Life Sciences 231 (2019) 116531

Contents lists available at ScienceDirect

Life Sciences

journal homepage: www.elsevier.com/locate/lifescie

Analysis of CRISPR/Cas system of *Proteus* and the factors affected the functional mechanism

Daofeng Qu[a], Shiyao Lu[a], Peng Wang[a], Mengxue Jiang[a], Songqiang Yi[b], Jianzhong Han[a,*]

[a] School of Food Science and Biotechnology, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China
[b] Jiangxi Animal Husbandry Technology Extension Station, Nanchang 330046, China

ARTICLE INFO

Keywords:
Proteus
CRISPR
Spacer
Cas
Mobile genetic element

ABSTRACT

Background: The *Proteus* is one of the most common human and animal pathogens. Clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins (CRISPR/Cas) are inheritable genetic elements found in a variety of archaea and bacteria in the evolution, providing immune function against foreign invasion.
Objectives: To analyze the characteristics and functions of the CRISPR/Cas system in *Proteus* genomes, as well as the internal and external factors affecting the system.
Methods: CRISPR loci were identified and divided into groups based on the repeat sequence in 96 *Proteus* strains by identification. Compared the RNA secondary structure and minimum free energy of CRISPR loci through bioinformatics, the evolution of *cas* genes, and the effects of related elements were also discussed.
Results: 85 CRISPR loci were identified and divided into six groups based on the sequence of repeats, and the more stable the secondary structure of RNA, the smaller the minimum free energy, the fewer base mutations in the repeat, the more stable the CRISPR and the more complete the evolution of the system. In addition, Cas1 gene can be a symbol to distinguish species to some extent. Of all the influencing factors, CRISPR/Cas had the greatest impact on plasmids.
Conclusions: This study examined the diversity of CRISPR/Cas system in *Proteus* and found statistically significant positive/negative correlations between variety factors (the RNA stability, free energy, etc.) and the CRISPR locus, which played a vital role in regulating the CRISPR/Cas system.

1. Introduction

Gram-negative Enterobacteriaceae bacteria widely distributes in nature and has a wide range of hosts. There are parasitic or symbiotic, epiphytic and saprophytic phenomena in humans, animals, and plants and they can easily be found in soil or water [27]. Some strains, like *Escherichia coli* and *Proteus* are important sources for the study of genetics and molecular biology.

Proteus includes five species, which are *Proteus vulgaris*, *Proteus mirabilis*, *Proteus viscous*, *Proteus pneumoniae*, and *Proteus hausmannii* [8]. Common proteobacteria and *Proteus mirabilis* are closely related to the clinic. *Proteus* food poisoning is one of the common food poisoning in China, the proteobacteria which caused food poisoning are mainly caused by *P. vulgaris* and *P. mirabilis* [6]. *Escherichia coli*, the representative strain of the genus *Escherichia*, is the most important and abundant type of bacteria in the intestine of humans and animals [15]. It is generally not pathogenic, and is a resident bacterium in human and animal intestines, under certain conditions, exactly, it can lead to intestinal infections [5].

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)/Cas system which comprises genomic (CRISPR) and proteomic (Cas) components are found in 75% bacteria and archaea, and mediate an adaptive immune response against invading viruses [4,7,20]. The genomic component is a DNA loci containing short fragments of targeted nucleic acid sequences (spacers) interspaced by short repeated sequences (repeats) [17]. The spacer sequences can be either foreign or self-origin [30]. The length of the repeat sequences varies between 25 and 40 nt, whereas the length of the spacer sequences varies between 21 and 71 nt [34]. As mentioned above, some spacers show high homology with foreign nucleic acids, but the origin of a significant percentage of spacers remains unknown [2].

The objective of this study was to gain the further insights into the character of CRISPRs in *Proteus* species by analyzing a collection of 96 unique strains. In this research, we characterized the CRISPR content and the presence of the mobile elements, regulators, etc., to explore the putative link between them.

* Corresponding author.
E-mail address: jianzhonghan1@163.com (J. Han).

https://doi.org/10.1016/j.lfs.2019.06.006
Received 24 March 2019; Received in revised form 28 May 2019; Accepted 3 June 2019
Available online 05 June 2019
0024-3205/

2. Materials and methods

2.1. Strains collection

We chose 96 *Proteus* strains and 50 *Escherichia coli* strains (for contrast) from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome/), and downloaded complete genomes and bioinformation of these strains with default parameters (Table S1). CRISPR loci and *cas* genes were searched in CRISPRs database (http://crispr.i2bc.paris-saclay.fr/crispr/) [28] and CRISPR Finder [13], then we obtained the flanking sequence, repeat and spacer nucleotide sequence of these strains.

2.2. Identification and analysis of CRISPR

The classification of confirmed CRISPR loci were divided into six groups [21] based on six different repeats, named CRISPR1~6. The typical repeats and terminal repeats of CRISPR were analyzed through multiple sequence alignment using ClustalX [31], and these six confirmed CRISPR loci were visualized with Weblogo (http://weblogo.berkeley.edu/logo.cgi). These repeat sequences were regarded as the specific gene signature for CRISPR. Secondary structure prediction of the most frequent sequence of each repeat was performed by RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) [9], and the Minimum Free Energy of the RNA was obtained with current limits of 7500 nt for section function calculations and 10,000 nt for MFE-only predictions.

2.3. Analysis of spacers

We collected the number and nucleotide base pairs of spacers in all strains and made a statistical correlation between spacers and the repeats, and IS finder, INTEGRALL, CRISPRTarget were used to analyze spacers (Biswas, Gagnon, 2013, [12]). To identify the spacer sequences matching sequences from mobile elements, the spacers were subjected to the standard BLASTN search (e-value threshold, 1.10^{-5}) in Genbank, genetic mobile elements were defined by identifying homologous sequences with an e-value $< 1.10^{-5}$ and < 10% difference in sequence length [32].

2.4. Phylogenetic tree of Cas1 gene in Proteus and Escherichia coli

We used Cygwin64 terminal to obtain the nucleotide sequence of *cas* gene from 10,000 bp upstream to 10,000 bp downstream in the CRISPR loci. The MEGA7.0 program was used to estimate nucleotide diversity and evolutionary distances as well as to build phylogenetic trees by the neighbor-joining method using the Jukes-Cantor distances. The *cas* gene has 45 families; *Proteus cas* gene belongs to I-E and I-F types. The *cas* gene of the publicly available *Proteus* genome was chosen to construct the *Proteus cas* gene tree in order to identify the *Proteus cas* gene, which was derived from 107 strains of *Proteus* that were downloaded from NCBI. Based on the *cas* gene sequence, BLAST was performed on *Proteus* and the *cas* gene clusters present in each strain were obtained. Through Cas1, strains with these genes were selected to construct a homologous tree of *Proteus cas* genes.

2.5. The distribution of mobile genetic elements and regulator in strains

To identify the distribution of phage and plasmid in *Proteus* and *Escherichia coli*, we used Prophinder (http://aclame.ulb.ac.be/Tools/Prophinder/) [24] and Addgene (http://www.addgene.org/) [19]. We submitted Genbank files of genome data obtained from NCBI to Prophinder and get the results for prophage prediction. The genetic bioinformation of strains was obtained by Cygwin64 Terminal software, and the number of elements such as Insert Sequence, transposon, and integron contained therein was counted. After all the data were integrated, the statistical correlation between the data and CRISPR was analyzed by using Principle Component Analysis (PCA).

3. Results

3.1. Geographical comparison of CRISPR alleles

We selected all *Proteus* and *Escherichia coli* complete genomes available from the NCBI database, totaling 146 strains (Table S1). According to CRISPR database and Guo et al. [14], confirmed CRISPR sites should contain at least two different spacers. The number of CRISPR loci varied from 1 to 5 depending on the strains. Most strains have CRISPR loci and *cas* genes. In *Proteus*, only 43 strains contain CRISPR/Cas system. Statistical analysis results showed that, in *Proteus*, the number of direct repeats was between 3 and 17, and the number of spacers ranged from 2 to 16; and in *Escherichia coli*, the number of spacers ranged from 2 to 21, and the number of repeats was between 3 and 22 (typically 6, 7;3, 4, Table S2).

3.2. The profiles of Proteus CRISPRs

The CRISPR loci were assigned into six groups based on the repeat sequence similarity, since the direct repeat length of CRISPR loci was similar within each locus by multiple sequence alignment analysis. It was indicated that CRISPR1, CRISPR 5 and CRISPR6 were the most common confirmed loci in all strains; the number of each repeat was 180, 147 and 60. These groups were taken into account in the current classification of CRISPR/Cas system (Table 1).

In order to better understand the features of these CRISPR groups, we used Weblogo to analyze the differences between repeats including terminal repeats, repeat variants and typical repeats in the same CRISPR group, so that we can see individual nucleotide base mutations in six different CRISPR groups (Fig. 1). It was described the results of CRISPR, which showed that CRISPR5 and CRISPR6 had less mutation and high frequency. From the analysis of the diversity of base mutations, the implications of these findings confirmed which CRISPR's structure was stable in these six CRISPR groups, and the fewer base mutations in the CRISPR repeat sequence, the more stable the CRISPR and the more complete the evolution of the system.

Previous researches have indicated that CRISPR repeats may form stable hairpin-like secondary structures (classical stem-loop) due to the partially palindromic nature [3], which contains a large and a small loop at both ends of each repeats of CRISPR [22]. The RNA secondary structure and minimum free energy (MFE) was detected for typical direct repeat sequences of each group through the RNAfold Web Server (Fig. 1). From the short review above, key findings emerged that in these 6 CRISPR groups, we showed that, their RNA secondary structures almost have two rings at each end and a stem in the middle, except CRISPR4 which only had a circle. The stem length in other Group was from 3 to 7, which appeared highly conservative. Secondary structure prediction of the most frequent sequence of the first and terminal repeats of each CRISPR was performed by RNAfold. The free energy of the thermodynamic ensemble was −12.22, −12.62, −0.42, −1.72, −12.57, −12.07 kcal/mol. Although the presence and number of CRISPRs were relatively constant, the number of repeats in each locus always varied with strains. According to the result, the CRISPR2 and CRISPR3 have the lowest MFE, which means they had the most stable RNA secondary structure.

3.3. The effect of spacer structure on CRISPR loci

According to statistics of the data, the total number of spacers in strains was 657 in *Proteus*, and 1057 in *Escherichia coli*, respectively. It was known that the number of spacers had a significant inverse correlation between strains, suggested that CRISPR/Cas have permitted phage insertion by lacking its own spacers. So we made an analysis on

Table 1
The information of the confirmed CRISPR1∼CRISPR6.

CRISPR	Type	Repeat sequence (5′-3′)	No. of strains	No. of repeats	Frequency (%)
CRISPR1	Typical repeat	CGGTTCATCCCCGTGCATACGGGGAACAC	27	145	80.56
	Repeat variants	AGGTTTATCCCCGTGTATACGGGGAACAC		1	0.56
		CGATTCATCCCCGTGCATACGGGGAACAC		12	6.67
		TGGTTTATCCCCGTATATACGGGGAACAC		1	0.56
		TAATTTATCCCCGTATACACGGGGAACAC		7	3.89
		CGGTTCATCCCCGTGCATACGGGGGAACA		1	0.56
	Terminal repeat	CGGTTTATCCCCGTGCATACGGGGAACAC		13	7.22
CRISPR2	Typical repeat	GTGTTCCCCGTATGCACGGGGATAAACCG	11	24	58.54
	Repeat variants	GTGTTCCCCGTATGCACGGGGATAAACCG		24	58.54
	Terminal repeat	GTGTTCCCCGTATGCACGGGGATGAATCG		11	26.83
		GTGTTCCCCGTGTATACGGGGATAAATTA		6	14.63
CRISPR3	Typical repeat	ATAATTGCCTTTAGGTTGATATTT	20	20	35.09
	Repeat variants	ATAATTTCCTTTAAGTTGATATTT		6	10.53
		ATAATTCCCTTTAGGCTGATATTT		2	3.51
		ATAGCTACCTTTAGGCTGATACTT		1	1.75
		ATACTGCACCAACGACGACCTTT		1	1.75
		CTCAATACTTTTATATTGATACTT		1	1.75
		TCCTTTTCCTTTTGGCTGATACTT		8	14.04
		GTACTGCATCAACGGCGACCTTT		1	1.75
		ATAGTTACCTTTAGGCTGATACTT		1	1.75
	Terminal repeat	ATAATTGCCTTTGGGTTGATAATT		14	24.56
		TAACTGCATCAACGGCGACCTTT		1	1.75
		GCTATAACCTTCCGCTTGATATTT		1	1.75
CRISPR4	Typical repeat	AAATATCAACCTAAAGGCAATTAT	9	12	38.71
	Repeat variants	GCTTTTCAGCCAAAAGGGAATTAT		9	29.03
		AAATATCAGCCTAAGGGGAATTAT		9	29.03
	Terminal repeat	AAATATCAACTTAAAGGAAATTAT		1	3.23
CRISPR5	Typical repeat	GTGTTCCCCGTATGCACGGGGATGAACC	45	135	91.84
	Repeat variants	GTGTTCCCCGTATGCACGGGGATGAACC		135	91.84
	Terminal repeat	GTGTTCCCCGTATGCACGGGGATGAATC		12	8.16
CRISPR6	Typical repeat	ATCCCCGTATACACGGGGAACAC	8	54	90.00
	Repeat variants	ATCCCCGTATACACGGGGAACAC		54	90.00
	Terminal repeat	ATTCCCGTATACACGGGGAACAC		6	10.00

the number and length of spacer in CRISPR (Fig. 2a-f). Analysis of the spacer size distribution indicated that variability was greatest in Group 5, and Group 2 had the lowest variability ($P < 0.05$), and most groups had polymorphic unique spacers. The spacer size ranged 31 to 33 bp of Group 1 and Group 3, the typical spacer size was 32 bp, and 58 bp of Group 2. The typical size was 38 bp of Group 4; compared to a typical size of 42 bp of Group5 ranging from 39 bp to 60 bp. Group 6 had a large number of 32 bp spacers > 33 bp. The proportions of spacers of typical size was 94.04% (205 of 218), 100% (58 of 58), 90.53% (172 of 190), 68.42% (13 of 19), 58.82% (20 of 34), 92.75% (128 of 138), for Group 1 to 6, respectively. The spacer length has been shown to influence the activity of CRISPR loci. Our data suggested an inverse correlation between the size of repeat and spacer (Fig. 2g). The number of spacer regions changes the least, and based on the above results, CRISPR1 and CRISPR3 are the most stable ones.

From the perspective of matching bases with foreign gene sequences, we found in CRISPR 1–6, there were 130, 10, 82, 4, 7, and 46 unique spacer sequences, respectively, and the matching foreign sequences were 809, 42, 508, 36, 23, and 331. Most of these foreign sequences were derived from the insertion sequence (IS), transposon, plasmid, and phage; it also proved the formation mechanism of spacer. The occurrence of spacers matched elements involved in antibiotic resistance gene mobilization (e.g. IS5, Tn3. IntI). Together, the present finding confirmed that repeats are negatively related to spacer size and change the activity of CRISPR loci, but further investigation is required [11].

3.4. Distribution and structural feature of cas genes in strains

Cas proteins are of great necessity for the function of the CRISPR/Cas immune system and are indicators of the system's activity [16]. We searched for *cas* genes from 10,000 bp upstream to 10,000 bp downstream the CRISPR loci in the NCBI database, so different subtypes of *cas* gene structure were found in all strains, which belonged to CRISPR/Cas Type I-E and I-F. *Cas* gene has eight subtypes, Cse represents E.coli, Csy represents ypest [25]. Some CRISPR members lack the corresponding *cas* genes (*cas* I-E and I-F, respectively) and in very rare occasions are both simultaneously found, for instance, CRISPR6. In this section, we found that there were sixteen strains that have the *cas* genes in the order of Cas2-Cas1-Cas6-Cas5-Cas7-Cse2-Cse1-Cas3 located downstream of the repeat-spacer region, and eight strains have the *cas* genes in the order of Cas3-Cse1-Cse2-Cas7-Cas5-Cas6-Cas1-Cas2, four strains have the *cas* genes in the order of Cse1-Cse2-Cse4-Cas5-Cas6-Cas1-Cas2. There was an important finding that, in each type, it all in habitats Cas1 gene, thus we can boldly suggest that the Cas1 gene was ubiquitous in strains. Through comparing the Cas1 gene in different strains by constructing the homologous evolutionary tree, we did further research and analyzed the effect of Cas1 gene in the evolution of *Proteus*; the studies found that the *Proteus* between different genera had a close relationship (Fig. 3). From these result it was clear that Cas1 gene can classify all bacteria between species, according to the nucleotide similarity. So Cas can better classify bacteria than other genes, and nearly all bacteria contain Cas1. To better understand the features of the CRISPR/Cas system, six representative strains (*Escherichia coli* 042, *Proteus hauseri* ATCC 700826, *Proteus mirabilis* T21, *Proteus mirabilis* AR_0059, *and Escherichia coli* APEC O1, *Escherichia coli* UMN026) were chosen for the next research (Fig. 4). These strains have different *cas* genes structure, so the functions they play were also different.

3.5. The relationship between CRISPR and mobile genetic elements, regulators, and DNA-related enzyme

It is well known that the CRISPR/Cas system can protect against the invasion of mobile genetic elements (MGE), for instance, plasmid and bacteriophage, which often carries antibiotic resistance genes (ARG) (Bhaya, Davison, 2011). Previous studies [18] also indicated that

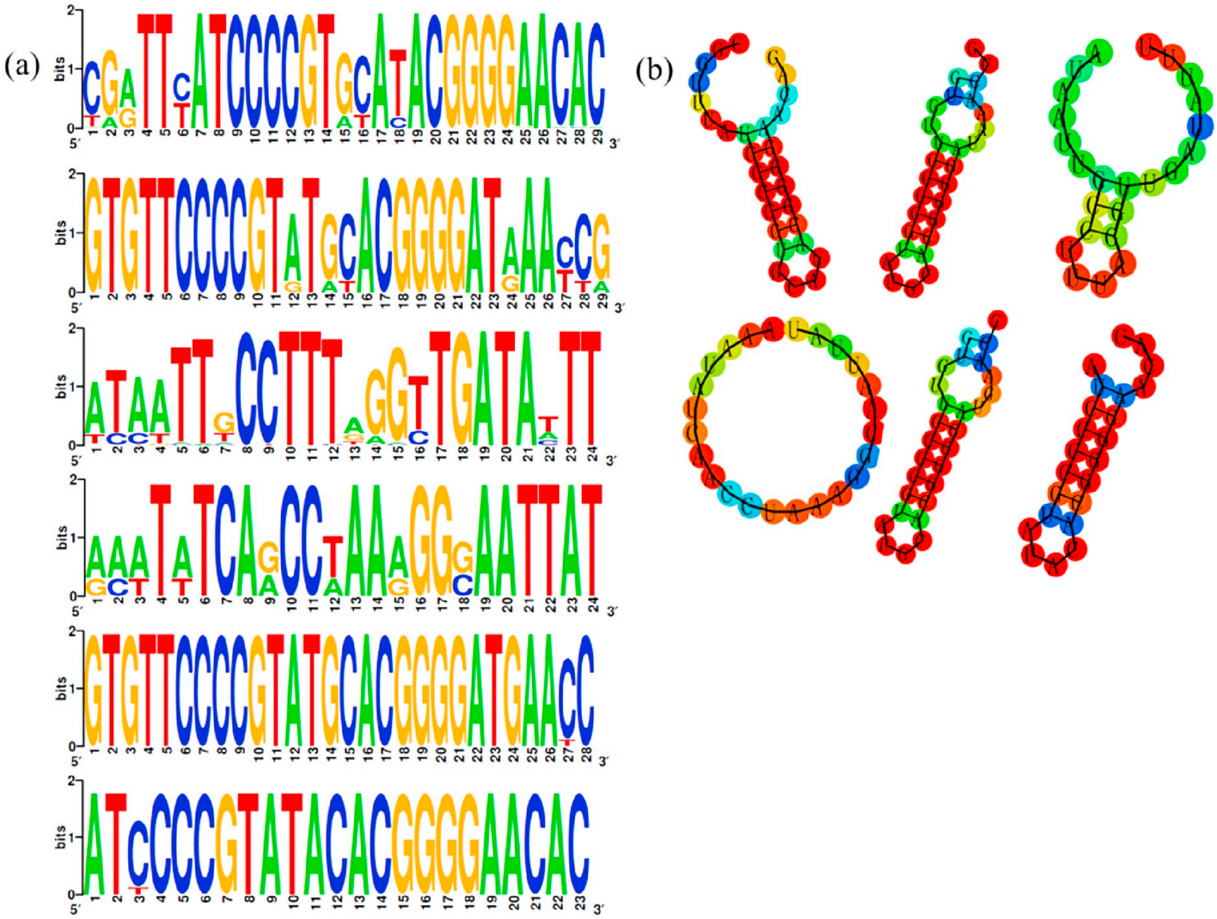


Fig. 1. (A) The Weblogo of repeats of CRISPR1∼CRISPR6. The sequence were first and terminal repeats of CRISPR1∼CRISPR6. (B) The secondary structure of repeats of CRISPR1∼CRISPR6. Secondary structure prediction of the most frequent sequence of the first and terminal repeats of each CRISPR was performed by RNAfold. The free energy of the thermodynamic ensemble was −12.62, −1.72, −0.42, −12.57, −12.07 kcal/mol.

mobile elements promote high variation of bacterial CRISPR loci, and CRISPR can defense plasmid or phage attack in host; bacteria have evolved CRISPR-mediated adaptive immunity against large populations of mobile elements in the nature. We have found in this research that some strains lacked CRISPR and these strains possess significantly more phages and plasmids than CRISPR harboring strains. Meanwhile, DNA-related enzymes also played a crucial role in the transcription and translation of CRISPR systems. Therefore, we wanted to further evaluate this problem—which factor, genetic mobile elements, regulators, or DNA-related enzyme, has the greatest impact on CRISPR, in other words, the most significant by means of conducting a PCA research to have an overview of the data. Through correlation analysis, the subtype I-E or I-F *cas* genes and the CRISPR were closely linked, the correlation coefficient was as high as 82.7% in *Proteus*. In addition, the correlation between plasmids and CRISPR was higher than that of phages, IS, transposons, and integrons in the analysis of mobile genetic elements, and the coefficients were 32.4%, 14.5%, 20.9%, 17.2%, respectively, but the result was the exactly opposite of the data in *E. coli*, what's amazing was that it's completely negative compared with *Proteus*. We would like to make a guess analysis of it later. The regulators like TetR/AcrR transcriptional regulator and DNA-related enzyme showed negative and positive correlation, respectively, also the coefficients were all relatively small. In Fig. 5, this PCA indicated that the strain phylogeny and the CRISPR activity were correlated and the MGE and ARG were also correlated, points in the same region indicated that they had a certain similarity. For example, compared with the first region, the strain in the third region did not contain CRISPR structure, and Spacer and *cas* genes played a major role in PCA in Fig. 5B. Differently, diversity was shown in *E. coli* (Fig. 5A), it may be due to the complete evolution of *Escherichia coli* and the relatively complete database. We can see that the principal components of the first region and the third region are *cas* gene and IS respectively. In the second and fourth regions, DNA-related enzymes and regulators played a major role in the strains, but there were few strains in these two regions, most of which were in the other two regions. We concluded that the principle effect on the stability of CRISPR systems was its own structure and mobile genetic elements particularly plasmids.

3.6. The effect of CRISPR/Cas system on hybrid plasmids

From the above data, we suggested that the relationship between plasmid and CRISPR loci was closer than any other elements (Table 2). In consequence we began to analyze the characteristics and structure of two different plasmids (*Escherichia coli* NRG875C-pO83_CORR and *Proteus mirabilis* T21-pT212), to found the relationship between CRISPR and plasmid (Fig. 6). The pO83_CORR genome sequence contains two CRISPR loci without *cas* genes, meanwhile, pT212 genome sequence contains two CRISPR loci with type I-E *cas* genes, they had circularly closed DNA sequences and contain 153 and 215 total predicted open reading frames (ORF), respectively. Here we compared the results of the schematic maps of two plasmids, plasmid pO83_CORR had more replicons than pT212, the latter contains only one replicon (repA), hence plasmid pO83_CORR was a hybrid plasmid with four replicons. We knew from the results of previous studies that the non-binding transposon of the resistance gene can be transposed between chromosome and plasmid, or between the plasmids, which promoted the spread of

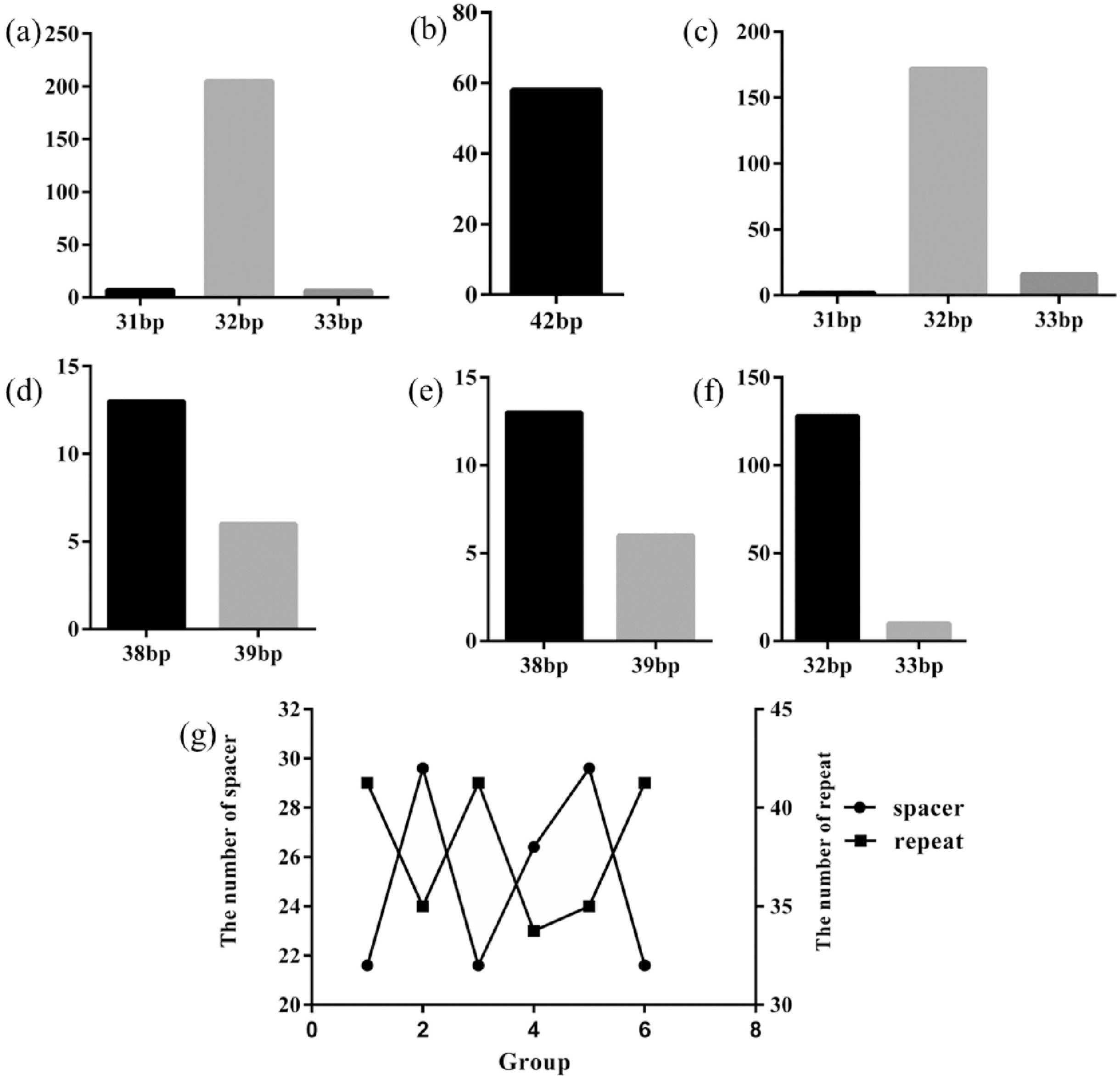


Fig. 2. Six Groups of CRISPR spacer size variability. The relationship between the size of repeat and spacer among six groups: (A) Group 1 spacers; (B) Group 2 spacers; (C) Group 3 spacers; (D) Group 4 spacers; (E) Group 5 spacers; (F) Group 6 spacers; (G) The x-axis represents the size of the CRISPR spacers, the y-axis represents the number of the CRISPR spacer. The size of repeat and spacer were inversely correlated.

the resistance gene between the bacteria by the hybrid plasmid. In Fig. 6a, there were more resistance genes such as tetA and tetR. The Tra gene can pull the distance between the conjugative plasmids and stabilize the binding. There were more Tra family genes, IS sequences, transposons and integron in the plasmid pO83-CORR, suggesting the diversity was great, and because of the existence of *cas* gene that it has a stronger capacity (more accessory modules). In the context of CRISPR acting as an immune system, differences in its activity among strains would be expected, for instance due to genetic diversity or the varied inducing factors they encounter in their respective habitat. These factors included the frequency when they faced invaders, the diversity of such invaders or the occurrence.

4. Discussion

The overarching goal of this study was to provide an in-depth analysis of type I-E and I-F CRISPR/Cas system in *Proteus* by comparison with *Escherichia coli*. All *Proteus* analyzed to date exclusively harbor a type I-E system, whereas some *E.coli* have been shown to contain type I-F. Together, through various analyses, the present findings displayed which inner or outer factor influences the CRISPR/Cas system's stability. It was found that plasmids affected them the most except for their internal structure.

Some CRISPR loci in *Proteus* are highly conserved, like CRISPR2, the frequency of base changes in the direct repeat sequence is minimal, and the minimum free energy is also the lowest. We suggested that longer repeats have a more stable secondary structure because there are more nucleotide base pairs. Moreover, the stability of RNA secondary structures may strengthen the function of CRISPR loci. The spacer diversity had been studied before, and many unique spacers changed rapidly in the *Proteus* and *Escherichia coli*. The spacer is inserted into the CRISPR loci near the leader sequence (Di, Ye, 2014). The older spacers especially the last spacers are more likely to be deleted to ensure the stability of the bacterial genome [10]. In *Proteus* strains, only CRISPR1 and CRISPR3 have a relatively high number of spacers. Depending on CRISPRTarget, the spacer-matched genes in CRISPR loci code for Cas protein. However, there are no *cas* genes around CRISPR4 and CRISPR6, the explanation of this situation might be that, CRISPR loci

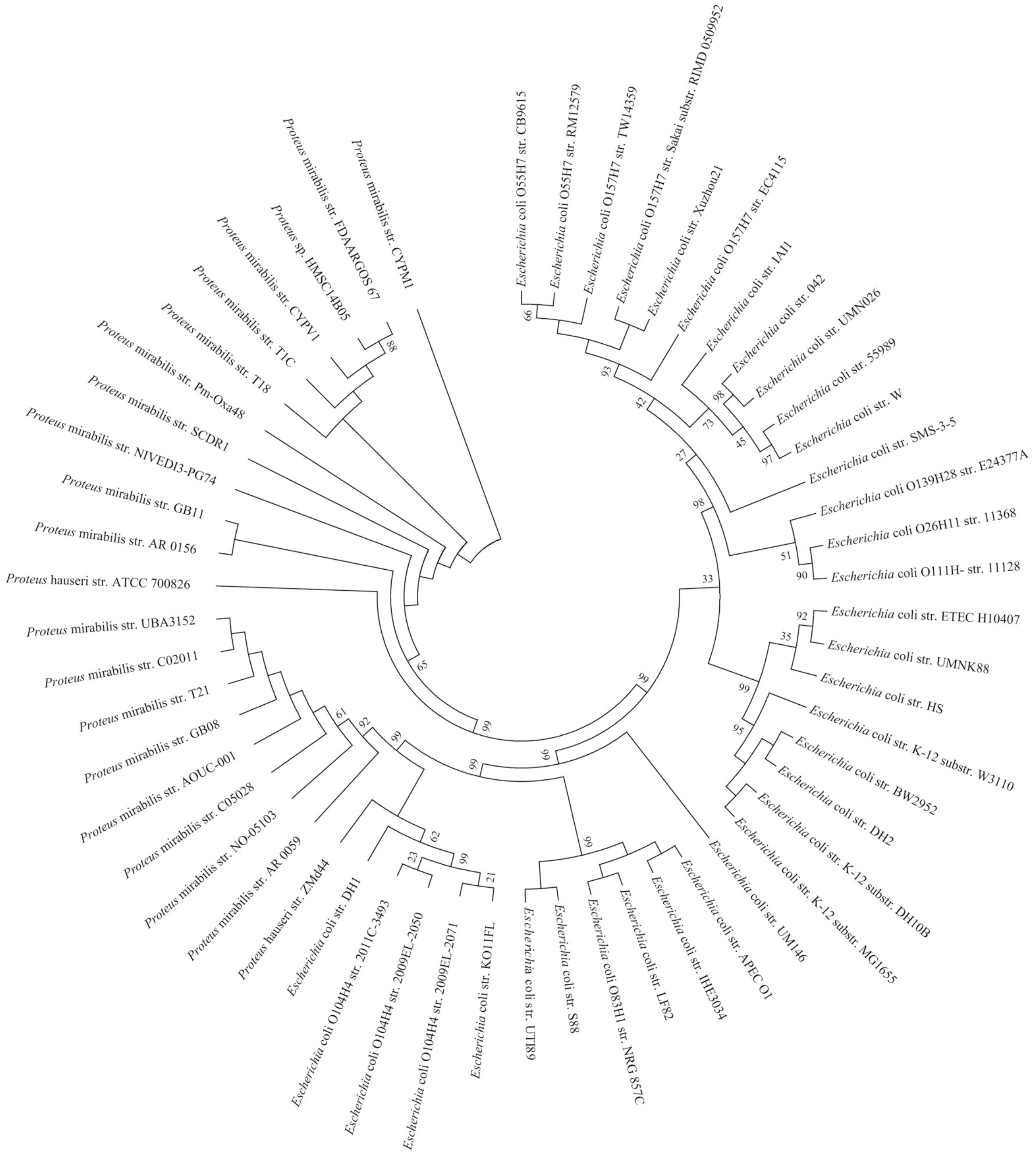
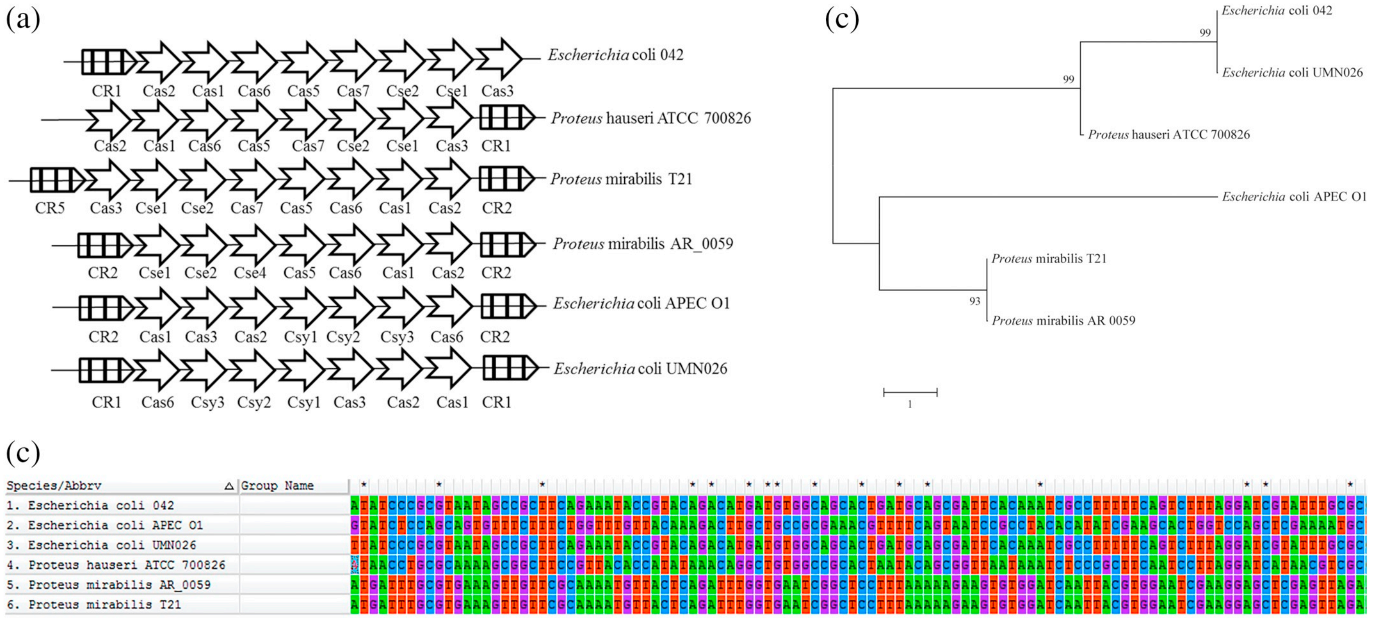

Escherichia coli O55H7 str. CB9615
Escherichia coli O55H7 str. RM12579
Escherichia coli O157H7 str. TW14359
Escherichia coli O157H7 str. Sakai substr. RIMD 0509952
Escherichia coli str. Xuzhou21
Escherichia coli O157H7 str. EC4115
Escherichia coli str. IAI1
Escherichia coli str. 042
Escherichia coli str. UMN026
Escherichia coli str. 55989
Escherichia coli str. W
Escherichia coli str. SMS-3-5
Escherichia coli O139H28 str. E24377A
Escherichia coli O26H11 str. 11368
Escherichia coli O111H- str. 11128
Escherichia coli str. ETEC H10407
Escherichia coli str. UMNK88
Escherichia coli str. HS
Escherichia coli str. K-12 substr. W3110
Escherichia coli str. BW2952
Escherichia coli str. DH2
Escherichia coli str. K-12 substr. DH10B
Escherichia coli str. K-12 substr. MG1655
Escherichia coli str. UM146
Escherichia coli str. APEC O1
Escherichia coli str. IHE3034
Escherichia coli str. LF82
Escherichia coli O83H1 str. NRG 857C
Escherichia coli str. S88
Escherichia coli str. UTI89
Escherichia coli str. KO11FL
Escherichia coli O104H4 str. 2009EL-2071
Escherichia coli O104H4 str. 2009EL-2050
Escherichia coli O104H4 str. 2011C-3493
Escherichia coli str. DH1
Proteus hauseri str. ZMd44
Proteus mirabilis str. AR 0059
Proteus mirabilis str. NO-05103
Proteus mirabilis str. C05028
Proteus mirabilis str. AOUC-001
Proteus mirabilis str. GB08
Proteus mirabilis str. T21
Proteus mirabilis str. C02011
Proteus mirabilis str. UBA3152
Proteus hauseri str. ATCC 700826
Proteus mirabilis str. AR 0156
Proteus mirabilis str. GB11
Proteus mirabilis str. NIVEDI3-PG74
Proteus mirabilis str. SCDR1
Proteus mirabilis str. Pm-Oxa48
Proteus mirabilis str. T18
Proteus mirabilis str. T1C
Proteus mirabilis str. CYPV1
Proteus sp. HMSC14B05
Proteus mirabilis str. FDAARGOS 67
Proteus mirabilis str. CYPM1
66
93
42
98
73
45
97
27
98
51
90
33
92
35
99
95
99
99
99
99
62
23
99
21
61
92
99
65
99
88

Fig. 3. The evolutionary tree of Cas1 of all strains. The Cas1 has 57 strains, respectively. The Cas1 genes sequence were obtained by searching for the complete genome sequences in Genbank. Strains in one branch indicate most evolutionary similarities, the branch represented that these sequences could be divides into groups by certain values and the percentage of each branch showed the sequence similarity, and the evolutionary distance scale of Cas1 in 0.10.

would select appropriate conditions for their host to stabilize the loci against external environment [33]. Therefore, this strain may participate in the defense against several antibiotics, but this hypothesis needs to be further investigated. Many studies demonstrated that there are only a few spacers that can match the proto-spacer perfectly due to many plasmids and phages, high rate of evolution of plasmids or phages [16,29]. We assume that non-coding region of spacer can regulate coding region of the same region. It has been suggested that splicing and recombination of gene segments is one way to acquire new spacers. Based on the spacer diversity and structure, other researchers figured out a way to classify spacers through combination of BLAST and CRISPRTarget, leading to easier analysis of spacers, depending on these, we can do more research. Thus, future studies on more spacers are needed to reveal the features of spacer-formation mechanisms.

The six subtypes of the *cas* gene located in the vicinity of CRISPR loci were identified from the NCBI database and were consistent with previous studied [7]. Analysis of *cas* genes either in absence/presence and the length of CRISPR strongly support the linkage between pairs of CRISPR system [23]. Like *Escherichia coli*, cluster of *cas* genes in *Proteus* appears adjacent to CRISPR loci that contain identical repeats to those of CRISPR2. The *cas* genes in all strains are not unchangeable. The phenomenon of depletion is as common as in *Escherichia coli*. The homology tree of Cas1 indicates the evolution of the *cas* genes, CRISPR in *Escherichia coli* evolves more stable than *Proteus*, absolutely, more *cas* genes and CRISPR loci in the former. Interestingly, the result postulated the potential co-evolution of Cas1 genes and CRISPR repeats, indicating a potential functional linkage between them.

This immune function was derived from the dynamic changes in the CRISPR loci. While CRISPR loci provide mobile genetic element antibiotics, some phage can continue to infect host cells that have been immunized [1]. The research found that this phenomenon was mainly related to specific site mutations in the genome.


(a)
CR1 Cas2 Cas1 Cas6 Cas5 Cas7 Cse2 Cse1 Cas3
Escherichia coli 042
Cas2 Cas1 Cas6 Cas5 Cas7 Cse2 Cse1 Cas3 CR1
Proteus hauseri ATCC 700826
CR5 Cas3 Cse1 Cse2 Cas7 Cas5 Cas6 Cas1 Cas2 CR2
Proteus mirabilis T21
CR2 Cse1 Cse2 Cse4 Cas5 Cas6 Cas1 Cas2 CR2
Proteus mirabilis AR_0059
CR2 Cas1 Cas3 Cas2 Csy1 Csy2 Csy3 Cas6 CR2
Escherichia coli APEC O1
CR1 Cas6 Csy3 Csy2 Csy1 Cas3 Cas2 Cas1 CR1
Escherichia coli UMN026
(c)
Escherichia coli 042
99
Escherichia coli UMN026
99
Proteus hauseri ATCC 700826
Escherichia coli APEC O1
Proteus mirabilis T21
93
Proteus mirabilis AR 0059
1
(c)

```
Species/Abbrv                △ Group Name
1. Escherichia coli 042        ATATCCCGCGTAATAGCCGCTTCAGAAATACCGTACAGACATGATGTGGCAGCACTGATGCAGCGATTCACAAATCGCCTTTTTCAGTCTTTAGGATCGTATTTGCGC
2. Escherichia coli APEC O1    GTATCTCCAGCAGTGTTTCTTTCTGGTTTGTTACAAAGACTTGCTGCCGCGAAACGTTTTCAGTAATCCGCCTACACATATCGAAGCACTGGTCCAGCTCGAAAATGC
3. Escherichia coli UMN026     TTATCCCGCGTAATAGCCGCTTCAGAAATACCGTACAGACATGATGTGGCAGCACTGATGCAGCGATTCACAAATCGCCTTTTTCAGTCTTTAGGATCGTATTTGCGC
4. Proteus hauseri ATCC 700826 ATAACCTGCGCAAAAGCGGCTTCCGTTACACCATATAAACAGGCTGTGGCCGCACTAATACAGCGGTTAATAAATCTCCCGCTTCAATCCTTAGGATCATAACGTCGC
5. Proteus mirabilis AR_0059   ATGATTTGCGTGAAAGTTGTTCGCAAAATGTTACTCAGATTTGGTGAATCGGCTCCTTTAAAAAGAAGTGTGGATCAATTACGTGGAATCGAAGGAGCTCGAGTTAGA
6. Proteus mirabilis T21       ATGATTTGCGTGAAAGTTGTTCGCAAAATGTTACTCAGATTTGGTGAATCGGCTCCTTTAAAAAGAAGTGTGGATCAATTACGTGGAATCGAAGGAGCTCGAGTTAGA
```

Fig. 4. (A) Conservation of cas genes subtypes across six different strains. Complete genomes of six strains were available in Genbank. The arrows represented cas genes, the CR was abbreviators CRISPR, all schematics were not drawn to scale. (B). The homology tree of Cas1 of six represented strains. (C) Base alignment of six strains of Cas1 gene.

At present, people's research on CRISPR has found that the CRISPR system is related to the transfer of horizontal genes in order to better adapt to the environment (Guo, Wang, 2015). *S. aureus* found that the CRISPR system can limit horizontal gene transfer and prevent the spread of drug resistance genes between strains. Bacteria can resist the stress of living environment by acquiring virulence genes, and CRISPR also affects the virulence of bacteria. There have been significant advances and breakthroughs in the CRISPR/Cas research, but there are still many problems that have not been solved, such as how CRISPR/Cas evolved, one that bacteria are obtained from ancient bacteria, and that

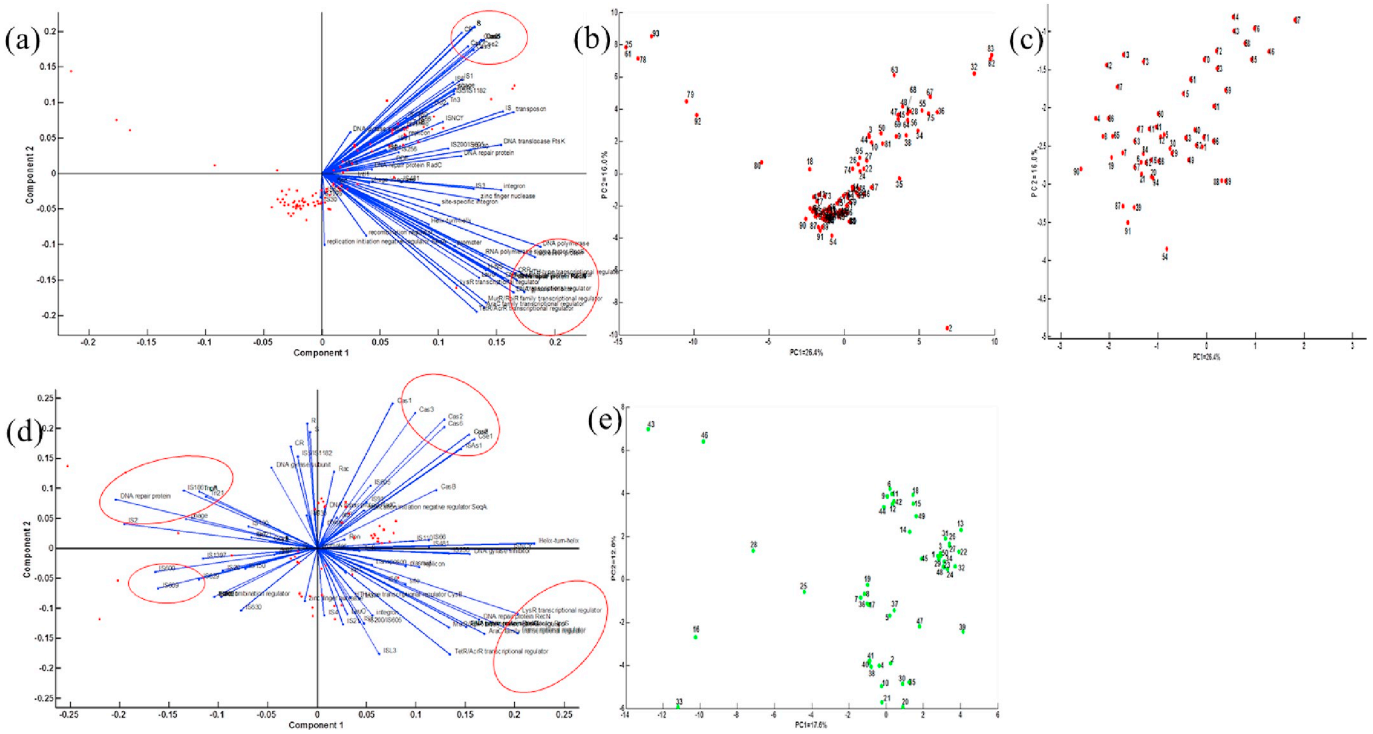


Fig. 5. Principle Component Analysis to compare *Proteus* and *Escherichia coli* about the correlation of CRISPR and mobile genetic elements, regulators and DNA-related enzymes. (A) Correlation vector plot of factors (blue lines) of the principle component axes and plot of *Proteus* strains (red dots). The vector indicates the correlation of each factor to the first principle component axes. Significantly correlated parameters are circled; (B and C). An enlarged view of the point which representing the *Proteus* strain. (D) Correlation vector plot of factors (blue lines) of the principle component axes and plot of *Escherichia coli* strains (red dots). (E) An enlarged view of the point which representing the *Escherichia coli* strain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
The distribution of plasmid in *Proteus* and *Escherichia coli.*

Species	Type	The number of plasmid	The number of replicon	CRISPR	No CRISPR
Proteus	Hybrid plasmid	6	13	4	2
	Normal plasmid	12	7	4	8
Escherichia coli	Hybrid plasmid	35	89	35	0
	Normal plasmid	50	24	50	0

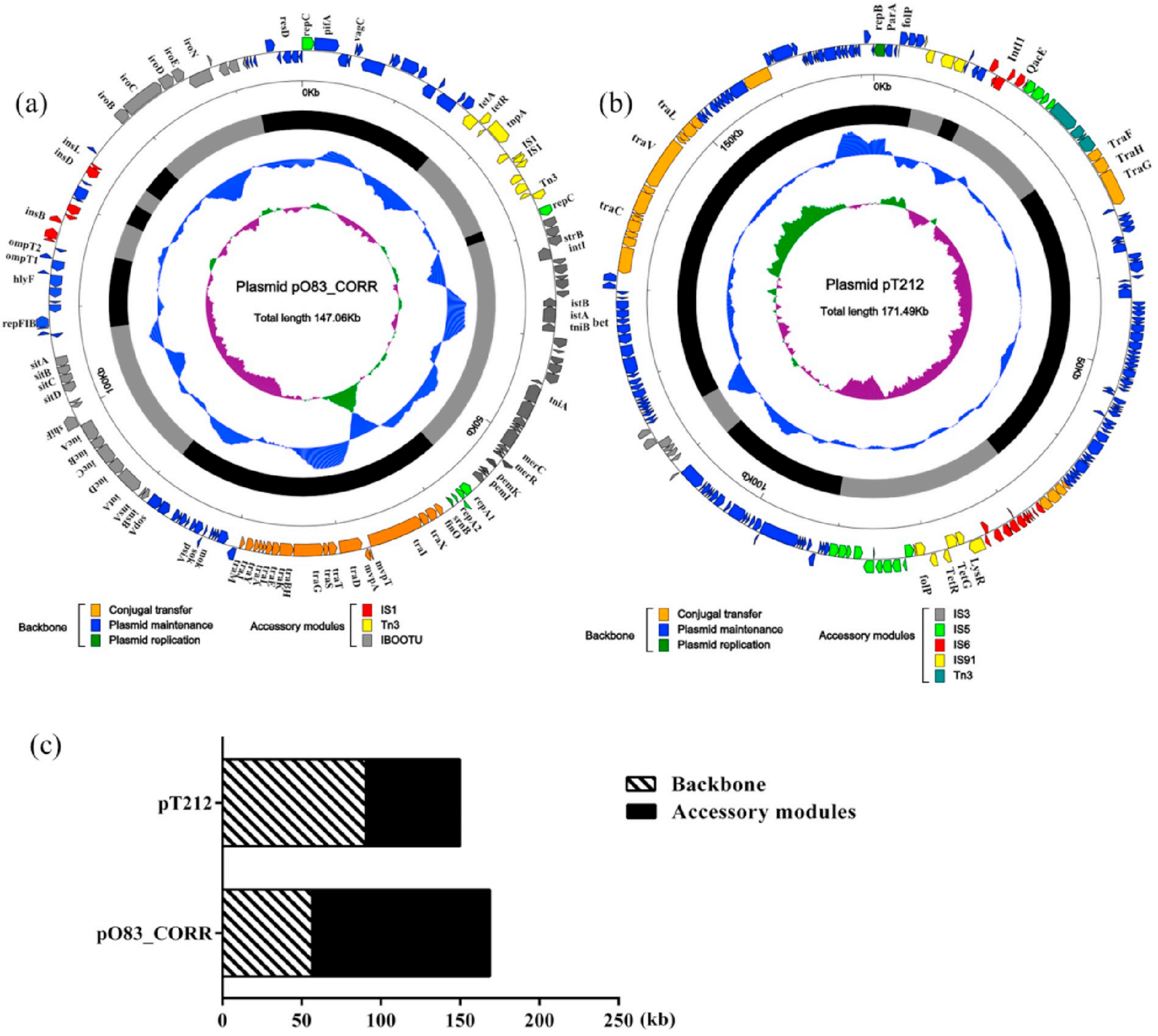


Fig. 6. Schematic diagrams of (A) plasmid pO83-CORR and (B) plasmid pT212. Genes are denoted by arrows and colored based on gene function classification. The innermost circle presents GC-Skew[(G-C)/(G + C)] with a window size of 500 bp and a step size of 20 bp. The blue circle presents GC content. Shown also are backbone and accessory module regions. The blue arrows represent plasmid maintenance, the green arrows represent plasmid replication. (C) The black box represents Backbone fragment, and the diagonal striped box represent Accessory modules fragment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the study is derived from transposons [26]. Why do some *Escherichia coli* have I-E CRISPR/Cas, and some *Escherichia coli* have type I-F, some of which are available, and some of which are not. Why *Escherichia coli* are from I-E CRISPR/Cas rich in spacer sequences, but in a laboratory condition, it is inactive? Under what conditions can the CRISPR/Cas system in *Escherichia coli* are activated? At present, what we know may be just the tip of the iceberg. There may be more interesting basic biological functions in the future that need to be further explored and discovered.

Declaration of Competing Interest

None to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lfs.2019.06.006.

References

[1] Y. Azuma, A. Hosoyama, M. Matsutani, N. Furuya, H. Horikawa, T. Harada, et al., Whole-genome analyses reveal genetic instability of Acetobacter pasteurianus, Nucleic Acids Res. 37 (2009) 5768–5783.

[2] M. Babu, N. Beloglazova, R. Flick, C. Graham, T. Skarina, B. Nocek, et al., A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair, Mol. Microbiol. 79 (2011) 484–502.

[3] D. Bhaya, M. Davison, R. Barrangou, CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation, Annu. Rev. Genet. 45 (2011) 273–297.

[4] A. Biswas, J.N. Gagnon, S.J. Brouns, P.C. Fineran, C.M. Brown, CRISPRTarget: bioinformatic prediction and analysis of crRNA targets, RNA Biol. 10 (2013) 817–827.

[5] H.W. Boyer, D. Roulland-Dussoix, A complementation analysis of the restriction and modification of DNA in Escherichia coli, J. Mol. Biol. 41 (1969) 459–472.

[6] E.A. Brewer, C.N. Dellarocas, A. Colbrook, W.E. Weihl, Proteus: a high-performance parallel-architecture simulator, ACM SIGMETRICS 20 (1991) 247–248.

[7] M. Burmistrz, K. Pyrä, CRISPR-Cas systems in prokaryotes, Pol. J. Microbiol. 64 (2015) 193–202.

[8] M.M. D'Andrea, T. Giani, D.A.L. Henrici, N. Ciacci, M. Gniadkowski, V. Miriagou, et al., Draft genome sequence of Proteus mirabilis NO-051/03, representative of a multidrug-resistant clone spreading in Europe and expressing the CMY-16 AmpC-type β-lactamase, Genome Announc. (2016) 4.

[9] R.B. Denman, Using RNAFOLD to predict the activity of small catalytic RNAs, Biotechniques 15 (1993) 1090–1095.

[10] H. Deveau, J.E. Garneau, S. Moineau, CRISPR/Cas system and its role in phage-bacteria interactions, Annu. Rev. Microbiol. 64 (2010) 475.

[11] H. Di, L. Ye, H. Yan, H. Meng, S. Yamasak, L. Shi, Comparative analysis of CRISPR loci in different listeria monocytogenes lineages, Biochem. Biophys. Res. Commun. 454 (2014) 399–403.

[12] R. Ge, G. Mai, W. Pu, M. Zhou, Y. Luo, Y. Cai, et al., CRISPRdigger: detecting CRISPRs with better direct repeat annotations, Sci. Rep. 6 (2016) 32942.

[13] I. Grissa, G. Vergnaud, C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats, Nucleic Acids Res. 35 (2007) W52.

[14] X. Guo, Y. Wang, G. Duan, Z. Xue, L. Wang, P. Wang, et al., Detection and analysis of CRISPRs of Shigella, Curr. Microbiol. 70 (2015) 85–90.

[15] D. Hanahan, Studies on transformation of Escherichia coli with plasmids, J. Mol. Biol. 166 (1983) 557–580.

[16] P. Horvath, Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus, J. Bacteriol. 190 (2008) 1401.

[17] P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea, Science 327 (2010) 167–170.

[18] Y. Hu, X. Yang, J. Li, N. Lv, F. Liu, J. Wu, et al., The bacterial mobile resistome transfer network connecting the animal and human microbiomes, Appl. Environ. Microbiol. 82 (2016) 6672–6681.

[19] J. Kamens, The Addgene repository: an international nonprofit plasmid and data resource, Nucleic Acids Res. 43 (2015) 1152–1157.

[20] E.V. Koonin, Y.I. Wolf, Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus-host coevolution, Mol. BioSyst. 11 (2014) 20–27.

[21] M. KS, H. DH, B. R, B. SJ, C. E, H. P, et al., Evolution and classification of the CRISPR-Cas systems, Nat. Rev. Microbiol. 9 (2011) 467.

[22] V. Kunin, R. Sorek, P. Hugenholtz, Evolutionary conservation of sequence and secondary structures in CRISPR repeats, Genome Biol. 8 (2007) R61.

[23] A. Levy, M.G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, et al., CRISPR adaptation biases explain preference for acquisition of foreign DNA, Nature 520 (2015) 505–510.

[24] G. Limamendez, J. Van Helden, A. Toussaint, R. Leplae, Prophinder: a computational tool for prophage prediction in prokaryotic genomes, Bioinformatics 24 (2008) 863–865.

[25] N. Listed, Proceedings of CRISPR evolution, mechanisms and infection, June 17-19, 2013, United Kingdom, Biochem. Soc. Trans. 41 (2013) 1383.

[26] M.L. Ostriahernández, C.J. Sánchezvallejo, J.A. Ibarra, G. Castroescarpulli, Survey of clustered regularly interspaced short palindromic repeats and their associated Cas proteins (CRISPR/Cas) systems in multiple sequenced strains of Klebsiella pneumoniae, BMC. Res. Notes 8 (2015) 332.

[27] D.L. Paterson, Resistance in gram-negative bacteria: enterobacteriaceae, Am. J. Infect. Control 34 (2006) (S20-S8).

[28] S.S. Rahmatabadi, N. Nezafat, M. Negahdaripour, N. Hajighahramani, M.H. Morowvat, Y. Ghasemi, Studying the features of 57 confirmed CRISPR loci in 29 strains ofEscherichia coli, J. Basic Microbiol. 56 (2016) 645–653.

[29] R. Sanozky-Dawes, K. Selle, S. O'Flaherty, T. Klaenhammer, R. Barrangou, Occurrence and activity of a type II CRISPR-Cas system in Lactobacillus gasseri, Microbiology 161 (2015) 1752.

[30] A. Stern, L. Keren, O. Wurtzel, G. Amitai, R. Sorek, Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. 26 (2010) 335–340.

[31] J.D. Thompson, T.J. Gibson, D.G. Higgins, Multiple sequence alignment using ClustalW and ClustalX, Curr. Protoc. Bioinformatics (2002) (Chapter 2: Unit 2.3.1-2.3.22).

[32] M. Touchon, S. Charpentier, D. Pognard, B. Picard, G. Arlet, E.P.C. Rocha, et al., Antibiotic resistance plasmids spread among natural isolates of Escherichia coli in spite of CRISPR elements, Microbiology 158 (2012) 2997–3004.

[33] P. Wang, B. Zhang, G. Duan, Y. Wang, L. Hong, L. Wang, et al., Bioinformatics analyses of Shigella CRISPR structure and spacer classification, World J. Microbiol. Biotechnol. 32 (2016) 38.

[34] A. Weinberger, M. Gilmore, CRISPR-Cas: to take up DNA or not—that is the question, Cell Host Microbe 12 (2012) 125–126.