The CRISPR-Cas System and its Relationship with MGEs in Klebsiella

Daofeng Qu (✉ daofeng@mail.zjgsu.edu.cn)

Zhejiang Gongshang University   https://orcid.org/0000-0002-6416-603X

Yuqiao Zhou

Zhejiang Gongshang University

Wei Zhou

Zhejiang Provincial Center for Animal Disease Control and Prevention

Jinzhi Zhou

Zhejiang Provincial Center for Animal Disease Control and Prevention

Ze He

Zhejiang Gongshang University

Enbao Zhang

Zhejiang Gongshang University

Xiner Zheng

Zhejiang Gongshang University

Shuaizhou Zong

Zhejiang Gongshang University

Shiyi Tian

Zhejiang Gongshang University

Jianzhong Han

Zhejiang Gongshang University

Abstract

Microorganisms have developed many strategies in the process of long-term defense against external attacks, one of which is the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated proteins (Cas) bacterial immunological system. In this paper, the whole genome of 300 strains of *Klebsiella* was collected, the CRISPR-Cas system in the strains was statistically analyzed, and the types and structures of CRISPR system in *Klebsiella* were explored, as well as the correlation between CRISPR and mobile genetic elements (MGEs). Through principal component analysis (PCA), we found that Cas gene, plasmids, integron, IS1, IS609 and DNA-related enzymes were closely related to CRISPR. Compared the structural characteristics of plasmids, the DinG family helicases, Cas6, Csf2, and IS5 were observed near the CRISPR loci in plasmid, which is also confirmed by the the results of PCA that they may be important factors affecting the plasmid whith CRISPR.

1. Introduction

Klebsiella is a gram-negative bacterium which is divided into three species in *Berger's Identification Bacteriology*: *Klebsiella* pneumoniae, *Klebsiella* acidophilus, *Klebsiella* native and *Klebsiella* phytophthora. The occurrence of infectious antibiotic resistance in this organism is a major problem worldwide. *K. pneumoniae* has a large antibiotic resistance gene pool, which they share with other *Enterobacteriaceae*, mainly through self-transferring plasmids (Navon-Venezia, et al. (2017). In these organisms, almost all modern antibiotic resistance (to carbapenems, cephalosporins, aminoglycosides, and now even colistin) is encoded in large (40-200 Kb) low-copy (16 per cell) conjugated plasmids (Moineau, 2015).

Over the past decades, researchers have studied some CRISPR-Cas systems and Cas proteins in detail. CRISPR-Cas is an adaptive immune system which stores memories of encounters with foreign DNA which are mostly mobile genetic elements (MGE) in unique spacing sequences extracted from the MGE and inserted into the CRISPR array (Moineau, 2015). Transcripts of CRISPR sequences are used to recognize homologous sequences and guide Cas nucleases to their unique targets when encountering familiar MGEs, resulting in inactivation of the latter (Barrangou, et al., 2017; Garcia-Martinez, et al., 2018). Like all defense mechanisms, the CRISPR-Cas system evolved in a long arms race with MGE, which has led to rapid evolution of some Cas gene sequences, primarily effect components, as well as significant diversity in the genetic composition and tissue of the CRISPR-Cas site. Like eukaryotic RNA interference and the argonte-centric defense mechanisms of prokaryotes, CRISPR-Cas belong to nucleic acid-oriented defense systems(Koonin, 2017; Swarts, et al., 2014; Takeuchi, et al., 2012). Among these mechanisms, however, CRISPR-Cas has the complete ability to create immune memories, representing true adaptive immunity. The complexity and diversity of the CRISPR-Cas system implies a complex evolutionary history.

The CRISPR system is divided into two types (polysubunit effector complex, mono-protein effector module) and six types (Types I, III, IV and Types II, V, VI), of which type I and type III are more studied. (Lino, et al., 2018; Makarova, et al., 2020; Ostria-Hernandez, et al., 2015). In the type I system, the CRISPR RNA (crRNA) complex recognizes the target DNA, which is then cleaved by Cas3. In type III system, the Cas10 protein is assembled into a complex that recognizes and cuts targets. The genomic CRISPR locus consists of three parts: trans-activated CRISPR RNA (tracrRNA) genes, Cas genes, and CRISPR repeat and spacers.

Some studies have identified type I-E and I-F CRISPR-Cas in many gram-negative *Enterobacteriaceae*, in addition, there are some studies have found that *Kebsiella pneumoniae* contains the CRISPR-Cas plasmid system (Medina-Aparicio, et al., 2018; Ostria-Hernandez, et al., 2015; Qu, et al., 2019). So far, there have been few studies on the type IV system in which such genes located on mobile elements. The relationship between CRISPR-Cas and MGEs is very complex. Some MGEs contribute to the origin and evolution of CRISPR-Cas, and conversely, the CRISPR-Cas system and its components are absorbed by some MGEs (Faure, et al., 2019).

This study examined the diversity of CRISPR-Cas system in *Klebsiella* strains, the relationship between MGEs and CRISPR-Cas system, especially the plasmid CRISPR-Cas were analyzed. Plasmid CRISPR-Cas directed against other plasmid, it may provide

another level of incompatibility in plasmid communities. Both plasmid and chromosomal CRISPR-Cas are evidently important determinants of the epidemiology of large antibiotic resistance plasmids in *Klebsiella*.

2. Material And Methods

2.1 Strains collection

In the National Center for Biotechnology Information database (https://www.ncbi.nlm.nih.gov/), we randomly chosed 300 strains *Klebsiella* strains. The whole genome download was saved as a FASTA format. Upload the whole genome sequence in the CRISPR Cas + + website (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index) to get each strain of CRISPR- Cas information (including the CRISPR locus, Cas gene, repetitive sequence and spacer, etc.).

2.2 Identification and analysis of CRISPR

Typical CRISPR repeats were sorted and stored in FASTA format, and ClustalX was used for multiple sequence alignment analysis. The confirmed CRISPR loci were divided into 7 categories according to the different repeat sequences, which were named as CRISPR1-7 (Makarova, et al., 2011). Web logo (http://weblogo.berkeley.edu/logo.cgi) was used to visualize the identified CRISPR site. These repeats are thought to be specific genetic markers for CRISPR. The secondary structures of single stranded RNA or DNA sequences were predited with RNAfold Web Server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Current limits are 7,500 nt for partition function continuously and 10,000 nt for minimum free energy only predicitions. MGEA7.0 software was used to construct phylogenetic trees of repeated sequences in *Klebsiella* CRISPR- Cas system for genetic evolution analysis.

2.3 Spacer sequence analysis

In order to identify the spacing sequences matched with the mobile elements, the spacing sequences in the CRISPR loci were sorted and saved in FASTA format. The spacing sequences were BLASTN search in Genbank using standard BLASTN search, e-value $<10^{-5}$ homologous sequences with a 10% difference in sequence length , identifying the genetic moving element.

2.4 Phylogenetic tree of Cas1 gene in *Klebsiella*

Most Cas genes in *Klebsiella* belong to type I-E and type I-F. Cas gene loci were obtained through the CRISPR Cas ++ (https://crisprcas.i2bc.paris-saclay.fr/) website, and the corresponding Cas1 gene sequence was extracted from the whole genome sequence with BioEdit software and stored as FASTA file. The MGEA7.0 program was used to estimate nucleotides diversity and evolutionary distance, as well as to construct phylogenetic trees by using the Neighbor connection approach of the Juke-Cantor distance.

2.5 The distribution of mobile genetic elements and regulator instrains

The FASTA file of the whole genome sequence obtained from NCBI was submitted to RAST website for gene annotation, and the obtained results were saved in the form of table. The gene of NCBI strain was used to count the insertion sequence, transposon, integron and DNA related enzymes. After all the data were integrated, the statistical correlation between the data and CRISPR was analyzed using principal component analysis.

2.6 Plasmids analysis

In order to analyze the characteristic structure of CRISPR-Cas system, the plasmids containing CRISPR-Cas were further analyzed. Two CRISPR-Cas plasmid genome sequences were downloaded from the NCBI database and uploaded to RAST to annotate the genes of the two plasmids. Based on the result of RAST, the sequences were uploaded to IS Finder, INTEGRALL, CRISPR Target, etc., respectively, to supplement the information of movable elements and drug-resistant genes, and to draw the plasmid structure diagram. (Ge, et al., 2016)

The correlation between plasmid CRISPR-Cas and mobile genetic elements was also analyzed by principal component analysis, the steps were same as described above (Qu, et al., 2019).

3. Results

3.1 Geographical comparison of CRISPR alleles

702 *Klebsiella* strains were selected from all available strains in NCBI database, 300 of which were randomly selected. The CRISPR locus was identified according to CRISPR Cas ++, which should contain at least two distinct intervals. The number of CRISPR loci varies from one to three, depending on the strain. Of the *Klebsiella* strains collected, only 95 had the CRISPR locus, and 12 had the CRISPR locus in their plasmids. The results of statistical analysis showed that the number of direct repeats in *Klebsiella* was between 3 and 66, and the number of spacers was between 2 and 65.

3.2 The profiles of *Klebsiella* CRISPRs

The CRISPR loci were divided into 7 groups according to repeat sequence similarity. Through multi-sequence alignment analysis, the direct repeat length of CRISPR gene in each locus was similar. The results showed that CRISPR2, CRISPR3 and CRISPR6 were the most common confirmed loci in all strains. The number of repetitions was 180, 147 and 60, respectively (Table 1).

The results from describing the CRISPR cohort suggest that CRISPR2 and CRISPR3 have fewer mutations and higher frequencies (Fig 1). From the diversity analysis of base mutations, the significance of these findings is to confirm that the CRISPR structure is relatively stable in the seven groups, that the fewer base mutations in the CRISPR repeat sequence, the more stable the CRISPR, and the more complete the systemic evolution.

CRISPR repeats may form stable hairpin like secondary structures (classical stem rings), with each CRISPR repeat containing a large ring and a small ring at each end. In these 7 CRISPR groups, there are two rings at each end of the RNA secondary structure and a stem in the middle, which is 5-7 in length and highly conserved (Fig 2). The free energies of the thermodynamic ensemble are -11.60, -11.70, -15.20 -11.80, -12.70, -14.20, -8.60 kcal/mol. Although the presence and number of CRISPRs is relatively constant, the number of repeats at each site always varies from strain to strain. The results show that CRISPR2 and CRISPR3 have the lowest MFE, meaning they have the most stable RNA secondary structure.

3.3 The effect of spacer structure on CRISPR loci

According to data statistics, the total number of spacers in *Klebsiella* strains was 2549. Spacing length has been shown to affect the activity of the CRISPR locus. Our data shows a negative correlation between the size of the repeat and the interval block (Fig 3 h). According to the above results, the most stable ones are CRISPR1 and CRISPR3.

From the perspective of base matching with exogenous gene sequences, we found that CRISPR 1-7 had 51, 197, 111, 176, 100, 68, and 8 special spacing sequences, respectively, and the exogenous matched sequences were 262, 1512, 1617, 1040, 751, 1197, and 73, respectively. Most of these foreign sequences come from insertion sequences (IS), transposons, plasmids and phages. The formation mechanism of spacer was proved. The presence of spacers was matched with elements associated with antibiotic resistance gene mobilization (e.g., IS5, Tn3). Taken together, the current findings confirm that repeated sequences are negatively correlated with the size of the spacer block and alter the activity of the CRISPR site, while further research is needed.

3.4 Distribution and structural feature of Cas genes in strains

Cas protein is an important component of CRISPR/ Cas immune system function and an indicator of immune system activity. In this section, we found that there are 29 strains have the *cas* genes in the order of Cas3-Cse1-Cse2- Cas6- Cas7- Cas5- Cas1-Cas2, and there are 10 strains have the *cas* genes in the order of Cas2- Cas1- Cas5- Cas7- Cas6-Cse2-Cse1- Cas3,10 strains have the *cas* genes in the order of Cas3-Cse1-Cse2- Cas7- Cas5- Cas6- Cas1- Cas2, 8 strains have the *cas* genes in the order of

Cas2- Cas1- Cas6- Cas5- Cas7-Cse2-Cse1- Cas3 and 1 strain has the *cas* genes in the order of Cas1-Cas3-Cas 2-Csy1-Csy2-Csy3- Cas 6,1 strain has the *cas* genes in the order of Cas 3-Cse1- Cas 6- Cas 5- Cas 1- Cas 2.

One important finding is that in almost every type, the Cas1 gene is present in its habitat, so it is safe to assume that the Cas1 gene is ubiquitous in *Klebsiella*'s CRISPR-Cas system. By constructing homologous evolutionary trees, Cas1 gene of different strains was compared to conduct further research and analyze the role of Cas1 gene in *Klebsiella* evolution. Thus, Cas1 gene can be used to roughly classify bacteria among species according to nucleotide similarity. Compared to other Cas genes, Cas1 is more representative, because almost all bacteria contain Cas1.

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

CRISPR-Cas systems are known to resist MGE invasions, such as plasmids and phages, which often carry antibiotic resistance genes (ARG). The relationship between CRISPR-Cas system and MGEs is complex and diverse. MGEs can promote the high variation of CRISPR locus in bacteria, and CRISPR can defend against MGE attacks.

We performed principal component analysis (PCA) on whether CRISPR has an impact on genetic moving elements, regulators, and DNA-related enzymes. Through PCA, we found that Cas gene was most closely related to CRISPR (the coefficients were 56%-62%), and plasmids, integrons, IS1, IS609 and DNA-related enzymes had high correlation with CRISPR (the coefficients were 29.4%, 26.6%, 20.4%, 42.1% and 35.4%, respectively).

3.6 The effect of CRISPR-Cas system on plasmids

To observe the structural characteristics of the CRISPR-Cas system, we analyzed the plasmid containing CRISPR (P15WZ-82_Vir) and the plasmid without CRISPR (pKpvST101_5), and compared the distribution of mobile genetic elements and regulatory factors on the two plasmids. We compared the graphical results of the two plasmids, and observed that both plasmids contained multiple replicons, as well as a comparable number of *Tra* family genes, IS sequences, transposons, and integrons, indicating similar levels of diversity of the two plasmids. Interestingly, we also found that the CRISPR sites on the P15WZ-82_VIR plasmid contained DinG family helicase, type I-E CRISPR-associated protein Cas6/Cse3/CasE, CRISPR-associated protein Csf2 and IS5.These genes and mobile elements were not found in pKpvST101_5, a plasmid that does not contain the CRISPR site. So we consider that these genes and elements may be important factors that influence the emergence or evolution of CRISPR on plasmids.

3.7 The effect of MGEs on CRISPR-Cas system of plasmids

According to these results, we think the DinG family helicase, Cas6/Cse3/CasE, Csf2, IS5 have some relationship with the emergence of CRISPR. In order to comprehensively analyze which MGEs and gene can affect the emergence and evolution of CRISPR plasmid type, we collected the full genetic sequences of 10 plasmids that contained CRISPR and 10 plasmids that did not, we annotated these gene sequences and got the result of principal component analysis. PCA results showed that DinG family helicase, Cas6/Cse3/CasE, Csf2, IS5 and plasmid CRISPR had relatively high coefficients (81.2%, 53.3%, 73.8%, 47.1%, respectively).

4. Discussion

In this paper, the distribution, type and spacing sequence of CRISPR-Cas system in *Klebsiella* were studied. Among the collected strains, about one third contained CRISPR-Cas system, and most of the CRISPR-Cas system belonged to type I-E. When the CRISPR locus was analyzed, nine strains were found to have CRISPR gene loci in plasmids. After statistical analysis of the influencing factors of CRISPR-Cas system, plasmids were found to have great influence on its stability.

The CRISPR-Cas system provides bacteria with adaptive immunity against plasmids and other MGEs. In *Klebsiella*, the plasmid specific spacer obtained from the CRISPR spacer of chromosomes can provide immunity to the plasmid for the strain. Huang et al proposed in the study of multi-drug resistant *Klebsiella* pneumoniae that drug-resistant genes could be integrated

from the plasmid to the chromosome by using the CRISPR-Cas system (Huang, et al., 2017). Muhammad and Jonathan also concluded that obtaining new spacer sequences in the CRISPR-Cas array could induce the degradation of its targeted plasmids in the host, prompting the transfer of drug-resistant genes on the plasmids to chromosomes or other related mobile genetic elements under the pressure of antimicrobial selection (Kamruzzaman, et al., 2019). Many strains carry multiple plasmids, and without the selection of plasmids genes, the acquisition of new plasmids reduces the growth rate and competitiveness of the plasmid-carrying host, thus placing a burden on the host [18]. Obtaining plasmid-mediated CRISPR spacers targeting other plasmids and host chromosomes may facilitate the collaborative integration of plasmids with each other or into host chromosomes, thereby improving the stability and compatibility of plasmids (Kamruzzaman, et al., 2019).

The type-IV CRISPR Cas system is equivalent to a simplified version of the type-I CRISPR Cas system, with a genetic makeup similar to that of type-I. However, the Cas protein sequences of type-IV systems are quite different from those of other type-I systems, so they are classified as different systems(Makarova, et al., 2015) . nition in the type- IV system is unclear (Newire, et al., 2019).

Type-IV system has two variants (subtype IV-A and subtype IV-B), both of which contain highly differentiated effector module genes of Cas5 (Csf3), Cas7 (Csf2) and Cas8-like large subunit (Csf1), but subtype IV-A also encodes the DinG family helicase. All of the complete genomes that characterize the type-IV CRISPR Cas system are encoded by bacterial plasmids, bacteriophages, or other uncharacterized integrated elements (Faure, et al., 2019). In addition, some type-IV CRISPR Cas loci encode predictive enzymes of ADP Ribosyl transferase family (ART), including bacterial toxins. Together with the type-IV system's Cas proteins, these enzymes may help suppress the host CRISPR Cas or other defense systems, ensuring the stability of plasmids and prophages(Shabbir, et al., 2016) .

The type-IV CRISPR-Cas system on plasmids lacks target enzymes (Cas3 or Cas10 genes).In the study of Muhammad Kamruzzaman and Jonathan R. Iredell, it was mentioned that the positive plasmids of the CRISPR-Cas system of type-IV *Klebsiella* pneumococci were only found in the bacteria of type I-E chromosome CRISPR-Cas, which made up for the lack of target cutting function in the CRISPR plasmids, considering that there may be a cross between the plasmid and the CRISPR chromosome (Kamruzzaman, et al., 2019). However, in this study, we found that some strains only contained the type-IV CRISPR-Cas system, which may be the result of the continuous evolution of CRISPR under environmental pressure and some MGEs. The impact of MGEs on the CRISPR Cas system occurs in a number of independent situations, including the ability to eliminate interference. But the actual effect of the derived CRISPR-Cas system on plasmids remains to be discovered.

The CRISPR-Cas system was found not only in plasmids, but also in other MGEs. Including phages, Tn7 transposition elements and intergated joint elements (ICEs) (Koonin, et al., 2020). Recruitment of CRISPR-Cas defense systems by different MGEs may have contributed to the evolution of MGEs and defense systems. Some CRISPR adaptation modules (e.g., Cas1, Cas2, Cas4, etc.) are thought to have evolved from different transposons. Transposons are an extensive MGE that can be reproduced by recombinases that insert elements into new locations in the host genome. Translocation is involved in DNA replication, DNA repair, and sometimes reverse transcription(Faure, et al., 2019). Most of the CRISPR-Cas carried by MGEs only retained some of their original functions, and the CRISPR-Cas system was preserved in the evolution of MGEs by inhibiting the host defense to gain an advantage in the conflict with MGEs. There is a complex functional and evolutionary relationship between CRISPR-Cas and MGEs, including the similarity between CRISPR-Cas function and the various nuclease reactions in the life cycle of MGEs(Faure, et al., 2019). Much of the biological information involved needs further exploration and discovery.

5. Conclusions

This study focuses on the CRISPR-Cas system in *Klebsiella* to explore various factors affecting CRISPR and the relationship between CRISPR and mobile genetic elements. The analysis shows that CRISPR interferes with and protects against foreign mobile devices, while some genes and mobile genetic elements may also have significant influence on the emergence and evolution of CRISPR. Explored various types of CRISPR-Cas systems in Klebsiella, which is prevalent worldwide, it's of great significance to research the plasmid - mediated resistance transmission of Klebsiella in the future

Declarations

Competing interests

The authors declare that have no competing interests.

Availability of data and material

All data needed to evaluate the conclusions in this paper are present either in the main text.

Consent for publication

Not applicable.

Authors' contributions

D.Q. conceived the project; Y.Z., W.Z., J.Z. participated in its design; Y.Z., W.Z., Z.H., E.Z., X.Z., S.Z. and S.T. performed the data mining and analyses; and Y.Z. and D.Q. wrote and revised the manuscript with input from all authors. The author(s) read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

References

Barrangou, R., & Horvath, P. (2017). A decade of discovery: CRISPR functions and applications. *Nat Microbiol, 2*, 17092.

Faure, G., Shmakov, S. A., Yan, W. X., et al. (2019). CRISPR-Cas in mobile genetic elements: counter-defence and beyond. *Nat Rev Microbiol, 17*(8), 513-525.

Garcia-Martinez, J., Maldonado, R. D., Guzman, N. M., et al. (2018). The CRISPR conundrum: evolve and maybe die, or survive and risk stagnation. *Microb Cell, 5*(6), 262-268.

Ge, R., Mai, G., Wang, P., et al. (2016). CRISPRdigger: detecting CRISPRs with better direct repeat annotations. *Sci Rep, 6*, 32942.

Huang, W., Wang, G., Sebra, R., et al. (2017). Emergence and Evolution of Multidrug-Resistant Klebsiella pneumoniae with both blaKPC and blaCTX-M Integrated in the Chromosome. *Antimicrob Agents Chemother, 61*(7).

Kamruzzaman, M., & Iredell, J. R. (2019). CRISPR-Cas System in Antibiotic Resistance Plasmids in Klebsiella pneumoniae. *Front Microbiol, 10*, 2934.

Koonin, E. V. (2017). Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol Direct, 12*(1), 5.

Koonin, E. V., Makarova, K. S., Wolf, Y. I., et al. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet, 21*(2), 119-131.

Lino, C. A., Harper, J. C., Carney, J. P., et al. (2018). Delivering CRISPR: a review of the challenges and approaches. *Drug Deliv, 25*(1), 1234-1257.

Makarova, K. S., Haft, D. H., Barrangou, R., et al. (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol, 9*(6), 467-477.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol, 13*(11), 722-736.

Makarova, K. S., Wolf, Y. I., Iranzo, J., et al. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol, 18*(2), 67-83.

Medina-Aparicio, L., Davila, S., Rebollar-Flores, J. E., et al. (2018). The CRISPR-Cas system in Enterobacteriaceae. *Pathog Dis, 76*(1).

Moineau, J. E. S. A. H. M. S. (2015). The CRISPR-Cas immune system and genetic transfersReaching an equilibrium. *Microbiology Spectrum, 3*, 9.

Navon-Venezia, S., Kondratyeva, K., & Carattoli, A. (2017). Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol Rev, 41*(3), 252-275.

Newire, E., Aydin, A., Juma, S., et al. (2019). Identification of a Type IV CRISPR-Cas system located exclusively on IncHI1B/ IncFIB plasmids in Enterobacteriaceae

Ostria-Hernandez, M. L., Sanchez-Vallejo, C. J., Ibarra, J. A., et al. (2015). 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*, 332.

Qu, D., Lu, S., Wang, P., et al. (2019). Analysis of CRISPR/Cas system of Proteus and the factors affected the functional mechanism. *Life Sci, 231*, 116531.

Shabbir, M. A., Hao, H., Shabbir, M. Z., et al. (2016). Bacteria vs. Bacteriophages: Parallel Evolution of Immune Arsenals. *Front Microbiol, 7*, 1292.

Swarts, D. C., Makarova, K., Wang, Y., et al. (2014). The evolutionary journey of Argonaute proteins. *nature structural & molecular biology, 21*(9), 743-753.

Takeuchi, N., Wolf, Y. I., Makarova, K. S., et al. (2012). Nature and intensity of selection pressure on CRISPR-associated genes. *J Bacteriol, 194*(5), 1216-1225.

Tables

Table1. The information of the confirmed CRISPR1~CRISPR7

CRISPR	Type	Repeat sequence (5'-3')	No. of strains	No. of repeats	Frequency (%)
CRISPR1	Typical repeat	GAAACACCCCCACGTGCGTGGGGAAGAC	10	127	57.21
	Repeat variants	AGAAACACCCCCACGTGCGTGGGGAAGAC		12	5.41
		AGAAACACCCCCACGCGTGTGGGGAAGAC		20	9.01
		AGAAACACCCCCACGCATGTGGGGAAGAC		4	1.80
		AGAAACACCCCCACGCGTGTGGGGAAGA		10	4.50
		GAAACACCCCCACGCGTGTGGGGAAGAC		25	11.26
	Terminal repeat	GAAACACCCCCACGCATGTGGGGAAGAC		24	10.81
CRISPR2	Typical repeat	GTCTTCCCCACGCACGTGGGGGTGTTTC	38	405	44.90
	Repeat variants	GTCTTCCCCACACGCGTGGGGGTGTTTCT		166	18.40
		GTCTTCCCCACGCACGTGGGGGTGTTTCT		149	16.52
		GTCTTCCCCACACGCGTGGGGTGTTTCT		7	0.78
		TCCCCACACGCGTGGGGGTGTTTCT		10	1.11
		GTCTTCCCCACACACGTGGGGGTGTTTC		12	1.33
	Terminal repeat	GTCTTCCCCACATGCGTGGGGGTGTTTC		78	8.65
		GTCTTCCCCACACGCGTGGGGGTGTTTC		57	6.32
		GTCTTCCCCACGTGCGTGGGGGTGTTTC		18	2.00
CRISPR3	Typical repeat	GTGTTCCCCGCGCCAGCGGGGATAAACCG	17	612	97.45
	Repeat variants	GTGTTCCCCGCGCTAGCGGGGATAAACCG		6	0.96
		GTGTTCCCCGCGCCAGCGGGGATAAACTGG		10	1.59
	Terminal repeat				
CRISPR4	Typical repeat	GTATTCCCCCCGCATGCGGGGGGTTATCGG	13	79	36.74
	Repeat variants	AGTATTCCCCCCGTGTGCGGGGGGTTATCGG		5	2.33
		CCCCCCGCATGCGGGGGGTTATCGG		19	8.84
	Terminal repeat	GTATGTCCCCGCTGGCGGGGGGTTATCGG		12	5.58
		GTATTCCCCCGCTTGCGGGGGGTTATCGG		29	13.49
		GTATTCCCCCCGTGTGCGGGGGGTTATCGG		71	33.02
CRISPR5	Typical repeat	CCGATAACCCCCGCATGCGGGGGGGAATAC	11	70	45.45
	Repeat variants	CGATAACCCCCGCATGCGGGGGGGAATAC		10	6.49
		CCGATAACCCCCGCACACGGGGG		43	27.92
	Terminal repeat	CCGATAACCCCCGCATGCGGGGGGGAATACT		9	5.84
		CCGATAACCCCCGCATGCGGGGG		11	7.14
		CCGATAACCCCCGCAAGCGGGGGGGAATAC		11	7.14

CRISPR6	Typical repeat	CGGTTTATCCCCGCTGGCGCGGGGAACAC	12	354	93.65
		GGTTTATCCCCGCTGGCGCGGGGAACAC		7	1.85
	Repeat variants	CGGTTTATCCCCGCTCACGCGGGGAACAC		7	1.85
	Terminal repeat	CGGTTTATCCCCGCTAGCGCGGGGAACAC		10	2.65
CRISPR7	Typical repeat	GTTCACTGCCGTACAGGCAGCTTAGAAA	7	44	55.00
		TAAGCTGCCTGTACGGCAGTGAA		4	5.00
	Repeat variants	TTTCTAAGCTGCCTGTACGGCAGTGAACA		4	5.00
	Terminal repeat	ATTCGGCTTGAGAGCCGTTTCCA		4	5.00
		GGGAATAAGTCACTGAAAGTAA		4	5.00
		GGAAACGGCACAAAAGCCGAATCGAAGCAAGTTACTGAAAATAA		4	5.00
		GGAATCGGCGCAAAAGCCGAATCGAAGCAAGTTACTGAAAATAA		4	5.00
		GCATTCGTCCCAAGAGCCGAA		4	5.00
		GGGAATAAGGCATTGAAAGTTA		4	5.00
		GCATTCGTCCCAAGAGCCGAT		4	5.00

Figures
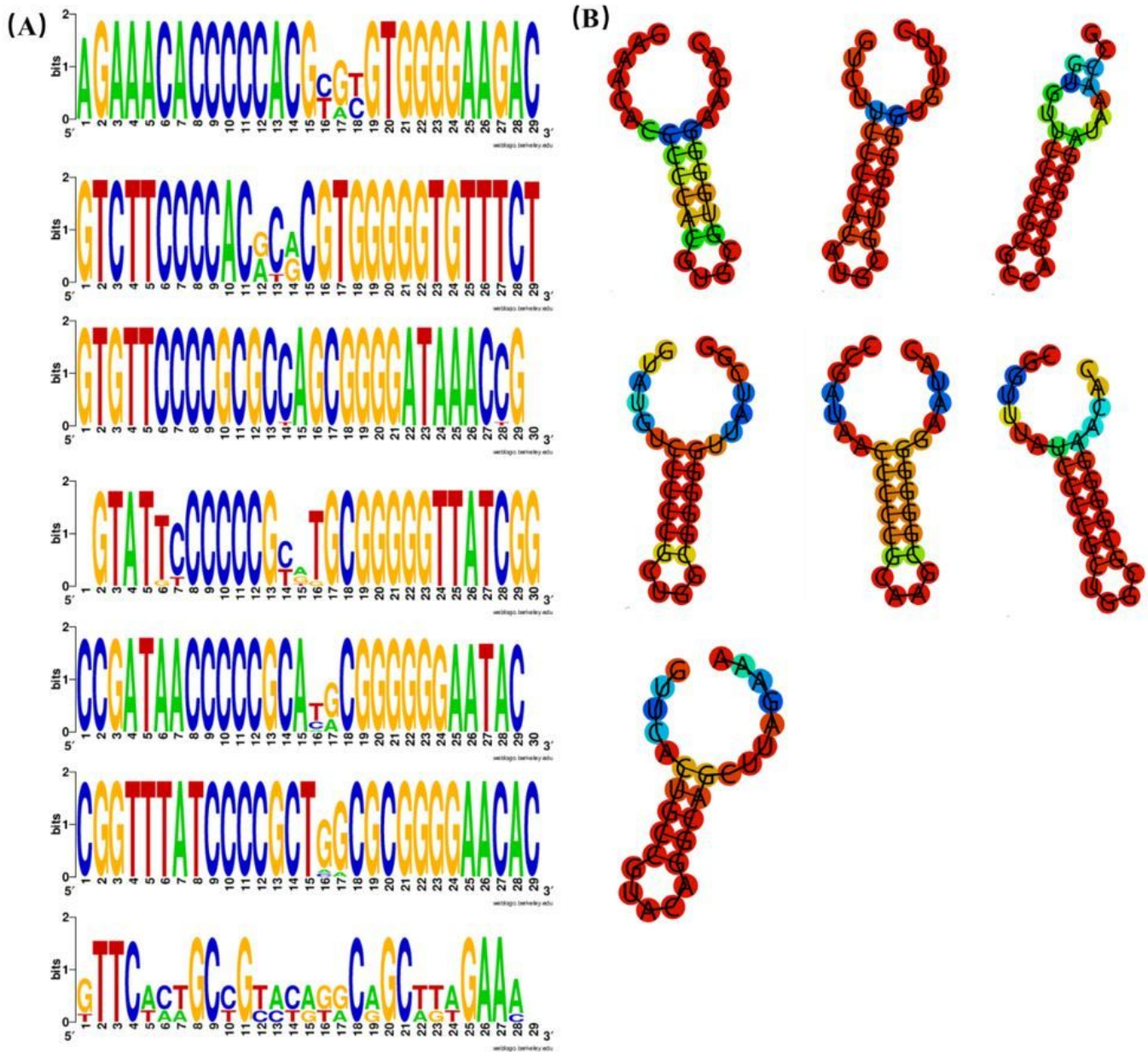
Figure 1

(A) The weblogo of repeats of CRISPR1~CRISPR7. The sequence were first and terminal repeats of CRISPR1~CRISPR7. (B) The secondary structure of repeats of CRISPR1~CRISPR7. 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 -11.60, -11.70, -15.20 -11.80, -12.70, -14.20, -8.60 kcal/mol.
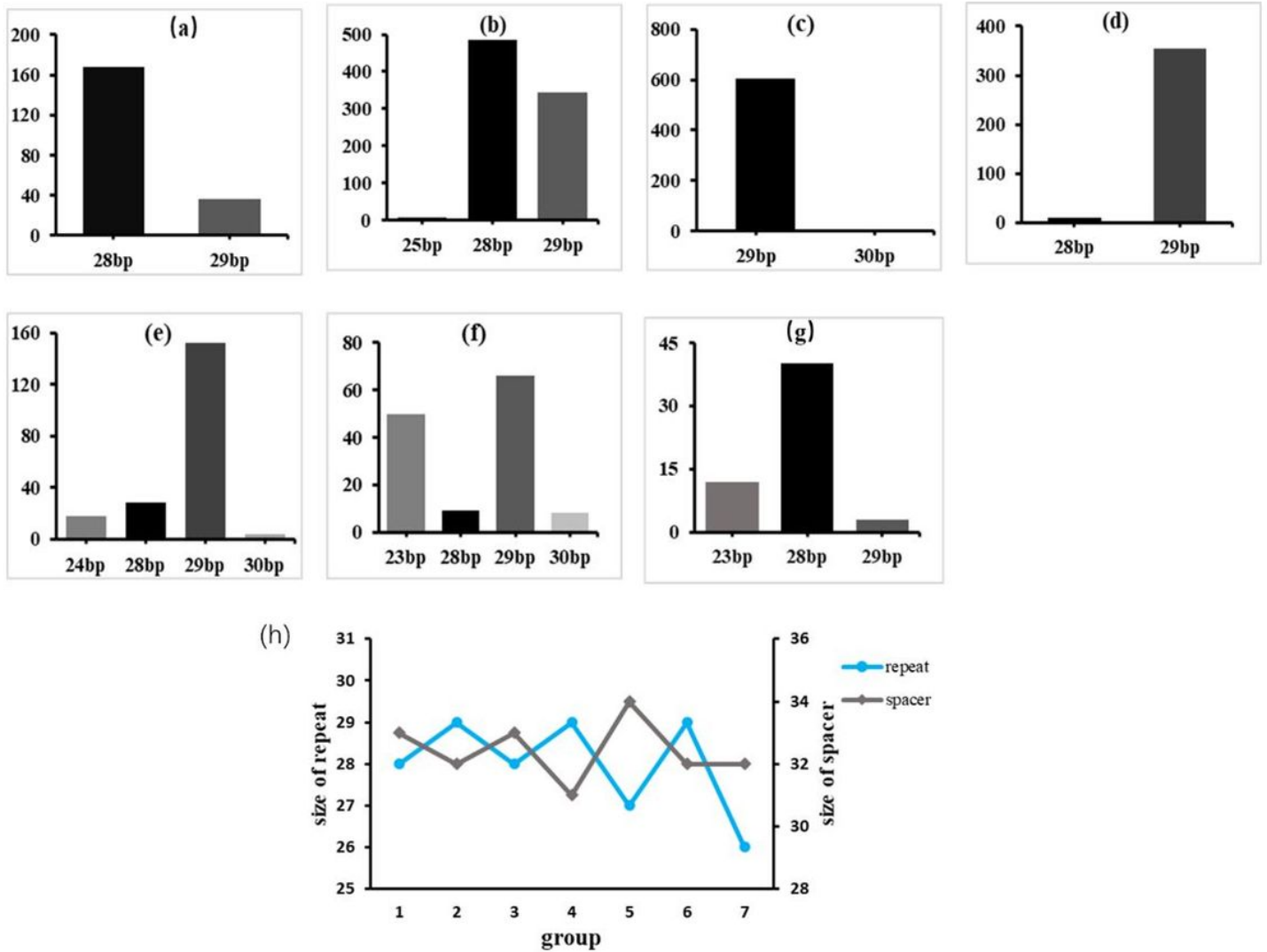
Figure 2

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) Group 7 spacers; (h) 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.
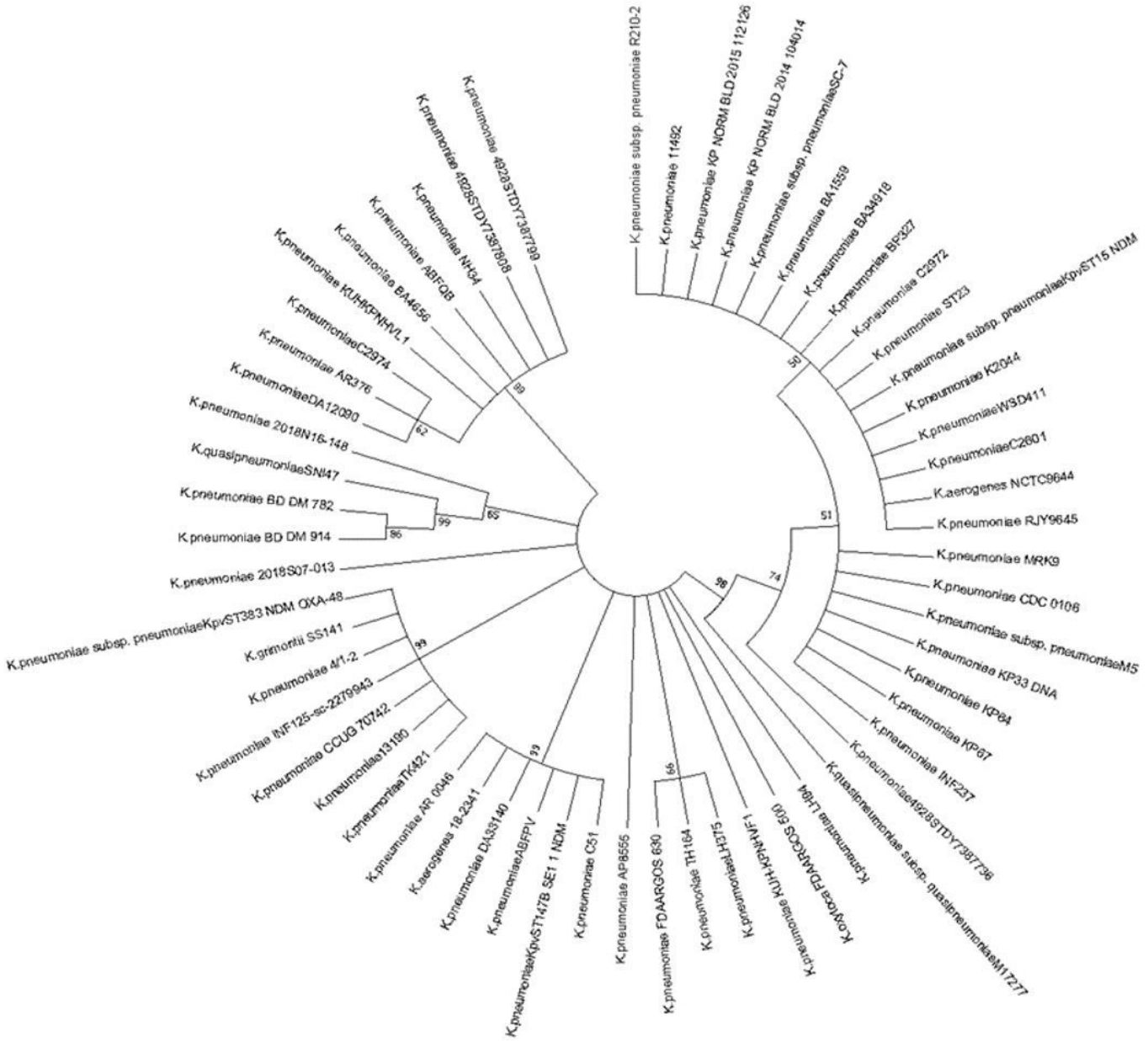
Figure 3

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.
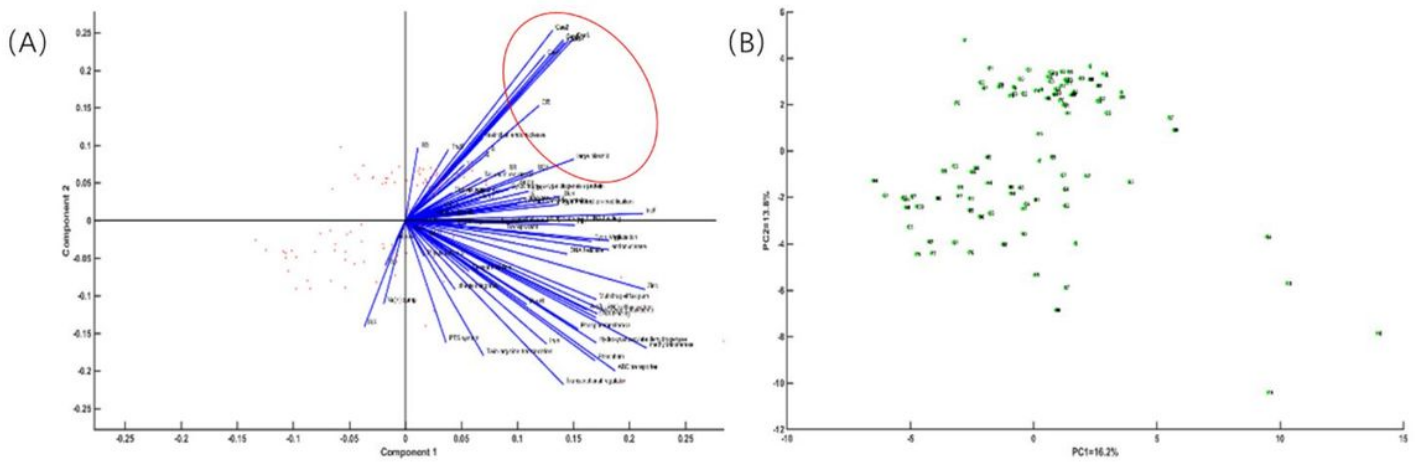
Figure 4

Principal component analysis of CRISPR of Klebsiella in relation to mobile genetic elements, regulators and DNA-related enzymes. (A) Correlation vector diagram of each factor of principal component axis (blue line) and Klebsiella strains (red dot). A vector represents the correlation of each factor with the axis of the first principal component. (B) An enlarged view of the point representing Klebsiella.
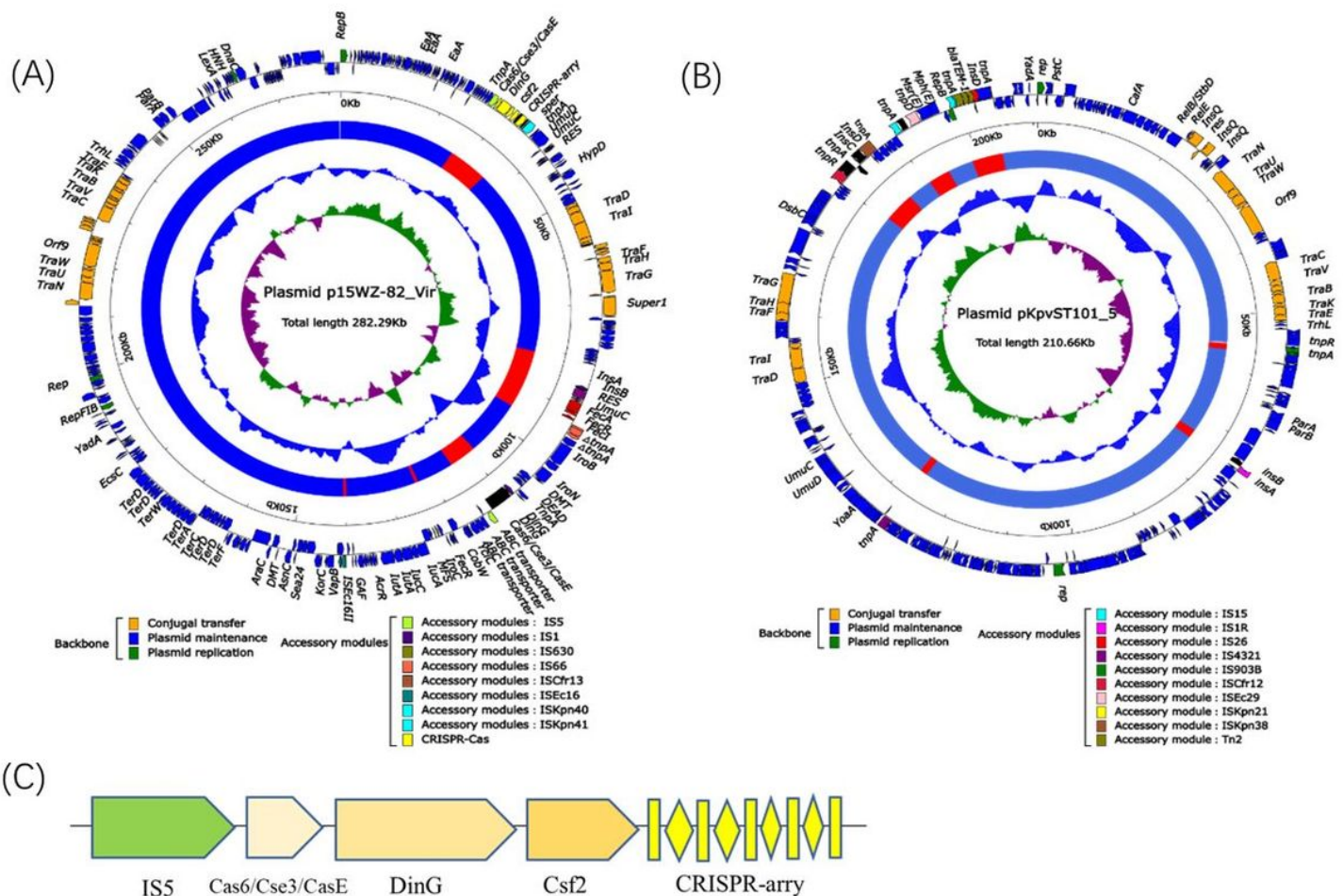

Figure 5

(A) Schematic diagram of the Plasmid p15WZ-82_Vir, (B) Schematic diagram of the Plasmid pKpvST101_5. According to the function of the gene classification, use arrows to represent the gene, and color. (C) Local structure of plasmid CRISPR.
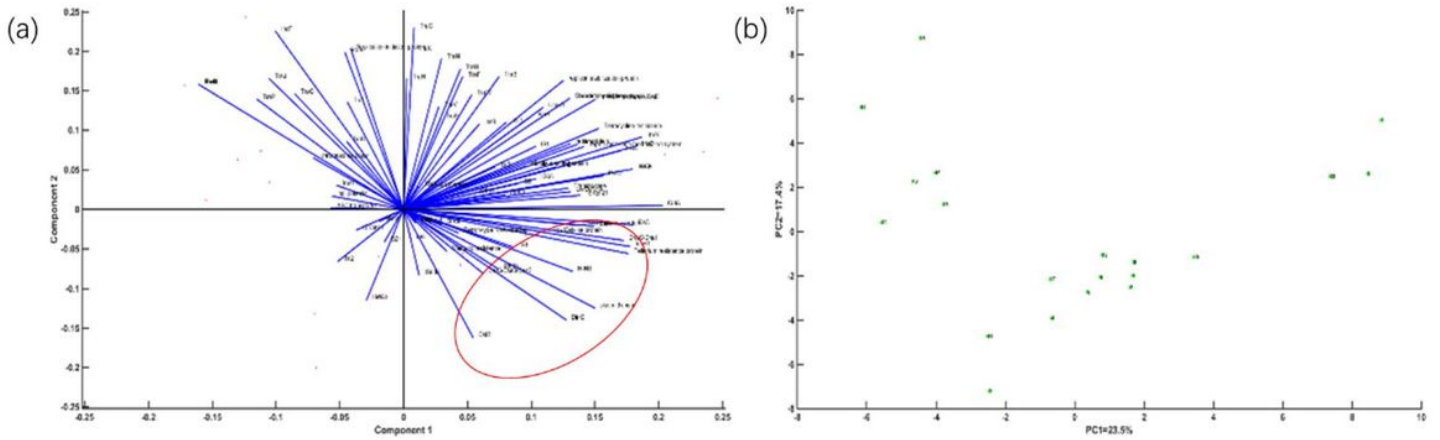

Figure 6

Principal component analysis of the CRISPR plasmid of Klebsiella with mobile genetic elements, regulatory factors and DNA-related enzymes. (A) Correlation vector diagram of principal component axis (blue line) and each factor of Klebsiella strain (red dot). The vector represents the correlation of each factor to the axis of the first principal component. (B) represents an enlarged view of Klebsiella's point.