Occurrence and diversity of CRISPRs in Laribacter hongkongensis isolates from animals, environment and diarrhea patients in southern China

Li Wang
Southern Medical University https://orcid.org/0000-0001-6738-4078
Ling Wang
Southern Medical University
Zhihua Liu
Southern Medical University Nanfang Hospital
Huijie Guo
Southern Medical University
Youzhao Liu
Guangdong Center for Disease Control and Prevention
Zhiyun Wang
Baiyun district Center for Disease Control and Prevention
Qing Chen
Southern Medical University
Jing Hu (hjalzh@smu.edu.cn)

Research article

Keywords: CRISPR, cas, spacers, Laribacter hongkongensis, bacteriophage

Posted Date: May 29th, 2019

DOI: https://doi.org/10.21203/rs.2.9713/v1

Abstract

Background Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (Cas) proteins are functional elements of archaea and bacteria, and they form the genetic barrier that reduces the transformation of horizontal genes by an unknown mechanism. Results we searched for CRISPRs in 118 Laribacter hongkongensis strains isolated from patients, animals, and water reservoirs. Two CRISPR loci, designated CRISPR4.1 and CRISPR4.2, were identified in L. hongkongensis strains. A CRISPR4.1/cas system was detected in 91.5% (108/118) of the isolates and belonged to the I-F/Ypest subtype of CRISPR/cas systems, while the remaining ten strains only possessed cas genes without the CRISPR4.1 array. The CRISPR4.2 locus was an orphan locus and existed in 72.0% (85/118) L. hongkongensis strains. Meanwhile, a total of 2562 spacers and 980 unique spacers, arranged in 77 alleles, were found, including 1613 (579 unique, 40 alleles) for CRISPR4.1 and 949 (401 unique, 37 alleles) for CRISPR4.2. The results showed that limited spacers had matches in the plasmid (34), phage (19) and bacteria chromosomal sequences (4) from GenBank databases. Consequently, we found the diversity and activity of CRISPRs from human and frog isolates were closer and higher, respectively, than those of the fish isolates. Conclusions CRISPR4.1 and CRISPR4.2 exist extensively in L. hongkongensis and belong to the I-F/Ypest subtype. Our results indicate that the CRISPR4.1/cas system is a functional unit, and CRISPR4.2 is perhaps a degenerated and nonfunctional locus. Considering that the CRISPR4.1 locus is relatively active, diverse, and conserved at different levels, the biological functions, types, and epidemiological characteristics of CRISPR loci may be worth studying.

Background

Clustered regularly interspaced short palindromic repeats (CRISPRs) are recently discovered to be common in the genomes of bacteria and archaea, where they provide acquired immunity against foreign DNA [1] . CRISPR loci are usually composed of palindromic direct repeats [2] that are regularly interspaced by flexible sequences called spacers. The CRISPRs are classified into 12 major groups based on repeat sequence similarity and the ability to produce and stabilize RNA secondary structures [3]. CRISPR-associated proteins (*cas*) genes, which encode functional proteins, such as helicases, polymerases, nucleases, and polynucleotide binding, are often adjacent to CRISPR and contribute to the propagation and functioning of CRISPR [4]. They include eight CRISPR/*cas* subtypes, according to gene order as well as gene content. Many spacers of CRISPR derived from proto-spacers are part of foreign DNA [5-8], suggesting that spacers may provide a historical record of mobile element exposure. Spacer sequences show the phenomenon of insertion and selective elimination in the course of bacterial evolution and induce structural polymorphism in CRISPR, which show differences between different strains of the same species [9-11]. Thus, the structure of the CRISPR may be a potential target for tracing the origin and evolution of bacteria.

As a newly discovered species, *Laribacter hongkongensis* (*L. hongkongensis*) is a member of the Neisseriaceae family of β-subclass Proteobacteria, and it is a novel emerging bacterium that is closely associated with community-acquired gastroenteritis and travelers' diarrhea [12, 13]. In 2001, this facultative anaerobic, motile, Gram-negative, nonsporulating, urease-positive, S-shaped bacillus was isolated first from an alcoholic cirrhosis patient in Hong Kong [14] and then obtained from feces samples from gastro-enteritis patients in Asia, Central America, Africa and Europe [15, 16]. The tourist histories of the patients may suggest that *L. hongkongensis* is globally distributed [17-20].

In addition, *L. hongkongensis* has been obtained from diverse environments related to freshwater. It has been widely discovered in the intestinal contents (60%) of universally consumed freshwater fish, Chinese tiger frogs in Hong Kong and water reservoirs in Guangzhou [21]. Freshwater fish and edible frogs are considered two likely major reservoirs of this bacterium and main sources of human infections [22, 23]. In our previous study, apart from patients, freshwater fish and frogs in Guangdong province of China [22, 24, 25], *L. hongkongensis* was also isolated from two sewer rats in Guangzhou city [26]. However, whether *L. hongkongensis* adapts to diverse environmental conditions to protect it in different hosts and habitats remains unknown. The distribution and structural characteristics of the CRISPR/*cas* system of *L. hongkongensis* that may contribute to this adaptation have not been reported.

Here, we investigate the occurrence and distribution of the CRISPRs in 118 *L. hongkongensis* strains from patients, animals, and the environment in southern China, which included 6 available genomes and the 112 *L. hongkongensis* isolates. We further describe the structural features of the CRISPR systems and analyze the activity, diversity, and evolution of CRISPR/*cas* systems in those strains with CRISPRs.

Results

Identification of two CRISPR loci by analysis of 6 available *L. hongkongensis* genomes

First, we analyzed the genome sequence of *L. hongkongensis* LHGZ1 (CP022115.1) and detected two confirmed CRISPR systems, which were distant to each other (>344 kbp) (**Fig. 1**). Both of them exhibited several parallel direct repeats (GTTCACTGCCGGACAGGCAGCTCAGAAA), which could be classified into cluster 4 on the basis of sequence identity[3]. We named them CRISPR4.1 and CRISPR4.2, respectively. For those two specific genomic structures, we only confirmed the existence of putative *cas* genes in the flanking sequence from the CRISPR4.1 arrays, which consisted of the successive co-oriented *cas* genes (*cas1-cas3-csy1-csy2-csy3-csy4*) and were located between the LHGZ1-1184 (*rve_3,* rve_3 domain-containing protein) gene and the LHGZ1-1195 (*sts,* sulfate transporter subunit) gene. Furthermore, downstream of the *csy4* gene, a leader sequence of 170 bp, which was rich in AT (52.33%), was present in the CRISPR4.1 system of *L. hongkongensis* LHGZ1. The CRISPR4.1 system of the *L. hongkongensis* LHGZ1 had 31 spacers, approximately 32 bp (32–34 bp) in length, and almost all repeats (28 bp in length) of the system were identical, though there were several variations. The leader, the spacer sequences, and the repeat sequences are given in **Fig. 1**.

We then analyzed the other 5 (GZ-NF250, HLHK9, JM-679, GZ-R251, GZ-Y19) available genome sequences of the *L. hongkongensis* strains for the presence of a CRISPR4.1/*cas* system. Interestingly, four of the investigated *L. hongkongensis* strains were positive for CRISPR4.1, strain HLHK9 being the exception. Strain HLHK9 had the majority of *cas* genes that LHGZ1 had but no direct repeat or spacer sequence (**Fig. 1**). The order of *cas* genes was *csy1-csy2-csy3-csy4'-csy4-cas1'*. Among these *cas* genes, *cas3* was deleted, and *cas1'* had only 111-bp, which were not identical to any 111-bp stretch of *cas1*. Surprisingly, the *csy4'* gene and *csy4* gene were divided by a 9 bp stretch, and *csy4'* only had 165 bp identical to *csy4.* Nevertheless, the leader sequence was still present and similar (94% DNA identity) to that of strain LHGZ1, although it did not exhibit any direct repeat or spacer DNAs. The last direct repeat was located next to an *sts* gene homologous to strain LHGZ1 in the remaining 4 investigated *L. hongkongensis* isolates (**Fig. 1**). All four showed a parallel *cas* gene group with 90%-98% DNA similarity to LHGZ1, and the direct repeat was also similar to the LHGZ1 strain (**Fig. 1**). Leader sequences (93%-97% DNA similarity to LHGZ1) were also found in the four strains. The results in **Fig. 1** indicate that all strains exhibited disparate numbers of spacer sequences.

As shown in **Fig. 1**, the CRISPR4.2 locus of *L. hongkongensis* LHGZ1 existed downstream of a gene (LHGZ1-1563) encoding a Phage_TAC_7 domain-containing protein (Phage_TAC_7) and upstream of a gene (LHGZ1-1567) encoding phage tail protein (Pht.p). In contrast to the CRISPR4.1/*cas* system, CRISPR4.2 was an orphan locus, only consisting of repeats and spacers, with no *cas* genes and no leader sequence, though it shared identical consensus repeats with the CRISPR4.1/*cas* system. Fourteen spacers of approximately 32 bp (32-33 bp) were present. In CRISPR4.2, in contrast to CRISPR4.1, the last seven bases of the repeats were variable. Of the other 5 available genome sequences of *L. hongkongensis*, strains GZ-NF250, GZ-R251 and GZ-Y19 exhibited CRISPR4.2 loci, with identical repeats (except the last repeat) and different spacers. Strains JM-679 and HLHK9 did not exhibit CRISPR4.2 loci.

The distribution of the two CRISPR loci in the 118 *L. hongkongensis* strains

To confirm the presence of the two CRISPR loci among various sources of *L. hongkongensis*, we amplified and sequenced the *rve_3* and *sts* regions of 112 *L. hongkongensis* isolates for detection of CRISPR4.1, as well as Phage_TAC_7-Pht.p region for CRISPR4.2 **(Table 2)**. Table S1 summarizes the CRISPR4.1/*cas* systems of all 118 *L. hongkongensis* strains. Of the 42 frog isolates, 37 were CRISPR4.1+/CAS-Y+, and five were CRISPR4.1-/CAS-Y+, with the units varying from five to 39. Of the 66 fish isolates, 63 had repeats (CRISPR4.1+) accompanied by Cas-Y genes (CAS-Y+), while three were CRISPR4.1-/CAS-Y+, and the number of units varied from two to 25. Of the seven human isolates, five were CRISPR4.1+/CAS-Y+, two were CRISPR4.1-/CAS-Y+, and the units ranged from 11 to 33. The two sewer rat isolates (GZ-R398, GZ-R251) and one water reservoir isolate (GZ-SHUI) all were CRISPR4.1+/CAS-Y+, with the units being two, 17, and 12, respectively. According to the chi-square test, there was no statistically significant difference in CRISPR distribution between fish-origin, frog-origin and human-origin strains (p=0.092). Due to the limited number of strains of sewer rat and water reservoir, we did no statistical analysis on them.

PCR amplification of the Phage_TAC_7-Pht.p region for all 112 *L. hongkongensis* isolates and analysis of the six available *L. hongkongensis* genomes revealed that 85 had CRISPR4.2, and 33 lacked a CRISPR in this region. Of the 42 frog isolates, 66 fish isolates, seven human isolates, two rat isolates and one water reservoir isolate, 31, 47, four, two, and one harbored repeat-spacer arrays, respectively. The repeat content of each strain was also quite different: frog isolates, fish isolates, human isolates, rat

isolates varied from four to 37, four to 15, 15 to 23, five to six units, respectively, and the one water reservoir isolate had five units. There was no statistically significant difference in CRISPR distribution between fish-origin, frog-origin and human-origin strains according to the chi-square test (p=0.068). Further statistics on the distribution of CRISPR4.1 and CRISPR4.2 showed that the positive rate of CRISPR4.1 (91.53%) was higher than that of CRISPR4.2 sites (72.03%, $p < 0.001$) (**Table 2**).

CRISPR repeat sequences are locus-specific

In our study, Ypest *cas* subtypes corresponded strictly to our structured repeat cluster 4, showing consistency with the previous observation that the classification of the typical CRISPR repeats (the most frequent repeat) is consistent with that of the Cas proteins. In general, the two CRISPR repeats had an identical and relatively conserved sequence of 28 bp, whereas we could observe the variations within the repeat sequences throughout a CRISPR locus. Interestingly, there were three typical CRISPR repeats in the CRISPR4.1 locus (**Table 3**). Most strains had the first two typical CRISPR repeats, only two strains invariably starting with A instead of G. Two variants had one mismatch in the 3' end of the CRISPR4.1 repeats. In CRISPR4.2, the last direct repeat differed from the others in the last seven bases, but was similar among all the different strains analyzed. Sequence degeneration existed at the 3' end of the terminal repeat. The percentage of atypical repeats was 0.17% for CRISPR4.1 and 8.22% for CRISPR4.2, suggesting that CRISPR4.2 repeat sequence degeneration was much higher than CRISPR4.1, while CRISPR4.1 appeared to have highly conserved repeat sequences. Although the existence and number of CRISPR loci were fairly regular, the sequence of repeats in each locus was extremely variable in those strains. In our study, two phenomena were found: either the number of repeats was extremely low (only two) or it was relatively high (up to 39).

Spacer diversity in *L. hongkongensis* CRISPR loci

Spacers separate consecutive repeats and comprise the diversity of CRISPRs. In CRISPR4.1, 1613 spacers, including 40 combinations, acting as alleles, existed in the 118 *L. hongkongensis* strains with repeats, resulting in 37.03% diversity. Some 579 (35.90%) different spacer sequences (unique spacers) existed, and the mean number of spacers was 14.94±7.45. For CRISPR4.2, 949 spacers spread over 37 alleles, resulting in 43.53% diversity, and 401 (42.26%) unique spacers were found. The number of spacers, the mean number of spacers, the number of unique spacers and the number of alleles among different-origin isolates are shown in **Table 4**, **Table 5**.

Among the two CRISPR spacers, (i) the mean number of spacers of CRISPR4.1 was higher than that of CRISPR4.2 (p=0.000); (ii) the unique-spacer proportion of CRISPR4.1 was lower than that of CRISPR4.2 (p=0.001); (iii) there was no statistically significant difference between the number of alleles in CRISPR4.1 and that in CRISPR4.2 (p=0.361), as shown in **Table 6**.

Interestingly, a further comparative analysis of the two CRISPRs' spacers of isolated strains from different sources revealed that in CRISPR4.1 (**Table 7**), (i) the average number of spacers between human-origin and frog-origin strains was higher than that of fish-origin strains (all $p < 0.05$), but there was no statistically significant difference between frog-origin and human-origin strains (p=0.176); (ii) the number of unique spacers in the fish-origin, frog-origin and human-origin strains was 87 (11.73%), 410 (56.63%) and 115 (97.46%), respectively. According to the chi-square test, the distribution difference of unique spacers in different-origin strains was statistically significant (all $p < 0.0125$), indicating that the human-origin strains had the largest proportion of unique spacers, followed by frog-origin strains; and (iii) the number of alleles in both human-origin and frog-origin strains was greater than that in fish-origin strains (all $p < 0.0125$). In CRISPR4.2 (**Table 8**), (i) the average number of spacers of human-origin strains was higher than that of fish-origin strains ($p < 0.05$), but there was no statistically significant difference between frog-origin strains and fish-origin or human-origin strains ($p = 0.179$); (ii) the number of unique spacers in the fish-origin, frog-origin and human-origin strains was 55 (12.70%), 304 (70.86%) and 72 (98.63%), respectively. The chi-square test showed that the distribution difference of unique spacers was statistically significant in the three different-origin strains, (all $p < 0.0125$), indicating that unique spacers in human-origin strains accounted for the largest proportion, followed by frog; and (iii) the number of alleles of both frog-origin and human-origin strains was higher than that of fish-origin strains ($p < 0.0125$). The spacers of the two loci from the reservoir water strain was consistent with that of the fish-origin strain.

Taken together, 980 unique spacers existed in the two CRISPR loci. Over half of the unique spacers (349 out of 579 in CRISPR4.1, 286 out of 401 in CRISPR4.2) were strain-specific (white boxes), i.e., present in only one strain (**Fig. 2**). We found that these sequences were always situated at the terminal leader of the CRISPR, with only limited sequences located in the inner positions (**Fig. 2**). These data suggest the acquisition of a new structure of spacers, with the new component being inserted between the end of the leader sequence and the start of the direct repeat, specifically for strains LHGZ1, SZ-W64, SZ-W71, SZ-W66, SZ-W48, etc. Meanwhile,

we found the most identical conserved spacers on the other end of the CRISPR (**Fig. 2**). In some situations, the spacers, which were located at the end of the locus, were normal to different-origin strains that may have had the same ancestor. For example, CRISPR4.1 from strains SZ-W51, SZ-W54, SZ-W59, SZ-W45, and SZ-W68 may have developed from the same ancestor (**Fig. 2A**). In addition, CRISPR4.2 of SZ-Y49c, SZ-Y27d, JM-Y10, GZ-Y8, and JM-Y16 strains possibly came from a parental strain and likely differentiated after the completion of the 12 spacers they share (**Fig. 2B**). As above, the most multiplicity of CRISPR4.1 spacers was proximal in the leader sequence (left). However, spacers located away from the leader sequence were the most frequently seen (**Fig. 2A**). Similar to CRISPR4.1, the highest diversity of CRISPR4.2 spacers also existed on the left, suggesting that there previously was a leader region on the spacers' right. Within a special CRISPR locus, besides the difference coming from the existence of multiply spacers, we also found possible deletions in some spacers. Some of the strain clusters are listed in **Figure 2**, and it looks like deletions have happened, resulting in the deletion of the inner spacers, and the remaining spacers have moved further toward the trailer at the end of the locus sequence.

Analysis of CRISPR spacer sequences

Referring to spacers, here we label spacers as "CRISPR array No." - "spacer No." (**Table 9**). The spacers of CRISPR are considered to have originated from mobile genetic elements, which are named proto-spacers [5, 27]. To elucidate the ecological function of the CRISPR, we sought proto-spacers of the CRISPR in the NCBI database (methods mentioned above). Despite the abundant availability of phage (virus), plasmid and bacteria sequences, we only matched proto-spacers for 18 unique spacers matched 53 mobile elements and four nonmobile elements among 980 unique spacers, including 34 plasmids, 19 phages, and four bacteria (**Table 9**). Interestingly, among plasmid proto-spacers, half of them (17) corresponded to *Xylella fastidiosa* strain plasmids, while the phage proto-spacers were diverse. It was also notable that although the general partiality to plasmids was obvious for CRISPR4.1 (30 out of 37), the CRISPR4.2 proto-spacers had a different prevalence, which was a trend toward phages (12 out of 16). Among the matched spacers, most spacers were associated with several, even ten, known sequences (SZ-W33-4.1-6) of plasmids and phages. Meanwhile, we also found that some different spacers from different strains could match the same mobile element. Indeed, SZ-W33-4.1-6 and SZ-W57-4.1-5/6 proto-spacers were found in the same plasmids, such as *Xylella fastidiosa* 9a5c plasmid pXF51 and *Xylella fastidiosa* strain Pr8x plasmid pXF39. This suggested that these strains have become resistant to the same phages and plasmids and that these plasmids and phages are widespread in *L. hongkongensis* strains.

Analysis of encoded products of proto-spacers

After the analysis of the encoded product of the proto-spacer (**Table 9**), we summarized the coding products of the spacer matching gene, usually being a variety of protein families, including (i) some coding products were necessary for the survival and growth of microorganisms, such as "DNA polymerase", "capsid maturation protease", "secretion system protein TraC", "DNA repair protein" and "DNA-binding protein"; some were involved in adapting to the environment, such as "conjugal transfer protein TraG". (ii) Some of the genetic coding products were functional proteins, for example, "ATPase", "DNA cytosine methyl transferase". (iii) Some of the coding products were "hypothetical protein" with unknown functions. There were also many spacer coding products that were currently not put into categories, including "hypothetical protein," which need to be further analyzed and studied. Finally, there were some genetic fragments that were noncoding regions or pseudogenes, which also require further study.

Discussion

The CRISPR/*cas* system has been described in many different species of bacteria [28, 29]. The CRISPR/*cas* system protects prokaryotes against invading plasmids and viruses and provides a historical perspective of foreign genetic element exposure [29, 30]. No study about the CRISPR/*cas* system in *L. hongkongensis* has been reported. In our study, we try to provide the first description of the structure and potential function of CRISPR/*cas* systems in 118 *L. hongkongensis* strains.

The I-F CRISPR type (CRISPR4.1 and CRISPR4.2) existed in most of the *L. hongkongensis* strains. Between the two CRISPR loci present in *L. hongkongensis* strains, diversity was observed at many levels, including 108 isolates that contained a CRISPR4.1/*cas* system, which included repeat genes and complete *cas*. The number of repeats for CRISPR4.1 varied from 2 to 39, and it was higher than other bacteria and infrequent in enterobacteria [28, 31], indicating that the CRISPR4.1/*cas* system was functional. Yet 10 isolates were CRISPR4.1-/CAS-Y+. The reason may be similar to that for the HLHK9 strain, incomplete *cas* genes resulting in lower *cas* activity in those strains. For CRISPR4.2, although there were no *cas* genes around the locus and the number of repeats was less than CRISPR4.1, 85 of 118 strains possessed this CRISPR, and the repeats were also very abundant. The explanation may be that

CRISPR4.2 chooses suitable conditions for the host to stabilize the locus against the external environment. The two loci (CRISPR4.1 and CRISPR4.2) were distant to each other (344 kbp), so they were functionally independent of each other according a previous study, suggesting that the two loci are compatible, that is, they can exist simultaneously [29]. Previous data have suggested that the enzymatic machinery of a specific locus cannot be effective in conjunction with the CRISPR genetic content of another [29]. Here, we provide data indicating that each *cas* system may be directly linked to a particular CRISPR repeat sequence. Considering the CRISPR4.2 without *cas* genes around, we speculate that CRISPR4.2 is a degenerate locus.

The two CRISPR loci shared identical repeats, which were classified into cluster 4. Our data show that the *cas* gene group was possibly directly connected with a peculiar CRISPR repeat, which is concurrent with the CRISPR structures and *cas* genes demonstrated by Kunin et al [3]. Sequence alignment between CRISPR loci showed that, although the repeats were generally highly conserved around the locus, polymorphism still existed. In the CRISPR4.2 locus, the last repeats were distinct from the others in the last seven bases but were alike in all the strains, which also existed in *Francisella tularensis*. This was unexpected because it is accepted that each repeat shares a structure with the previous repeat [32]. We speculate that there may be some sort of joint point between the repeat sequence and the spacer sequence. The marked difference between the first and last repeats suggests that these repeats have something to do with the evolution of the bacteria [28]. This observation is consistent with our assumption that bacterial evolution is accompanied by base alterations in CRISPR repeats.

For *cas* genes, there was no definite relationship between *cas* genes and the source, amount or host identity of repeats, since CRISPR4.1 was always accompanied by the presence of *cas* genes. Some Cas proteins might be required for the novel repeat-spacer entity unit to interact with molecules by CRISPR repeats. Other Cas proteins are likely related to spacer-encoded resistance, which may be modulated with an RNA-guided mechanism [33]. Further studies surveying the molecular mechanism of the role of CRISPRs and the functional connection between special *cas* genes and a peculiar CRISPR repeat should be done.

Spacers flank consecutive repeats and constitute the most diverse part of CRISPR. The spacer DNAs have been studied previously in many species [34, 35]. Our study showed that when corresponding CRISPR loci in other strains were compared, the spacer sequences were abundant in *L. hongkongensis* strains (**Table 5**), showing the polymorphism and evolutionary variability of the system. In total, 2562 spacers were found in the two CRISPR loci. A total of 579 and 401 unique spacers, arranged in 40 and 37 alleles, were found for CRISPR4.1 and CRISPR4.2, respectively. These spacers were also arranged in a number of alleles in isolates of different origin (**Table 5**). These results demonstrated that less than half of the CRISPR array was specific to most strains. The degree of spacer polymorphism, in terms of both total number of unique spacers and the total number of unique spacer arrangements, for a given CRISPR locus was directly correlated with its activity. Thus, CRISPR4.1 was more active than CRISPR4.2. This was based on several results: (i) the CRISPR4.2 repeat sequences were more degenerated than CRISPR4.1; (ii) the average number and maximum number of CRISPR4.1 spacers were higher than CRISPR4.2 spacers; (iii) *cas* genes accompanied CRISPR4.1, while CRISPR4.2 was an orphan locus. Moreover, our data indicate that the CRISPR polymorphism of human isolates and frog isolates was more closely related and more extensive than that of fish isolates, and the reasons are similar as the above-mentioned: (i) the numbers of spacers for human isolates and frog isolates were higher than fish isolates for CRISPR4.1 and CRISPR4.2, suggesting there was more genetic diversity of *L. hongkongensis* isolates of frog origin and human origin than those of fish origin, which is consistent with previous studies based on PFGE and MLST [22, 24]; (ii) there was no significant difference between frog isolates and human isolates in the quantitative distribution of spacers in the two CRISPR loci; (iii) the ratio of unique spacer sequences in fish isolates was the lowest, while human isolates had the highest, in the two CRISPR loci; (iv) the ratio of alleles of fish isolates was the lowest, meaning that the spacers of fish strains were relatively highly homogeneous. Although the number of the isolates from other origins except for fish and frogs was relatively small, the diversity of the isolates from other host origins was also observed, each having its specific allele. Overall, the CRISPR loci were complex in isolates of the three origins, the highest level of polymorphism being found in the human isolates and frog isolates, followed by fish isolates. It is possible to speculate that frogs are more closely related to humans, and their clones could be better adapted to human hosts.
Our study supports that the spacers were added into the CRISPR locus adjoining the leader sequence, resulting in more diversity than the sequences at the end of CRISPR. This meant that spacer sequences were chronological records that reflected previous encounters with foreign genetic elements. Although continual insertion of new spacers responds to foreign genetic elements, the stability of the bacterial genome always keeps relatively stable, since the deletion of original and valueless spacers, notably the terminal spacers, which possibly came from previous events, accompanied the entry of a novel spacer (**Fig. 2**). Furthermore, we speculated that recently acquired spacers might be important and have more opportunities to be kept in the environment to target

foreign genetic sequences. In addition, we have shown in strains JM-679, JM-W2, JM-W8, and GZ-W2 that CRISPR4.1 contained, respectively 4, 2, 2, and 4 of the same spacers in themselves, which likely involved responses to foreign genetic elements, leading to the same acquisition of spacers. In a word, CRISPR loci seemed to evolve with the process of gain and loss of repeat-spacer units.

In general, the spacer DNA is very specific for each strain and is therefore used for epidemiological genotyping [1, 36, 37]. The analysis of the spacer DNAs showed that matches in CRISPR4.1 (41) were stronger than in CRISPR4.2 (16), but in both, few (only 18) spacers had matches, including plasmids (34), phages (19), bacteria (4) (**Table 9**). The low degeneration of the older spacer sequences could explain the lower level of matches, which was found in the recently obtained spacers. Consequently, according to our results, the number of spacers for most strains that had matches were more than the average number of spacers, suggesting the acquisition of spacers from foreign elements. Beyond that, many CRISPR spacers matched the HLHK9 and LHGZ1 genomes (data not shown), implying acquisition of spacers from their own genomes. Many unique spacers did not match any genome, indicating there are many underexploited plasmids, phages, and bacteria, or else that the highly efficient evolution of plasmids and phages leads to proto-spacer base mutation, giving rise to recognition escape by spacers. Interestingly, our data also strongly suggest that CRISPR4.1 targeted mostly plasmids (30 out of 37), whereas CRISPR4.2 targeted phages (12 out of 16). Meanwhile, among the plasmids, most spacers corresponded to *Xylella fastidiosa* strain plasmids, while the phages were diverse. In the matched spacers, most spacers were related to ten reported sequences of phage and plasmids. Thus, those CRISPR spacers may offer wide-ranging protection against phages and plasmids, which could imply that these strains are one of the few strains conferring multiple drug resistance, as plasmids often carry drug resistance genes [38]. The decreasing diversity and amount of foreign elements in these environments may be the reason for the shortage of CRISPR4.2 activity.

Conclusion

In summary, two CRISPR loci, CRISPR4.1 and CRISPR4.2, exist extensively in *L. hongkongensis*, and the CRISPR4.1/*cas* system belongs to the I-F/Ypest subtype. The repeats in both CRISPR loci show conservation and variation. The results of spacer diversity and analysis of spacer sequences provide important information for elucidating the acquisition of spacers and the mechanism of the CRISPR/*cas* system. This study shows the diversity of CRISPR/*cas* and spacer characteristics, and CRISPR loci seem to have evolved by a process of addition and loss of repeat-spacer units in *L. hongkongensis* strains. Additionally, the diversity and activity of human-origin and frog-origin CRISPRs are more similar and higher than those of the fish isolates. Our results also indicate that the CRISPR4.1/*cas* system is a functional unit, and CRISPR4.2 perhaps is a degenerated and nonfunctional locus. Considering the higher spacer diversity and structure of the CRISPR4.1 array in *L. hongkongensis*, the potential function of a particular CRISPR locus for genotyping, epidemiological and evolution studies may warrant further investigation.

Methods

L. hongkongensis bacterial strains

A total of 118 *L. hongkongensis* isolates were characterized in this study. There were 66 strains from freshwater fish (65 grass carp and one bighead carp), 42 strains from frogs (one wild tiger frog, one giant spiny frog and 40 raising tiger frogs), seven strains from humans, two strains from sewer rats, and one strain from the water reservoir. The 66 fish strains and 42 frog strains were recovered previously from the guts of these animals randomly sampled from retail food markets in Shenzhen (20 fish and 34 frogs), Guangzhou (31 fish and 6 frogs), and Jiangmen (15 fish and 2 frogs), Guangdong province. Seven human *L. hongkongensis* isolates were all obtained from the feces of patients, including two strains from Guangzhou, two strains from Jiangmen, two strains from Hong Kong, and one strain from Hangzhou, Zhejiang province. The two sewer rat isolates were collected from the intestinal contents of sewer rats in Guangzhou. Finally, one strain from a water sample was recovered from Helong water reservoir in Guangzhou. All of the isolates were identified according to standard biochemical procedures and were analyzed by using a 16S rRNA gene-based PCR assay [15]. Biochemical tests of potential *L. hongkongensis* isolates showed that they reacted with arginine, catalase, cytochrome oxidase, dihydrolase, and urease and were hardly positive for sugar oxidation-fermentation [17]. These isolation details are summarized in additional file 1.

Data

Complete genome data of two *L. hongkongensis* strains (HLHK9, LHGZ1) available through the NCBI database (CP022115.1, NC_012559.1) and draft genome data of four strains, including JM-679 (MG209232), GZ-NF250 (MG209226, MG209227), GZ-Y19 (MG209230, MG209231), and GZ-R251 (MG209228, MG209229), were obtained by using an Illumina HiSeq Sequencing System (San Diego, California, America, Illumina) [39].

PCR amplification

DNA extracts were obtained by a rapid boiling method from the 112 confirmed *L. hongkongensis* isolates except the above six available complete genomes. These extracts were used as the templates for PCR amplification. The CRISPR and *cas* PCR primers in this study were synthesized following the principle of the alignment of the conserved sequences, which were found in the GenBank nucleotide sequence database according to the abovementioned accession numbers. Then, we obtained the primers as shown in Table 1. Primers of the 16S rRNA gene were used in positive-control reactions for each colony lysate. The cycling conditions of the two CRISPR arrays were (i) initial denaturation step of 5 min at 94°C; (ii) 32 cycles, with 1 cycle of 94°C for 1 min, primer annealing at 57°C for 1 min for CRISPR4.1 and 60°C for 1 min for CRISPR4.2, extension at 72°C for 1 min/kb PCR target; and (iii) a final extension of 10 min at 72°C. PCRs for *cas* genes were conducted with the steps (i) initial denaturation step of 5 min at 95°C; (ii) 32 cycles, with 1 cycle composed of denaturation at 94°C for 30 s and primer annealing at 56.5°C for 30 s (the specific primer annealing temperature is shown in Table 1); extension at 72°C for 1 min/kb for 30 s for the PCR target gene; and (iii) a final extension of 10 min at 72°C. The amplified PCR products were electrophoresed in agarose gels, stained with ethidium bromide and photographed under short-wavelength ultraviolet light.

DNA Sequencing and analysis

PCR products were applied with the Universal DNA PCR purification kit (Tiangen), and then we submitted the PCR products to Sangon Biotech Co., Ltd. (Shanghai, China) for DNA Sequencing. CRISPR arrays were manually detected in 118 *L. hongkongensis* strains using the CRISPR database (2) and CRISPRfinder [40] (http://crispr.u-psud.fr/Server/CRISPRfinder.php). The BLASTN program [41] was performed on the website of the NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) for sequence similarity analyses and comparisons to public sequences. ClustalX was used for multiple sequence alignments. We regarded two spacers with at least 90% identity in sequence length as the same spacer [42]. Proto-spacers were regarded as the structures that existed outside of a CRISPR locus, with the occurrence of at least 28 identities (only hits with > 90% identity) with spacers [38]. In this study, we applied CRISPRTarget and BLASTN running against the NCBI-NR database to analyze the spacers by using the parameters that the automatic application sets for short queries. The CRISPRTarget databases included GenBank-Phage, RefSeq-Plasmid, RefSeq-Microbial and RefSeq-viral, and the cutoff score was set to 24 [43]. Self-matches of *L. hongkongensis* CRISPR loci sequences, which were stored in BLAST and CRISPRTarget, were ruled out.

Statistical analyses

Data analysis was performed with Statistical Package for the Social Sciences (SPSS; Version 20.0), and $p < 0.05$ was regarded as statistically significant. The number of spacers of *L. hongkongensis* isolates were compared between two CRISPR loci using the independent-samples *t* test. Comparisons among more than two groups were performed using analysis of variance (ANOVA) and a post hoc test. Comparisons of the distribution of unique spacers and CRISPR alleles of the *L. hongkongensis* isolates from fish, frogs, and human for the two CRISPR loci were applied using Pearson's chi-square test.

Abbreviations

CRISPR: Clustered regularly inter*sp*aced short palindromic repeat; Cas: CRISPR-associated proteins; *L. hongkongensis*: *Laribacter hongkongensis*; MLST: Multi-locus sequence typing; PFGE: Pulsed Field Gel Electrophoresis.

Declarations

Funding

This work was supported by the National Natural Science Foundation of China [grant no. 81373052]. We are grateful to Kwok-yung Yuen, Woo Patrick C Y and Xiao-ping Ni for the generous gift of the three patient-derived strains of *L. hongkongensis* (HKU1, HLHK9, HZ-242), which served as reference strains in this study.

Availability of data and materials

The data generated during the current study are available. The analyzed data and scripts are included in the article, tables, figures and its additional files.

Author's contributions

Jing Hu conceived the study. Li Wang implemented the study and drafted the manuscript. Li Wang and Ling Wang wrote the custom scripts and performed statistical analysis. Ling Wang, Zhihua Liu and Huijie Guo interpreted the data. Youzhao Liu, Zhiyun Wang, Nancai Zheng and Qing Chen provided experimental materials. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. 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**(3):181-186.

2. Grissa I, Vergnaud G, Pourcel C: **The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats.** *BMC BIOINFORMATICS* 2007, **8**:172.

3. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats**. *GENOME BIOL* 2007, **8**(4):R61.

4. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of forty-five CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes**. *PLOS COMPUT BIOL* 2005, **preprint**(2005):e60.

5. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD: **Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin**. *MICROBIOLOGY+* 2005, **151**(Pt 8):2551-2561.

6. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C: **Short motif sequences determine the targets of the prokaryotic CRISPR defence system**. *MICROBIOLOGY+* 2009, **155**(Pt 3):733-740.

7. Szczepankowska A: **Role of CRISPR/cas system in the development of bacteriophage resistance**. *ADV VIRUS RES* 2012, **82**:289-338.

8. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies**. *MICROBIOLOGY+* 2005, **151**(Pt 3):653-663.

9. Tsui TK, Li H: **Structure Principles of CRISPR-Cas Surveillance and Effector Complexes**. *ANNU REV BIOPHYS* 2015, **44**:229-255.

10. Samson JE, Moineau S: **Bacteriophages in food fermentations: new frontiers in a continuous arms race**. *Annu Rev Food Sci Technol* 2013, **4**:347-368.

11. Bhaya D, Davison M, Barrangou R: **CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation**. *ANNU REV GENET* 2011, **45**:273-297.

12. Woo PC, Lau SK, Teng JL, Que T, Yung RW, Luk W, Lai RW, Hui W, Wong SS, Yau H *et al*: **Association of *Laribacter hongkongensis* in community-acquired gastroenteritis with travel and eating fish: a multicentre case-control study**. *The Lancet* 2004, **363**(9425):1941-1947.

13. Woo PCY, Ma SSL, Teng JLL, Li MWS, Kao RYT, Lau SKP, Yuen K: **Construction of an inducible expression shuttle vector for *Laribacter hongkongensis* , a novel bacterium associated with gastroenteritis**. *FEMS MICROBIOL LETT* 2005, **252**(1):57-65.

14. Yuen KY, Woo PCY, Teng JLL, Leung KW, Wong MKM, Lau SKP: **Laribacterhongkongensis gen. nov., sp. nov., a Novel Gram-Negative Bacterium Isolated from a Cirrhotic Patient with Bacteremia and Empyema**. *J CLIN MICROBIOL* 2001, **39**(12):4227-4232.

15. Feng J, Yan H, Chowdhury N, Neogi SB, Yamasaki S, Shi L, Hu J, Chen Q: **Identification and characterization of integron-associated antibiotic resistant *Laribacter hongkongensis* isolated from aquatic products in China**. *INT J FOOD MICROBIOL* 2011, **144**(3):337-341.

16. Ni XP, Ren SH, Sun JR, Xiang HQ, Gao Y, Kong QX, Cha J, Pan JC, Yu H, Li HM: ***Laribacter hongkongensis* Isolated from a Patient with Community-Acquired Gastroenteritis in Hangzhou City**. *J CLIN MICROBIOL* 2006, **45**(1):255-256.

17. Woo PCY, Kuhnert P, Burnens AP, Teng JLL, Lau SKP, Que T, Yau H, Yuen K: ***Laribacter hongkongensis*: a potential cause of infectious diarrhea**. *DIAGN MICR INFEC DIS* 2003, **47**(4):551-556.

18. Raja MK, Ghosh AR: ***Laribacter hongkongensis*: an emerging pathogen of infectious diarrhea**. *FOLIA MICROBIOL* 2014, **59**(4):341-347.

19. Woo PC, Teng JL, Tsang AK, Tse H, Tsang VY, Chan KM, Lee EK, Chan JK, Ma SS, Tam DM *et al*: **Development of a multi-locus sequence typing scheme for *Laribacter hongkongensis*, a novel bacterium associated with freshwater fish-borne gastroenteritis and traveler's diarrhea**. *BMC MICROBIOL* 2009, **9**:21.

20. Lau SK, Wong GK, Li MW, Woo PC, Yuen KY: **Distribution and molecular characterization of tetracycline resistance in *Laribacter hongkongensis***. *J Antimicrob Chemother* 2008, **61**(3):488-497.

21. Ni X, Sun J, Kong Q, Kong F, Brown M, Shen L, Cha J, Xiang H, Xu H, Jin H: **Isolation of *Laribacter hongkongensis* from Little Egrets (Egretta garzetta) in Hangzhou, China**. *LETT APPL MICROBIOL* 2011, **52**(5):465-467.

22. Feng J, Hu J, Lin J, Liu S, Chowdhury N, Zhang Q, Li J, Shi L, Yamasaki S, Chen Q: **The prevalence, antimicrobial resistance and PFGE profiles of Laribacter hongkongensis in retail freshwater fish and edible frogs of southern China**. *FOOD MICROBIOL* 2012, **32**(1):118-123.

23. Dorny P, Praet N, Deckers N, Gabriel S: **Emerging food-borne parasites**. *VET PARASITOL* 2009, **163**(3):196-206.

24. Wang Z, Zhu J, Liu Z, Liu Y, Zheng N, Feng M, Chen Q, Yu W, Jiang L, Hu J: **Multi-locus sequence typing of *Laribacter hongkongensis* isolates from freshwater animals, environment and diarrhea patients in southern China**. *INT J FOOD MICROBIOL* 2017, **245**:98-104.

25. Hu J ZJZN: **Detection of *Laribacter hongkongsis* in freshwater products in Jiangmen City, China.** *Disease Surveillance* 2013, **28**(07):532-535.

26. Liu Y, Chen W, Liu Z, Wang L, Ge J, Han Y, Chen Q, Hu J: **First dectection of *Laribacter Hongkongensis* from feces samples in sewer rat**. *Chinese Journal od Zoonoses* 2017(04):343-348.

27. Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF: **CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus**. *MBIO* 2015, **6**(2).

28. C. Dı ez-Villasenor CAJG: **Diversity of CRISPR loci in *Escherichia coli***. *MICROBIOLOGY+* 2010, **156**(5):1351-1361.

29. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes**. *SCIENCE* 2007, **315**(5819):1709-1712.

30. Marraffini LA, Sontheimer EJ: **CRISPR Interference Limits Horizontal Gene Transfer in *Staphylococci* by Targeting DNA**. *SCIENCE* 2008, **322**(5909):1843-1845.

31. Wang P, Zhang B, Duan G, Wang Y, Hong L, Wang L, Guo X, Xi Y, Yang H: **Bioinformatics analyses of *Shigella* CRISPR structure and spacer classification**. *World Journal of Microbiology and Biotechnology* 2016, **32**(3).

32. Yosef I, Goren MG, Qimron U: **Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli***. *NUCLEIC ACIDS RES* 2012, **40**(12):5569-5576.

33. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action**. *BIOL DIRECT* 2006, **1**:7.

34. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R: **Diversity, Activity, and Evolution of CRISPR Loci in *Streptococcus thermophilus***. *J BACTERIOL* 2008, **190**(4):1401-1412.

35. Touchon M, Rocha EP: **The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella***. *PLOS ONE* 2010, **5**(6):e11126.

36. Price EP, Smith H, Huygens F, Giffard PM: **High-Resolution DNA Melt Curve Analysis of the Clustered, Regularly Interspaced Short-Palindromic-Repeat Locus of Campylobacter jejuni**. *APPL ENVIRON MICROB* 2007, **73**(10):3431-3436.

37. Ginevra C, Jacotin N, Diancourt L, Guigon G, Arquilliere R, Meugnier H, Descours G, Vandenesch F, Etienne J, Lina G *et al*: **Legionella pneumophila Sequence Type 1/Paris Pulsotype Subtyping by Spoligotyping**. *J CLIN MICROBIOL* 2012, **50**(3):696-701.

38. Palmer KL, Gilmore MS: **Multidrug-resistant *enterococci* lack CRISPR-cas**. *MBIO* 2010, **1**(4).

39. Wright MN, Gola D, Ziegler A: **Preprocessing and Quality Control for Whole-Genome Sequences from the Illumina HiSeq X Platform**. *Methods Mol Biol* 2017, **1666**:629-647.

40. Grissa I, Vergnaud G, Pourcel C: **CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.** *NUCLEIC ACIDS RES* 2007, **35**(Web Server):W52-W57.

41. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *NUCLEIC ACIDS RES* 1997, **25**(17):3389-3402.

42. Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, Branger C: **CRISPR Distribution within the *Escherichia coli* Species Is Not Suggestive of Immunity-Associated Diversifying Selection**. *J BACTERIOL* 2011, **193**(10):2460-2467.

43. Biswas A, Gagnon JN, Brouns SJJ, Fineran PC, Brown CM: **CRISPRTarget_Bioinformatic prediction and analysis of crRNA targets**. *RNA BIOL* 2013, **10**(5):817-827.

Tables

Table 1 Primers for the two CRISPR loci and *cas* genes.

Genes	F-Regiona	Forward primer (5'–3')	R-Regionb	Reverse primer (5'–3')	annealing temperature (°C)
CRISPR4.1	the leader sequence	ATGGGTAGATGCTGACCTTT	sulfate transporter subunit *(sts)*	TGACTGACCAGAGGAGGAGT	57.0
	the *csy4* gene	CGCTTCTTGCTCTTTATCG			
CRISPR4.2	*Phage_TAC_7 (tac_7)*	CTGGCTCGCTGTTACAGCT	phage tail protein *(Pt.p)*	GCCAATGTTACAATATCAAACA	60.0
rev_3-cas1	*rve_3* domain-containing protein *(rev_3)*	TGTCACCGAGGGCTTTCTT	*cas1*	TCTTCCGTCAGCAAGTTGG	60.0
cas3	*cas3(5')*	CCTGGCAAGACAAGGCTTA	*cas3 (3')*	CCGAACTCAGTAAACGCAG	57.5
csy1- csy3	*csy1*	ACTACCTGCTTGCCTCCTT	*csy3*	ACAGCGTCTTGCTCTTGTC	56.6
csy4-the leader sequence	*csy4*	CGCTTCTTGCTCTTTATCG	the leader sequence	ACGGTATGGCTTGTCAGTC	56.5
16S rRNA	LPW264	GAGTTTGATCMTGGCTCAG	LPW265	GNTACCTTGTTACGACTT	55.0
distinguished sequence	upstream	ATC CCT AAG GCT AAT ACC CT	downstream	AAT CTC TTC GAC GTT CGG TA	53.0[45]

Table 2 The distribution of CRISPR4.1, CRISPR4.2 in 118 *L.hongkongensis* strains.

Strains	No. of isolates			CRISPR4.1				CRISPR4.2	
		+	-	positive rate (%)		+	-	positive rate (%)	
Fish-origin	66	63	3	95.45	$\chi2$=4.506 p=0.092	47	19	71.21	$\chi2$ = 0.767 P = 0.681
Frog-origin	42	37	5	88.10		31	11	73.81	
human-origin	7	5	2	71.43		4	3	57.14	
Sewer rat									
	2	2	0	100.00		2	0	100.00	
Water reservoir	1	1	0	100.00		1	0	100.00	
Total	118	108	10	91.53		85	33	72.03	$\chi2$ =15.043 P 0.001

Table 3 Analysis of CRISPR repeat sequences.

CRISPR	Type	DR sequence (5'-3')a	No. of stains	No. of sequenced repeats	Frequency (%)
CRISPR4.1	Typical repeat	GTTCACTGCCGGACAGGCAGCTCAGAAA	99	1595	92.68
		GTTCACTGCCGGACAGGCAGCTTAGAAA	7	107	6.63
		ATTCACTGCCGGACAGGCAGCTTAGAAA	2	16	0.93
	Repeat variants	GTTCACTGCCGGACAGGCAGCTCAGAAT	1	1	0.06
		ATTCACTGCCGGACAGGCAGCTCAGAAA	1	1	0.06
	Terminal repeat	TTTCACTGCCGGACAGGCAGCTTAGAAA	1	1	0.06
CRISPR4.2	Typical repeat	GTTCACTGCCGGACAGGCAGCTCAGAAA	85	949	91.78
	Repeat variants	-	-	-	-
	Terminal repeat	GTTCACTGCCGGACAGGCAGCGAAGTCT	83	83	8.03
		GTTCACTGCCGGACAGGCAGCTGAAGTC	2	2	0.19

Table 4 CRISPR4.1 spacer diversity in 118 *L. hongkongensis* strains.

Strains	No. of isolates	CRISPR4.1				
		No. of isolates with CRISPR(%)	No. of spacers	Mean no. of spacers±SD	No. of unique spacers (%)	No. of alleles(%)
Fish-origin	66	63 (95.45)	742	11.24±3.90	87 (11.73)	15 (23.81)
Frog-origin	42	37 (88.10)	724	17.24±10.74	410 (56.63)	26 (70.27)
human-origin	7	5 (71.43)	118	23.60±8.85	115 (97.46)	5 (100)
Sewer rat	2	2 (100.00)	17	8.50±10.61	17 (100.00)	2 (100)
Water reservoir	1	1 (100.00)	12	12±0.00	12 (100.00)	1 (100)
Total	118	108 (91.53)	1613	14.94±7.45	579 (35.90)	40 (37.04)

Table 5 CRISPR4.2 spacer diversity in 118 *L. hongkongensis* strains.

Strains	No. of isolates	CRISPR4.2				
		No. of isolates with CRISPR(%)	No. of spacers	Mean no. of spacers ± SD	No. of unique spacers (%)	No. of alleles(%)
Fish-origin	66	47 (71.21)	433	9.21±3.68	55 (12.70)	11 (23.40)
human-origin	7	4 (57.14)	73	18.25±3.50	72 (98.63)	4 (100)
Sewer rat	2	2 (100.00)	9	4.50±0.71	9 (100.00)	2 (100)
Water reservoir	1	1 (100.00)	5	5.00±0.00	5 (100.00	1 (100)
Total	118	85 (72.03)	949	11.16±6.63	401(42.26)	37 (43.53)

Table 6 Comparison of spacers between CRISPR4.1 and CRISPR4.2.

CRISPR	Distribution of CRISPR	Mean no. of spacers± SD	Unique spacers		Alleles	
			Yes	No	Yes	No
CRISPR4.1	108	14.94±7.45	579	1034	40	68
CRISPR4.2	85	11.16±6.63	401	548	37	48
	$\chi 2$=15.043	t=3.660	$\chi 2$=10.229		$\chi 2$=0.836	
	p<0.001	p<0.001	p<0.001		p=0.361	

Table 7 Comparison of CRISPR4.1 in different origin *L. hongkongensis* strains.

Strains	Distribution of CRISPR4.1	Mean no. of spacers± SD	Unique spacers		Alleles	
			Yes	No	Yes	No
Fish-origin	63	11.24±3.90	87	655	15	48
Frog-origin	37	17.24±10.74	410	314	26	11
Human-origin	5	23.60±8.85	115	3	5	0
	$\chi 2$=4.506	F=23.217	$\chi 2$=497.729		$\chi 2$=29.756	
	p=0.092	p<0.001	p<0.001		p<0.001	

Table 8 Comparison of CRISPR4.2 in different origin *L. hongkongensis* strains.

Strains	Distribution of CRISPR4.2	Mean no. of spacers± SD	Unique spacers		Alleles	
			Yes	No	Yes	No
Fish-origin	47	9.21±3.68	55	378	11	36
Frog-origin	31	13.84±8.78	304	125	21	10
Human-origin	4	18.25±3.50	72	1	4	0
	$\chi 2$=0.767	F=7.942	$\chi 2$=318.312		$\chi 2$=22.321	
	p=0.681	p=0.001	p<0.001		p<0.001	

Table 9 Spacer identities to mobile genetic and no-mobile genetic elements.

Strain	CRISPR and spacer no.a	Sequence identityb	RepresentativeBlastn hit	Area of identityin Blastn hit	Genbank accession
SW33	4.1-6	31/32	*Xylella fastidiosa 9a5c* plasmid pXF51	XFa0016 conjugal transfer protein	NC_002490.1
	4.1-6	31/32	*Xylella fastidiosa* strain U24D plasmid pXF51ud	XFUD_RS1302 conjugal transfer protein	NZ_CP009791.1
	4.1-6	31/32	*Xylella fastidiosa* strain Pr8x plasmid pXF39	XFUD_RS13020 conjugal transfer protein	NZ_CP009827.1
	4.1-6	31/32	*Xylella fastidiosa s*train J1a12 plasmid pXF51-J1	OY18_RS13950 type IV secretion protein	NZ_CP009825.1
	4.1-6	29/32	*Xylellafastidiosa MUL0034* plasmid unnamed2	P303_RS12040 conjugal transfer protein TraG	NZ_CP006739.1
	4.1-6	29/32	*Xylella fastidiosa subsp. fastidiosa GB514* plasmid	XFLM_RS10960 conjugal transfer protein TraG	NC_017561.1
	4.1-6	29/32	*Xylella fastidiosa* strain J1a12 plasmid pXF27-J1	OY18_RS13795 conjugal transfer protein TraG	NZ_CP009824.1
	4.1-6	29/32	*Xylella fastidiosa plasmid pXF-RIV19*	pXFRIV19_p06 TRAG protein	NC_014113.1
	4.1-6	29/32	*Xylella fastidiosa subsp. sandyi Ann-1* plasmid unnamed1	D934-RS12705 TRAG protein	NZ_CP006697.1
	4.1-6	29/32	*Xylella fastidiosa* plasmid pXF-RIV11	pXFRIV11_p06 TRAG protein	NC_014111.1
SW44	4.1-7	30/32	*Ralstonia* phage phiRSA1	ORF14 BAF52391.1 phage-related protein	AB276040
SW57	4.1-5	30/32	*Curvibacter sp.* AEP1-3 plasmid	AEP_RS20460 hypothetical protein	NZ_CP015699.1
	4.1-5	29/32	*Xylella fastidiosa* strain U24D plasmid pXF51ud	XFUD_RS13020 conjugal transfer protein	NZ_CP009791.1
	4.1-5	29/32	*Xylella fastidiosa* strain Pr8x plasmid pXF39	XFPR_RS12925 conjugal transfer protein	NZ_CP009827.1
	4.1-5	29/32	*Xylella fastidiosa* strain J1a12 plasmid pXF51-J1	OY18_RS13950 type IV secretion protein	NZ_CP009825.1
	4.1-5	29/32	*Xylella fastidiosa 9a5c* plasmid pXF51	XF_RS12215 conjugal transfer protein	NC_002490.1
	4.1-6	29/32	*Xylella fastidiosa* strain U24D plasmid pXF51ud	XFUD_RS12970 VirB4 family type IV secretion/conjugal transfer ATPase	NZ_CP009791.1
	4.1-6	29/32	*Xylella fastidiosa* strain J1a12 plasmid pXF51-J1	OY18_RS13905 VirB4 family type IV secretion/conjuga transfer ATPase	NZ_CP009825.1
	4.1-6	29/32	*Xylella fastidiosa* strain Pr8x plasmid pXF39	XFPR_RS12875 VirB4 family type IV secretion/conjugal transfer ATPase	NZ_CP009827.1
	4.1-6	29/32	*Xylella fastidiosa 9a5c* plasmid pXF51	XFa0007 VirB4 family type IV secretion/conjugal transfer ATPase	NC_002490.1
	4.1-5	30/32	*Martelella sp.* AD-3	AZF01_20600 hypothetical protein	CP014275.1
SW42	4.1-5	30/32	*Gordonia* phage Attis	SEA_ATTIS_4 capsid maturation protease	KU963247

	4.1-5	30/32	*Gordonia* phage SoilAssassin	BJD59_gp04capsid maturation protease	KU963246
	4.1-5	29/32	*Gordonia* phage BaxterFox	SEA_BAXTERFOX_4 capsid maturation protease	KU963263
	4.1-5	29/32	*Gordonia* phage Yeezy	SEA_YEEZY_4 capsid maturation protease	KU963249
	4.1-5	29/32	*Gordonia* phage Howe	PBI_HOWE_4 capsid maturation protease	KU252585
	4.1-5	30/32	*Martelella sp.* AD-3	AZF01_20600 hypothetical protein	CP014275.1
SW3	4.2-8	29/32	*Pseudomonas* phage vB_PaeS_PMG1	PMG1_00057 hypothetical protein	HQ711985
SW2	4.2-6	31/32	*Pseudomonas* phage F_ET2439spPa1651	NOR	KM389260
	4.2-6	30/32	*Pseudomonas* phage JBD93	BH777_gp28 DNA-binding protein	KU199709
	4.2-6	30/32	*Pseudomonas* phage JBD16C	JBD16C_25 DNA-binding protein	KU199707
	4.2-6	30/32	unidentified phage genome assembly 2P1	2P1_26 putative DNA-binding protein	LN907801
	4.2-6	30/32	*Pseudomonas* phage JBDs5	NOR	KM389464
SY34	4.2-19	32/32	*Enterobacteria* phage epsilon15	ABK13674.1 hypothetical protein	
ZW21	4.1-8	29/32	*Burkholderia* phage KS9	terminase gp2	FJ982340
ZY51	4.1-8	30/32	*Sphingobium japonicum* UT26S plasmid pCHQ1 DNA	SJA_RS19780 DNA cytosine methyltransferase	NC_014007.1
	4.1-8	29/32	*Burkholderia cenocepacia* strain ST32 plasmid pBCEN1232	TQ36_RS35975 DNA cytosine methyltransferase	NZ_CP011920.1
	4.1-8	29/32	*Burkholderia cenocepacia* strain VC1254 plasmid	A8F32_RS19600 DNA cytosine methyltransferase	NZ_CP019677.1
	4.1-8	29/32	*Cupriavidus sp.* USMAA1020	NOR	CP017754.1
	4.1-8	29/32	*Cupriavidus sp.* USMAHM13	NOR	CP017751.1
ZW28	4.2-19	30/32	*Enterobacteria* phage epsilon15	ABK13674.1 hypothetical protein	AY150271
ZY40	4.2-9	32/32	*Burkholderia* phage BcepMigl	BcepMigl_gp47 portal protein	JX104231
	4.2-9	32/32	*Burkholderia* phage DC1	DC1_00043 portal protein	JN662425
	4.2-3	31/32	*Alicycliphilus denitrificans* K601 plasmid pALIDE201	ALIDE2_RS23645 type-IV secretion system protein TraC	NC_015423.1
	4.2-3	31/32	*Acidovorax sp.* JS42 plasmid pAOVO01	AJS_RS20785 type-IV secretion system protein TraC	NC_008765.1
	4.2-3	31/32	*Acidovorax sp.* P4 plasmid pACP4.4	CBP36_RS21790 type-IV secretion system protein TraC	NZ_CP021370.1
	4.2-9	30/32	*Burkholderia cepacia* phage Bcep22	Bcep22_gp51 phage portal protein	AY349011

	4.2-3	30/32	*Pandoraea pnomenusa* strain MCB032 plasmid unnamed 1	A6P55_RS25085type-IV secretion system protein TraC	NZ_CP015372.1
	4.2-9	29/32	*Burkholderia cenocepacia* phage BcepIL02	BcepIL02_gp46 phage portal protein	FJ937737
SW45	4.1-6	32/32	*Azoarcus sp.* EbN1 plasmid 2	EB_RS22030 DNA repair protein	NC_006824.1
	4.1-6	30/32	*Brevibacterium sp.* Ap13 plasmid pAP13	AP13_p00150 putative Error-prone, lesion bypass DAN N polymerase V UmuC	NC_022590.1
	4.1-6	30/32	*Arthrobacter chlorophenolicus* A6 plasmid pACHL01	ACHL_RS21770 DNA polymerase V subunit UmuC	NC_011879.1
	4.1-6	29/32	*Kocuria turfanensis* strain HO-9042 plasmid unnamed4	AYX06_RS18795 hypothetical protein	NZ_CP014483.1
	4.1-6	29/32	*Aeromonas sp.* O23A plasmid pO23AP4	O23A_RS22685 protein UmuC	NZ_CP021658.1
	4.1-6	29/32	*Kocuria turfanensis* strain HO-9042 plasmid unnamed1	AYX06_RS17120 hypothetical protein	NZ_CP014481.1
	4.1-6	29/32	*Serratia marcescens* SM39 plasmid pSMC2	SM39_RS24820DNA polymerase V subunit UmuC	NZ_AP013065.1
	4.1-9	31/32	*Halolamina pelagica* strain CDK2 SY89_scf03	SY89_RS16325 hypothetical protein	NZ_LGUC01000003.1
623	4.1-22	30/32	*Chromobacterium violaceum* ATCC 12472	CV_0342 conserved hypothetical protein	AAQ58021.1
SW51	4.1-8	32/32	*Azoarcus sp. EbN1* plasmid 2	p2A244 DNA repair protein	CR555308.1
	4.1-8	31/32	*Klebsiella pneumoniae* strain MLST-15	CEV20_03835 DNA polymerase V subunit UmuC	CP022127.1

Figures

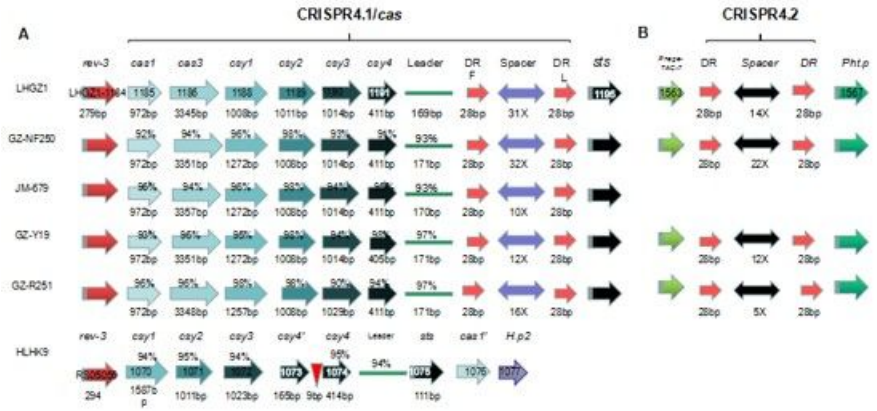

Figure 1

Organization and comparison of the CRISPR4.1/Cas systems and CRISPR4.2 locus of different L. hongkongensis strains. In the 6 strains analyzed, the CRISPR4.1/Cas system was located in the region between the LHGZ1-1183 (rve_3) gene and LHGZ1-1195 (sts). CRISPR4.2 was localized downstream of LHGZ1-1563 (Phage_TAC_7) and upstream of LHGZ1-1567 (Pht.p). The cas genes

are marked with blue arrows, and gene names, gene numbers and gene size are given above, within, and under the blue arrows, respectively. The leader sequence (green line), the direct repeats (red arrows), and the spacer sequences (violet arrows) are listed. The number of spacers is marked below the violet arrows. Sequence identity (%) of the cas genes and the leader to the representative DNA regions of L. hongkongensis LHGZ1 is given. A 9 bp insertion sequence is indicated by a red inverted triangle. Homologous genes are marked with the same color.

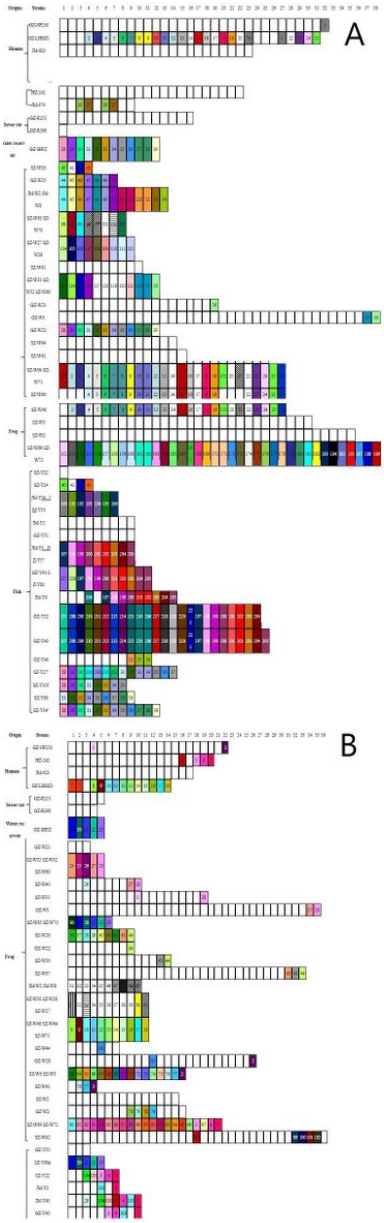

Figure 2

A. Sequence diagram of the spacers of CRISPR4.1. B. Sequence diagram of the spacers of CRISPR4.2. Graphic illustration of the CRISPR spacers in the L. hongkongensis strains. Transcription direction was from the leader (left) to arrays (right). Repeat sequence was not described. Structure of spacers were labeled with boxes, and those identical spacers in one strain or present in different strains were noted with a unique number. The latter was also marked with a particular color. The specific spacers (white boxes) only exist in one strain. More than four strains shared a special allele, just one was showed: a: CRISPR4.1 of SZ-Y27 was identical to that of SZ-Y51, SZ-Y62, SZ-Y101, GZ-Y4, GZ-Y14; b, CRISPR4.1 of SZ-Y34 was identical to a total of 40 fish origins strains, including SZ-Y58, SZ-Y19, SZ-Y49, SZ-Y42 etc; c, CRISPR4.2 of SZ-Y49 was identical to that of SZ-Y58, JM-Y27, JM-Y28, JM-Y29, JM-Y36, JM-Y62, JM-Y63, GZ-Y44, GZ-Y45, GZ-Y47, GZ-Y49, GZ-Y53, GZ-Y19; d, CRISPR4.2 of SZ-Y27 was identical to that of SZ-Y51, SZ-Y62, SZ-Y19, SZ-Y101, GZ-Y4, GZ-Y14, GZ-Y8, JM-Y16, JM-Y10.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1.xls