A comprehensive curation demonstrates the dynamic evolutionary patterns of prokaryotic CRISPRs

Guoqin Mai[1,*], Ruiquan Ge[1,2,*],Guoquan Sun[1],Qinghan Meng[1,2], Fengfeng Zhou[1,#]

1Shenzhen Institutes of Advanced Technology, and Key Lab for Health Informatics, Chinese Academy of Sciences,

2Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, Guangdong, P.R. China, 518055.

* These authors contribute equally to this work.

#Corresponding author: Fengfeng Zhou, phone: +86-755-86392200; fax: +86-755-86392299; e-mail: FengfengZhou@gmail.com, or ff.zhou@siat.ac.cn.Web site: http://www.healthinformaticslab.org/ffzhou/.

Material and Methods

Initial CRISPR annotations

The complete annotation of CRISPRs in microbial genomes was downloaded from the latest version of the database CRISPRdb(Grissa, et al., 2007), which was updated on April14, 2014. Among the 2,762 prokaryotic genomes, 4,065 CRISPRs are annotated in 1,302 genomes. The questionable structures in CRISPRdb are omitted. If a genome harbors CRISPRs, it has 3.12 CRISPRs on average.

SpacerDB consists of all the annotated CRISPR spacer sequences in the database CRISPRdb, and was downloaded from the web site of CRISPRdb (Grissa, et al., 2007) on April 14, 2014. CRISPR spacers are not random nucleotide sequences, and are supposed to originate from the foreign invasive elements like phages (Deveau, et al., 2010). So a DR flanking sequence matching a known spacer may also be a spacer.

Analysis techniques and tools

All the 2,762 prokaryotic genomes covered by CRISPRdb are re-annotated by the three existing programs, CRT, CRISPRFinder and PILER-CR. Inconsistency is observed between the three programs, mostly on the boundaries of annotated CRISPRs. So based on a union set of the annotated CRISPRs from CRISPRdb and the three programs, a comprehensive manual curation is conducted to screen for missing DR signals in the flanking regions.For an annotated CRISPR, the homologous copies of DRs were screened by the local copy of NCBI BLAST version 2.2.25 (Wheeler, et al., 2008). NCBI BLAST is also used to screen the homologous matches of a given spacer sequence.

A CRISPR is usually activated by the closest CRISPR associated (Cas) genes (Lawrence and White, 2011), and multiple CRISPRs may share the same group of Cas genes, if there is only one such group neighboring to these CRISPRs.

Figure S1. Structure of a CRISPR

Figure S1. Structure of a CRISPR. DR is the direct repeat flanking the spacer sequence. The grey box represents the DR sequence, and DRs of the same CRISPR are homologous to each other.The patterned boxes represent the spacer sequence.

Figure S2. 6 cases with one new DR

A few DRs were not detected in the flanking regions of CRISPRs, as demonstrated in Figure 1 (a) and (b). Six CRISPRs may have one missing DR in the flanking region, as in Figure 1 (a). For example, by screening for more DRs in the CRISPR flanking regions, we propose 10 spacers for the CRISPR NC_010125_ 2181482_2182111 in *Gluconacetobacter diazotrophicus* PAl 5, as in Figure S2. But CRISPRdb only detected 9 spacers for this CRISPR. The new DR is also

confirmed using the tool CRISPRFinder (Grissa, et al., 2007).Four other CRISPRs (NC_010125_62935_64899, NC_010125_2253748_2255112, NC_011365_388303_388536, NC_011365_460172_461964) in the same bacterial strain *Gluconacetobacter diazotrophicus* PAl 5missed one complete DR in their flanking regions too, as in the Figure S2.

> *Gluconacetobacter diazotrophicus* PAl 5_NC_010125_2181482_2182177_10_our results

```
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    CTGGCCGGTTAATTGCGTGACGGCGGGTTC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    ATCGCATGACCTTTGGTTTCGACCGGGTAT
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    TCGAAGACCGCGCTGGACGACATGGGGAAG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    CGATTCCGATACCTTGCGCGTGCGCACTGG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GCAAGATAACGGCCTTCGGCCACACGAAGA
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GCAGATTTCGATACCGGCAACGACGGTCTT
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GAGCGGAGGCAGCTTGTGGCCCAATATGGC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GATCGCGCATGATGGCCGGATCAACGCTCA
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    ATCGAGCGCCGGCCGGACGTCGTTGTCGCC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    CATGGTGTCGAGCTTGCTCGGCGCGTTCTC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC
```

> *Gluconacetobacter diazotrophicus* PAl 5_NC_010125_ 2181482_2182111_9_gold standard

```
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    CTGGCCGGTTAATTGCGTGACGGCGGGTTC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    ATCGCATGACCTTTGGTTTCGACCGGGTAT
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    TCGAAGACCGCGCTGGACGACATGGGGAAG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    CGATTCCGATACCTTGCGCGTGCGCACTGG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GCAAGATAACGGCCTTCGGCCACACGAAGA
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GCAGATTTCGATACCGGCAACGACGGTCTT
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GAGCGGAGGCAGCTTGTGGCCCAATATGGC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    GATCGCGCATGATGGCCGGATCAACGCTCA
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC    ATCGAGCGCCGGCCGGACGTCGTTGTCGCC
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC
```

> *Gluconacetobacter diazotrophicus* PAl 5_NC_010125_62935_64899_29_ our results

```
GTTTTAATCCCCGCTTCCGCGTGGGGAGCGAC        GACAACCTTGCGACTCCTGTCCGACGCGGTCGGC
GTTTCAATCCACGCTCCCGCACGGGGAGCGAC        GGGTGCGATCCTCCGCATGGTGCAGGAGGACATC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CGCCTCGTCGATGATGACGGGCGAATTCATCACGCC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        GATCACGCGCACGCACCGCAGCCACTCGTCGTAGG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CAACTGTCCGATATCTCGGCCAATCTCTCCAACG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CGTTTCTCGATACGCGGTCCAATCCCCTCCTTCTT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CAACGTCTCCGACCCTTGACGTGGTCTCACGCCT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        GAGCCGGGGACGAGCGTGTAATCCTGCTGCGCCG
.........
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CTTCTGTTCGGCATCGCCGGCGCGTCTGGTTCCG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        AAGATCCATGTCCTTCCAAGCGCATTCCCTAAAT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        GGGAAACGAACCGTGATGTCGTGGCGGAAAGAAAG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC
```

> *Gluconacetobacter diazotrophicus* PAl 5_NC_010125_ 62935_64832_28_gold standard

```
GTTTTAATCCCCGCTTCCGCGTGGGGAGCGAC        GACAACCTTGCGACTCCTGTCCGACGCGGTCGGC
GTTTCAATCCACGCTCCCGCACGGGGAGCGAC        GGGTGCGATCCTCCGCATGGTGCAGGAGGACATC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CGCCTCGTCGATGATGACGGGCGAATTCATCACGCC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        GATCACGCGCACGCACCGCAGCCACTCGTCGTAGG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CAACTGTCCGATATCTCGGCCAATCTCTCCAACG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CGTTTCTCGATACGCGGTCCAATCCCCTCCTTCTT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CAACGTCTCCGACCCTTGACGTGGTCTCACGCCT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        GAGCCGGGGACGAGCGTGTAATCCTGCTGCGCCG
.........
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        CTTCTGTTCGGCATCGCCGGCGCGTCTGGTTCCG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC        AAGATCCATGTCCTTCCAAGCGCATTCCCTAAAT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC
```

> *Gluconacetobacter diazotrophicus* PAl 5_NC_010125_2253747_2255112_22_our results

```
AGATTCATCCCTGCATATGCGAGGAACAC           AAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTCATCCCCGCACGTGCGGGAACACG           TCCCCCTGCTCGGCAACCCCAAACTCAAGAT
CGGTTCATCCCCGCACGTGCGGGGAACAC           GCCGCGACGGTGAAGGGGAGGTTGTTCGACGC
CGGTTCATCCCCGCACGTGCGGGGAACAC           GAGTCACCGGATTCCCATCTCATCCTAATATC
CGGTTCATCCCCGCACGTGCGGGGAACAC           CTTTTCTCAGCCGACACTCTCCATGCAAGAAC
CGGTTCATCCCCGCACGTGCGGGGAACAC           ACGCGGGTCTATGAGATTCCGATCATGCAGGA
CGGTTCATCCCCGCACGTGCGGGGAACAC           CATTGACCCGGCACCCTCCCCATGTGACGTTT
CGGTTCATCCCCGCACGTGCGGGGAACAC           TGGCTCGGGACCATCTTAGCGGACGACATGGC
.........
CGGTTCATCCCCGCACGTGCGGGGAACAC           TTGCGTCAAACGAAGCACAATATCAGGCGCAG
CGGTTCATCCCCGCACGTGCGGGGAACAC           TTGCCTCTAACGAGGCGCAGTATCAGGCACAG
CGGTTCATCCCCGCACGTGCGGGGAACAC           CCGTCGTTTCGATCGTCACGGCGATCATCGGC
CGGTTCATCCCCGCACGTGCGGGGA
```

> *Gluconacetobacter diazotrophicus* PAl 5_NC_010125_ 2253747_2255055_21_gold standard

```
AGATTCATCCCTGCATATGCGAGGAACAC           AAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTCATCCCCGCACGTGCGGGAACACG           TCCCCCTGCTCGGCAACCCCAAACTCAAGAT
CGGTTCATCCCCGCACGTGCGGGGAACAC           GCCGCGACGGTGAAGGGGAGGTTGTTCGACGC
CGGTTCATCCCCGCACGTGCGGGGAACAC           GAGTCACCGGATTCCCATCTCATCCTAATATC
CGGTTCATCCCCGCACGTGCGGGGAACAC           CTTTTCTCAGCCGACACTCTCCATGCAAGAAC
CGGTTCATCCCCGCACGTGCGGGGAACAC           ACGCGGGTCTATGAGATTCCGATCATGCAGGA
CGGTTCATCCCCGCACGTGCGGGGAACAC           CATTGACCCGGCACCCTCCCCATGTGACGTTT
CGGTTCATCCCCGCACGTGCGGGGAACAC           TGGCTCGGGACCATCTTAGCGGACGACATGGC
.........
CGGTTCATCCCCGCACGTGCGGGGAACAC           TTGCGTCAAACGAAGCACAATATCAGGCGCAG
CGGTTCATCCCCGCACGTGCGGGGAACAC           TTGCCTCTAACGAGGCGCAGTATCAGGCACAG
CGGTTCATCCCCGCACGTGCGGGGAACAC
```

> *Gluconacetobacter diazotrophicus* PAl 5_NC_011365_460172_461964_29_our results

AGATTCATCCCTGCATATGCGAGGA	ACACAAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTCATCCCCGCACGTGCGGGAA	CACGTCCCCCTGCTCGGCAACCCCAAACTCAAGAT
CGGTTCATCCCCGCACGTGCGGGGA	ACACGCCGCGACGGTGAAGGGGAGGTTGTTCGACGC
CGGTTCATCCCCGCACGTGCGGGGA	ACACCCTGATCCTGCGCCGACGCCGGAACCAGTCCG
CGGTTCATCCCCGCACGTGCGGGGA	ACACGCCGCGACGGTGAAGGGGAGGTTGTTCGACGC
CGGTTCATCCCCGCACGTGCGGGGA	ACACCCTGATCCTGCGCCGACGCCGGAACCAGTCCG
CGGTTCATCCCCGCACGTGCGGGGA	ACACGAGTCACCGGATTCCCATCTCATCCTAATATC
CGGTTCATCCCCGCACGTGCGGGGA	ACACCTTTTCTCAGCCGACACTCTCCATGCAAGAAC
.....	
CGGTTCATCCCCGCACGTGCGGGGA	ACACAACGTGGAATGGCACCAGATCGAAACGATGCT
CGGTTCATCCCCGCACGTGCGGGGA	ACACGCGAAGACCTGCGCCTGGAGCGCCTGCTTGTG
CGGTTCATCCCCGCACGTGCGGGGA	ACACCGTGGGGAATATCCTGACACACACATGGGCGT
CGGTTCATCCCCGCACGTGCGGGGA	

> *Gluconacetobacter diazotrophicus* PAl 5_NC_011365_460172_461907_28_gold standard

AGATTCATCCCTGCATATGCGAGGA	ACACAAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTCATCCCCGCACGTGCGGGAA	CACGTCCCCCTGCTCGGCAACCCCAAACTCAAGAT
CGGTTCATCCCCGCACGTGCGGGGA	ACACGCCGCGACGGTGAAGGGGAGGTTGTTCGACGC
CGGTTCATCCCCGCACGTGCGGGGA	ACACCCTGATCCTGCGCCGACGCCGGAACCAGTCCG
CGGTTCATCCCCGCACGTGCGGGGA	ACACGCCGCGACGGTGAAGGGGAGGTTGTTCGACGC
CGGTTCATCCCCGCACGTGCGGGGA	ACACCCTGATCCTGCGCCGACGCCGGAACCAGTCCG
CGGTTCATCCCCGCACGTGCGGGGA	ACACGAGTCACCGGATTCCCATCTCATCCTAATATC
CGGTTCATCCCCGCACGTGCGGGGA	ACACCTTTTCTCAGCCGACACTCTCCATGCAAGAAC
.....	
CGGTTCATCCCCGCACGTGCGGGGA	ACACAACGTGGAATGGCACCAGATCGAAACGATGCT
CGGTTCATCCCCGCACGTGCGGGGA	ACACGCGAAGACCTGCGCCTGGAGCGCCTGCTTGTG
CGGTTCATCCCCGCACGTGCGGGGA	

> *Gluconacetobacter diazotrophicus* PAl 5_ NC_011365_ 388303_388602 _4 _our results

AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	TCAGCCGGCTGGGCATTGCCCTACGAGAGG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	AGGGGAGCGCGTCACCTTCGACAAGGGCCT
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	CCTCGTTGCTCGCGCGCGGGTGCAGGGCTG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	TCCGCACGTTCACGCGTACCAGCATCGCGG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	

> *Gluconacetobacter diazotrophicus* PAl 5_ NC_011365_388303_388536_3_gold standard

AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	TCAGCCGGCTGGGCATTGCCCTACGAGAGG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	AGGGGAGCGCGTCACCTTCGACAAGGGCCT
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	CCTCGTTGCTCGCGCGCGGGTGCAGGGCTG
AGCCTACCATCGGCAAATCGGTAGGGAAACCACGGC	

```
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   TTTCCGGATTAGAATAATAAGGAACTTCCGGTACGGG
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   GGTATACTAAGCGTACCGTTCGTCTCAACATCGACTA
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   ACACCATATGCTCCGGGCAGAGGAGCATGTTCACACT
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   CCAGTCCTCCTCCTCAGGAGGACTGGCTTTTTCAAA
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   CACAATAAATCCCCCTTTGAAAACGAGTTCAAGC
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   TATGTAGCTGCATATCTACACCTCCACGGATATAAGAG
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   TCTTCTGGCTCTGGCACGGCGAGAACTATCTCGCCGT
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   TCCAAATTACCAAGACAAACATGCTTGTCATGGGTG
.........
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   GCCCATATCTCTTCGTGATGCTCCGCGGTCCTGCTG
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   AGCAGCATCTTTTTTCACCTCATGATATCATTATGAC
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC   TTTCTATACTGGGGATGTTGTACTTCCTTGCCTTCCTT
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC
```

> *Methanosaeta thermophila* PT_ NC_008553_670839_677982_97 _our results

> *Methanosaeta thermophila* PT _ NC_008553_670839_677982_96 _gold standard

Figure S3. 4 cases with two new DRs

Two DRs were added to each of 4 CRISPRs, as in Figure 1 (b). These DRs were missed by the database CRISPRdb mainly due to that one of the two DRs is only partially identical to the other DRs, as in Figure S3.One of the example CRISPR is NC_014152_2078344_2080300 in the bacterial genome *Thermincola sp.* JR, with 26 spacers. We propose two more DRs for this CRISPR, although one of the two new DRs is identical to the other DRs in half of its region. The mismatched region may be introduced by the gene conversion (Wang, et al., 2009)or homologous recombination (Liu and Huang, 2014)mechanism.Another piece of supporting evidence for the two new spacers comes from their BLAST matches to two known spacers in the other genomes in the SpacerDB with 91.3% and 94.4% in matching identity percentages, respectively.A spacer is supposed to originate from the foreign invasive elements. Since it is low in probability to have such almost identical sequences just by the random single nucleotide mutations,the two new candidate spacers are suggested to be real spacers originated from the same foreign invasive elements as the two homologous copies in the other genomes.Three more CRISPRs, *i.e.* NC_015865_1907425_1908328 in *Thermococcus sp.* 4557, NC_015738_2085666_2087297 in *Eggerthella sp.* YY7918, and NC_014209_791663_793738 in *Thermoanaerobacter mathranii subsp. mathranii str.* A3,are expanded with two more DRs for the same reason, as in the Figure S3.

Figure S4. 3 cases of long spacers with a partial DR

Each of 3 CRISPRs has an extraordinarily long spacer with a truncated DR inside, as demonstrated in Figure 1 (c). The representative example is the CRISPR NC_019693_6234891_6235861 with 12 spacers in the cyanobacterial strain *Oscillatoria acuminata* PCC 6304. Figure S4 illustrates that this CRISPR's ninth spacer harbors a partial DR copy with 70% length of the other DRs. And the two flanking sequences in this long spacer have reasonable lengths as spacers. So we propose that this CRISPR has 13 spacers, as in Figure S4.Similar cases are detected in two other CRISPRs, *i.e.* NC_008639_1625359_1633049 in *Chlorobium phaeobacteroides* DSM 266, and NC_007777_3904715_3905896 in*Frankia* sp. CcI3, as in the Figure S4.

>*Oscillatoria acuminata* PCC 6304_NC_019693_6234891_6235861_13_our results

```
GTTTCAATCCCGTTGCCGGGATTCATTATAATGAAAG   TCCCATGAGCATCACCAATATCCCAAGGATTAAATG
GTTTCAATCCCGCTGCCGGGATTCATTGTACTGAGAG   TCTGCTGATTCCCGCTATTACTCAGTCGGGGAAGACTAG
GTTTCAATCCCGTTGCCGGGATTCATTGTAATGAAAG   TCATACAAAGCATCACTTATGTATTGTTCTGAAAA
GTTTCAATCCCGTTGCCGGGATTCATTCTAATGAAAG   GAGGGCTTTTCAGGCTCCGTTTTGTGCGTGGCGATCG
GTTTTAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CAGCCCATTCTTTGGCTGTAAAGGGTATTTGGAGAA
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CTAAATAAACCAATCAGAGGGTTAATCATGACTTAT
GTTTCAATCCCGTTGCTGGGATTCATTGTACTGAAAG   CGCTTGTTGCGTGGCGATCGCACCGTGCGATCGTTTGGT
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   GAGCAATAAAAAGGAAGGGTGGTTAGTTCGTATTAA
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CGATCGCTCCGTCGAATCTTTATCT
          GTTGCCGGGATTCATTGTACTGAAAG   CTCAACCGGCCAAATCCTCACCGACTCGATTCAGTT
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   TCTATGCGGGGGAAATCTGGGAGGTGCTGAAACAGTGG
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CTGGGCGACAATCAGCCGGATTATGGTTTAAGTGGCT
GTTTCCATCCCGTTGCCGGGATTCATTGTACTGAAAG   TTCACTTGTTCAACGTCAACGACTCACCCCGACT
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG
```

>*Oscillatoria acuminata* PCC 6304_NC_019693_6234891_6235861_12_gold standard

```
GTTTCAATCCCGTTGCCGGGATTCATTATAATGAAAG   TCCCATGAGCATCACCAATATCCCAAGGATTAAATG
GTTTCAATCCCGCTGCCGGGATTCATTGTACTGAGAG   TCTGCTGATTCCCGCTATTACTCAGTCGGGGAAGACTAG
GTTTCAATCCCGTTGCCGGGATTCATTGTAATGAAAG   TCATACAAAGCATCACTTATGTATTGTTCTGAAAA
GTTTCAATCCCGTTGCCGGGATTCATTCTAATGAAAG   GAGGGCTTTTCAGGCTCCGTTTTGTGCGTGGCGATCG
GTTTTAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CAGCCCATTCTTTGGCTGTAAAGGGTATTTGGAGAA
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CTAAATAAACCAATCAGAGGGTTAATCATGACTTAT
GTTTCAATCCCGTTGCTGGGATTCATTGTACTGAAAG   CGCTTGTTGCGTGGCGATCGCACCGTGCGATCGTTTGGT
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   GAGCAATAAAAAGGAAGGGTGGTTAGTTCGTATTAA
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CGATCGCTCCGTCGAATCTTTATCTGTTGCCGGGATTCATTGTACTGAAAGCTCAACCGGCCAAATCCTCACCGACTCGATTCAGTT
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   TCTATGCGGGGGAAATCTGGGAGGTGCTGAAACAGTGG
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG   CTGGGCGACAATCAGCCGGATTATGGTTTAAGTGGCT
GTTTCCATCCCGTTGCCGGGATTCATTGTACTGAAAG   TTCACTTGTTCAACGTCAACGACTCACCCCGACT
GTTTCAATCCCGTTGCCGGGATTCATTGTACTGAAAG
```

> *Chlorobium phaeobacteroides* DSM 266_NC_008639_1625359_1633049_115_our results

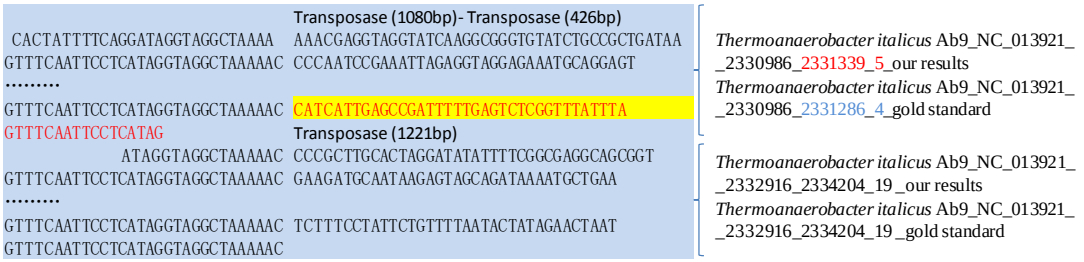
```
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        AGGCTATCCGTATCTGCGCTTTCACGTCTGTAAGCT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        CTGTCATTGATGGCCGATGATTTCCGTGAAATAG
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        AAGGCCAAATCGACCACATCGACGGCGACGGTCT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        GTGATGTCGCGTTCCGGCACGACGCGCTCAAAAA
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        CTGCATCCACGGCAGTTCAGGGGAGGGATTCGGTTAT
GTTTCAATCCACGCTCCCGC  AGGGGGCGCGAC      GGTGTCACTTACGGTTGTGTTTTTTGATCGGAATCGA
  TTCAATCCACGCCGCCCGCCAGGGGGGCTAC       CATTGAATACTACAAAGATGGAACCTCAACTACTC
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        CATGCTTGCCATTACCCTGATCGCAAGCCACCCAA
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        AAAAGGTGATTGTGATCACACACCTGCCAAGCAT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        TCGGATTGTACTCTCTTGCCAAGTGGGTGATCGGCA
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        AGTTGTCCAGGCCTTGCACGACCTCAATCAACGG
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        AGGCTATCCGTATCTGCGCTTTCACGTCTGTAAGCT
.........
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        GCATGGTCGAGAAAGAGTTGCCTGAATCGGCACCTGT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        CAATGCTTCTGTTAAGTTGAGCAAGCGCAATCACTG
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC        CACAGCCCGCCCTTATCAGCAGGCGTTGTTCGAGGC
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC
```

> *Chlorobium phaeobacteroides* DSM 266_NC_008639_1625359_1633049_114 _gold standard

```
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC AGGCTATCCGTATCTGCGCTTTCACGTCTGTAAGCT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC CTGTCATTGATGGCCGATGATTTCCGTGAAATAG
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC AAGGCCAAATCGACCACATCGACGGCGACGGTCT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC GTGATGTCGCGTTCCGGCACGACGCGCTCAAAAA
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC CTGCATCCACGGCAGTTCAGGGGAGGGATTCGGTTAT
GTTTCAATCCACGCTCCCGCAGGGGGCGCGAC GGTGTCACTTACGGTTGTGTTTTTTGATCGGAATCGATTCAATCCACGCCGCCCGCCAGGGGGCGCTACCATTGAATACTACAAAGATGGAACCTCAACTACTC
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC CATGCTTGCCATTACCCTGATCGCAAGCCACCCAA
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC AAAAGGTGATTGTGATCACACACCTGCCAAGCAT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC TCGGATTGTACTCTCTTGCCAAGTGGGTGATCGGCA
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC AGTTGTCCAGGCCTTGCACGACCTCAATCAACGG
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC CTCGTGATCAAGGGTATGCAGGTGAGCGTAGACATC
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC AGGCTATCCGTATCTGCGCTTTCACGTCTGTAAGCT
.........
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC GCATGGTCGAGAAAGAGTTGCCTGAATCGGCACCTGT
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC CAATGCTTCTGTTAAGTTGAGCAAGCGCAATCACTG
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC CACAGCCCGCCCTTATCAGCAGGCGTTGTTCGAGGC
GTTTCAATCCACGCGCCCGCAGGGGGCGCGAC
```

Figure S5. 59 cases with DRs broken by a transposon

Quite a number of CRISPRs acquired transposon insertions, and were broken into two CRISPRs in the CRISPRdb annotations, as in the Figure 1(d). All the 59 cases are demonstrated in the Figure S5.Our curation shows that there are 59 CRISPRs with flanking DRs inserted by transposons, *e.g.* insertion sequence (IS) elements (Siguier, et al., 2006; Zhou, et al., 2008)or miniature inverted repeat transposable elements (MITEs)(Chen, et al., 2008). Figure S5 illustrates that the 1,221-bp IS element inserts into the DR sequence of the CRISPR in the genome *Thermoanaerobacteritalicus* Ab9. The 4-bp tandem duplication ATAG in the DR sequence also supports that this IS copy was recently translocated here. A 180-bp MITE element is also observed to be within the DR sequence of a CRISPR in *Microcystis aeruginosa*NIES-843, and the 5-bp tandem duplication CTATT flanking the MITE should be produced during its recent translocation, as in Figure S5. Summary of all the 59 transposon insertions in CRISPRs may be found in the Figure S5.



GGGACTTAGGCATTTTCGGGCAGCAAAAAAGGCTCAAACCATTACGGGACAAAGGGTTAACCTCGATTTATGATTTTCCTTGATCTATCTGGGTTTCAGCGATTTTG
GGCTTCTCGAAGTAAATCCCAAGCAGGGATTGG GGTTGAGGCTTTTCTCGATTTGTTTTTTGTCTAAGTCCCA (180bp, appears 19 times)

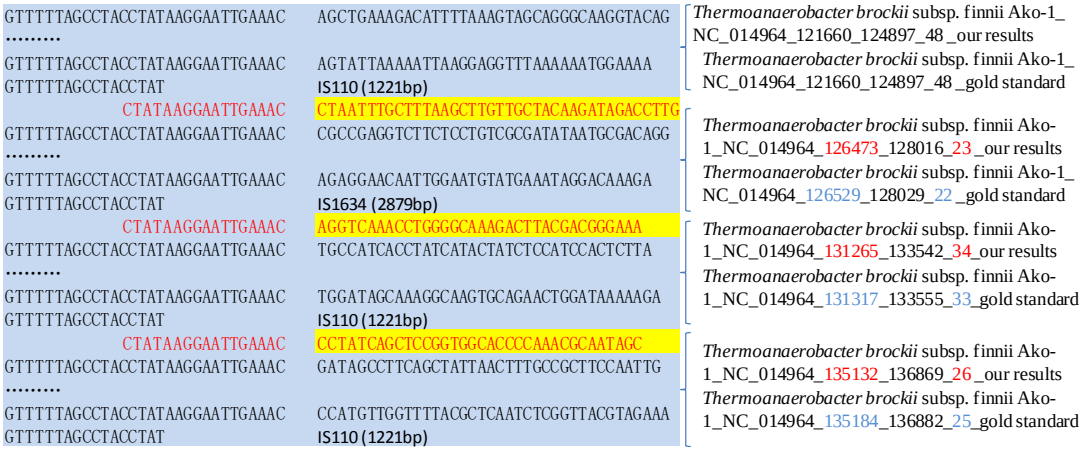
TTGTTATACTCGAAAAAGCTCGGAAGTATAAGAGTTTGTGCTGCCTATTTTTTCGGAAAAATAGGTTGTAATGGGAGACTCTTTGTTTTCCTAGACAAAAATAATTACT
TTTTCAAAAAAACACTTTTCTATTTTTTTGTTGGGTATTTTATTCTCTGGAAGTCCCT(167bp, appears 8 times)

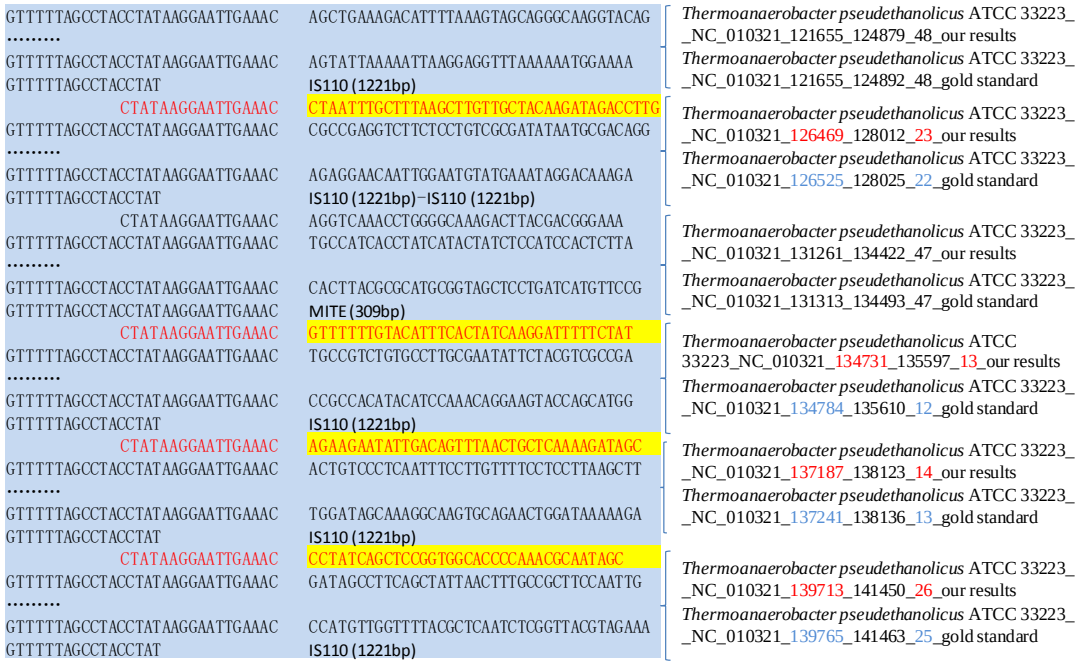

GTTTCAATTCCTCATAGGTACGCTGAGAAC GCATTGGCTTTCATCTCCCCCTCGGCATGCCCGGC
.....
GTTTCAATTCCTCATAGGTACGCTGATAAC GCTCTGGATGCTTGGAAGAAATGCAAGGAACGGAATA
GTTTCAATTCCTCATAG IS110 (2800bp)
AGGTACGCTGAGAAC ACATAATATTCCTTGTATGCATAAACCACGCCGTCAGGA
GTTTCAATTCCTCATAGGTACGCTGAGAAC TCTGCACATTGGGGGCATAAAGCGACGCCCCCAAAACC
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC TTCTCGGCTGGCAGGCTGATTTAGAGGAGGTGTTCTAA
GTTTCAATTCCTCATA IS1634(1701bp)
TCATAGGTACGCTGAAAAC CGGTAGATAGTGCCGACGAGGCTCTT
GTTTCAATTCCTCATAGGTACGCTGAAAAC CGATATTGTTCAAACTTTGCATTCATTTCGTTCCT
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC GAAATATACCATGCGGGGGTGTTTCAGTTCCGGTTCA
GTTTCAATTCCTCATAG IS110 (1215bp)
AGGTACGCTGAAAAC TAAAACTCGTATTCCTTCCCAGTCTTTTTCTTTGGGCA
GTTTCAATTCCTCATAGGTACGCTGAAAAC ACGGAACCACTCCTGCCAATATGCTTCTTTTTCGTTTC
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC CCCGGACAATACCCTCATGATACCGTCCTCGGTTCTTC
GTTTCAATTCCTCATAGGTACGCTGAAAAC
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC TTTGACCACCACCTCATCCTTAGCAAATGCTAACGGA
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC ACATTTTCAATTTTATGTAGATTGAAATATCGTTCTG
GTTTCAATTCCTCATAG MITE (300bp)
AGGTACGCTGAAAAC CGGTGTGTGATTGTATGAAAGATGACGATGACCAGGGC
GTTTCAATTCCTCATAGGTACGCTGAAAAC ATGTGCCGGAGGTGTGCCCGAACTGCGCAGATATCC
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC GCATGGACTGTGAGCGGATGTACGTGCTCTGGGAGTA
GTTTCAATTCCTCATA IS1634(1701bp)
TCATAGGTACGCTGAAAAC GCCGAGAAAATGTTCTCGGTGGCCTGGGATTACGCT
GTTTCAATTCCTCATAGGTACGCTGAAAAC GATGCCCTCCTTTCTTGCGATTGAACGGCAAGAAGTG
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC CGGACCGCTTCAACCCCAGGCAGGAGGTTTTTGACCTG
GTTTCAATTCCTCATAG MITE (300bp)
AGGTACGCTGAAAAC GACTGGCAGAGATTCGGGATGTCGTGAGGCAGTATC
.....
GTTTCAATTCCTCATAGGTACGCTGAAAAC GAGGGAGAATGCCGCATGGTCGGGTCGAGTTCCGCGG
GTTTCAATTCCTCATAGGTACGCTGAAAAC
Thermacetogenium phaeum DSM 12270_NC_018870_ _2439391_2440133_11_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _2439391_2440146_11_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _2443573_2445992_36_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _2443641_2446006_35_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2447892_2448160_4 _our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2447944_2448173_3_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2449719_2451060_20_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2449704_2451060_20_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _2453564_2455390_27_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _2453564_2455337_26_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2455691_2458143_37_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2455676_2458157_37_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _2460043_2463456_51_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _2460106_2463402_49_gold standard
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2463757_2464503_11_our results
Thermacetogenium phaeum DSM 12270_NC_018870_ _ 2463742_2464503_11_gold standard

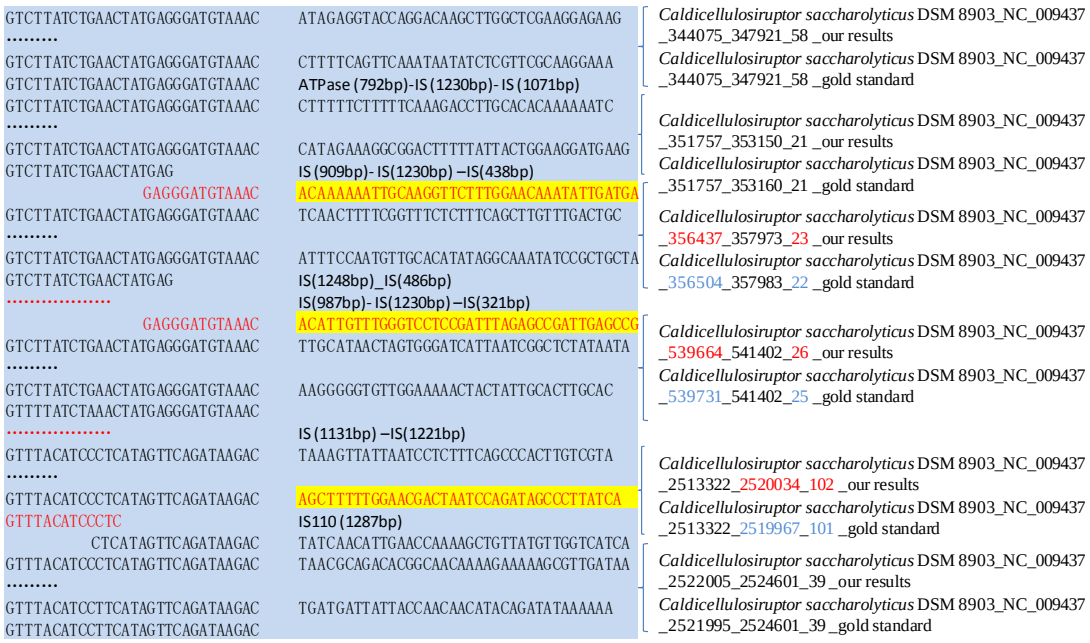
ACATTTGCGCAATATCTCATCGCAAATGTTTCTGTTCAGAGCTTGCTGG（49bp， appears 3 times ）

ACCCAGCCATCATATAATCGGCTGGGTTTCTGGTCCGAGCTTGCTGG（47bp， appears 4 times ）


GTTTCAATTACCGTAGGTACTATCAGAAC
TTTCAAAGCCTTAAAGGTACTATCAGAAC
.....
GTTTCAATTCCTTATAGGTACTATCAGAAC
TTCAATTCCTTATAGGTACTATCAGAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAG
ATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAG
ATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAG
ATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAG
ATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAG
ATAGGTAAGATCAAA
GTTTCAATTCCTCATAGGTACGATCAAAAC
.....
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAG
ATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
GTTTCAATTCCTCATAGGTACGATCAAAAC
TCATATAGGTGCATTCCTTCATTCGACATAATCCGCAAA
TTGGACAAGCTAAAGCTAGGGGCTTTACGGCGACATCCTA
CGATACGGGCGATGTCATAGTGATACTTCGCCCAGATGAGA
IS110(1105bp)
TTCACACTTCGGCATGTCGCGCCCCTCGGACTTACGAC
TATAGTGCTCGGGTTCAAGGACGCCGGAATATACCGG
IS110(1227bp)
GGAGCCTAAACCATGCGTCAATATCGCGGACGAAT
GAGCAGCCGGCGCCGCAATCCGATGCACAGACGCAG
ACTCCTGCAGCGCCTCGAGGCGCGCCAGCCGACTTT
IS110(1227bp)
CATCAGCGCGCCGACAACGGCGCCCACGTCGACACCG
GCAAACGGAACGCGTGCCGGCGCCGCCGTGGTTTGT
CGGGACGATCTATGTCATGTTCTCGGCGGTGGAAGT
IS110(1227bp)
AAAGGAATGGGCAGACATCGTGCTGTTCCTGAATTA
CATGTCGAGCTCCACCATGGAGAGCCCGATCCGCCGCGC
CTGTATCGTATATTCAGTTCCTGATCCTGCGAAGCATT
IS110(1228bp)
GTTACGGCTGTGCTCCACTAGCCGCTTCCACGCTTC
GTGATACTGTCCGAATACTTCGTGACCGTCATATCCGT
GTGAGCCATCAGAAATATCTCTCGTTCCCTATCCGT
IS(1293bp)-IS(1278bp)-IS110 (1224bp)
GATCGAAAAGCGAGAAACGATTGAGTTGAACGTTGACGA
GAAGGAGTTTACCGTCGCCGTCATAACGCTGGAGT
TGAGGTGGTTGTGAATTATAGAAGCCTGAAAACAGG
IS110(1227bp)
AAACCTACATCCGCATGGACTACACTGGTACAGATCGCTAC
GATGCCCTGAAGTTCCCAGACGGCCTGATACGCCA
TCAGTTTGGCCATGTATGTGATCTTCGGCATTTCGT
IS110(1225bp)
ACCGATGCCGAGAACATGGAGCTAAGAATCCGTATG
TTTGCATGTACATGGGCAGGCGCGGAACAGAGATCG
GGCCACGATCTCACCCACGCTGTCGTCACCGGTGT
Thermobacillus composti KWC4_NC_019897_ _3270701_3271073_5 _our results
Thermobacillus composti KWC4_NC_019897_ _3270768_3271074_4 _gold standard
Thermobacillus composti KWC4_NC_019897_ _3272406_3273774_20_our results
Thermobacillus composti KWC4_NC_019897_ _3272406_3273720_19 _gold standard
Thermobacillus composti KWC4_NC_019897_ _3275324_3277594_34 _our results
Thermobacillus composti KWC4_NC_019897_ _3275311_3277541_33_gold standard
Thermobacillus composti KWC4_NC_019897_ _3279144_3280208_16 _our results
Thermobacillus composti KWC4_NC_019897_ _3279131_3280155_15 _gold standard
Thermobacillus composti KWC4_NC_019897_ _3281758_3283032_19 _our results
Thermobacillus composti KWC4_NC_019897_ _3281745_3282977_18 _gold standard
Thermobacillus composti KWC4_NC_019897_ _3284583_3285001_6_our results
Thermobacillus composti KWC4_NC_019897_ _3284570_3285001_6 _gold standard
Thermobacillus composti KWC4_NC_019897_ _3298570_3299110_8_our results
Thermobacillus composti KWC4_NC_019897_ _3298557_3299057_7_gold standard
Thermobacillus composti KWC4_NC_019897_ _3300646_3300996_5_our results
Thermobacillus composti KWC4_NC_019897_ _3300647_3300943_4_gold standard
Thermobacillus composti KWC4_NC_019897_ _3302548_3302761_3_our results
Thermobacillus composti KWC4_NC_019897_ _3302535_3302761_3_gold standard


GTTTTTAGCCTACCTATAAGGAATTGAAAC
.....
GTTTTTAGCCTACCTATAAGGAATTGAAAC
GTTTTTAGCCTACCTAT
CTATAAGGAATTGAAAC
GTTTTTAGCCTACCTATAAGGAATTGAAAC
.....
GTTTTTAGCCTACCTATAAGGAATTGAAAC
GTTTTTAGCCTACCTAT
CTATAAGGAATTGAAAC
GTTTTTAGCCTACCTATAAGGAATTGAAAC
.....
GTTTTTAGCCTACCTATAAGGAATTGAAAC
GTTTTTAGCCTACCTAT
CTATAAGGAATTGAAAC
GTTTTTAGCCTACCTATAAGGAATTGAAAC
.....
GTTTTTAGCCTACCTATAAGGAATTGAAAC
GTTTTTAGCCTACCTAT
AGCTGAAAGACATTTTAAAGTAGCAGGGCAAGGTACAG
AGTATTAAAAATTAAGGAGGTTTAAAAAATGGAAAA
IS110 (1221bp)
CTAATTTGCTTTAAGCTTGTTGCTACAAGATAGACCTTG
CGCCGAGGTCTTCTCCTGTCGCGATATAATGCGACAGG
AGAGGAACAATTGGAATGTATGAAATAGGACAAAGA
IS1634 (2879bp)
AGGTCAAACCTGGGGCAAAGACTTACGACGGGAAA
TGCCATCACCTATCATACTATCTCCATCCACTCTTA
TGGATAGCAAAGGCAAGTGCAGAACTGGATAAAAAGA
IS110 (1221bp)
CCTATCAGCTCCGGTGGCACCCCAAACGCAATAGC
GATAGCCTTCAGCTATTAACTTTGCCGCTTCCAATTG
CCATGTTGGTTTTACGCTCAATCTCGGTTACGTAGAAA
IS110 (1221bp)
Thermoanaerobacter brockii subsp. finnii Ako-1_ NC_014964_121660_124897_48 _our results
Thermoanaerobacter brockii subsp. finnii Ako-1_ NC_014964_121660_124897_48 _gold standard
Thermoanaerobacter brockii subsp. finnii Ako-1_NC_014964_126473_128016_23 _our results
Thermoanaerobacter brockii subsp. finnii Ako-1_ NC_014964_126529_128029_22 _gold standard
Thermoanaerobacter brockii subsp. finnii Ako-1_NC_014964_131265_133542_34_our results
Thermoanaerobacter brockii subsp. finnii Ako-1_NC_014964_131317_133555_33_gold standard
Thermoanaerobacter brockii subsp. finnii Ako-1_NC_014964_135132_136869_26 _our results
Thermoanaerobacter brockii subsp. finnii Ako-1_NC_014964_135184_136882_25_gold standard


GTTTTTAGCCTACCTATAAGGAATTGAAAC AGCTGAAAGACATTTTAAAGTAGCAGGGCAAGGTACAG
GTTTTTAGCCTACCTATAAGGAATTGAAAC AGTATTAAAAATTAAGGAGGTTTAAAAAATGGAAAA
GTTTTTAGCCTACCTAT IS110 (1221bp)
CTATAAGGAATTGAAAC CTAATTTGCTTTAAGCTTGTTGCTACAAGATAGACCTTG
GTTTTTAGCCTACCTATAAGGAATTGAAAC CGCCGAGGTCTTCTCCTGTCGCGATATAATGCGACAGG
GTTTTTAGCCTACCTATAAGGAATTGAAAC AGAGGAACAATTGGAATGTATGAAATAGGACAAAGA
GTTTTTAGCCTACCTAT IS110 (1221bp)−IS110 (1221bp)
CTATAAGGAATTGAAAC AGGTCAAACCTGGGGCAAAGACTTACGACGGGAAA
GTTTTTAGCCTACCTATAAGGAATTGAAAC TGCCATCACCTATCATACTATCTCCATCCACTCTTA
GTTTTTAGCCTACCTATAAGGAATTGAAAC CACTTACGCGCATGCGGTAGCTCCTGATCATGTTCCG
GTTTTTAGCCTACCTATAAGGAATTGAAAC MITE (309bp)
CTATAAGGAATTGAAAC GTTTTTTGTACATTTCACTATCAAGGATTTTTCTAT
GTTTTTAGCCTACCTATAAGGAATTGAAAC TGCCGTCTGTGCCTTGCGAATATTCTACGTCGCCGA
GTTTTTAGCCTACCTATAAGGAATTGAAAC CCGCCACATACATCCAAACAGGAAGTACCAGCATGG
GTTTTTAGCCTACCTAT IS110 (1221bp)
CTATAAGGAATTGAAAC AGAAGAATATTGACAGTTTAACTGCTCAAAAGATAGC
GTTTTTAGCCTACCTATAAGGAATTGAAAC ACTGTCCCTCAATTTCCTTGTTTTCCTCCTTAAGCTT
GTTTTTAGCCTACCTATAAGGAATTGAAAC TGGATAGCAAAGGCAAGTGCAGAACTGGATAAAAAGA
GTTTTTAGCCTACCTAT IS110 (1221bp)
CTATAAGGAATTGAAAC CCTATCAGCTCCGGTGGCACCCCAAACGCAATAGC
GTTTTTAGCCTACCTATAAGGAATTGAAAC GATAGCCTTCAGCTATTAACTTTGCCGCTTCCAATTG
GTTTTTAGCCTACCTATAAGGAATTGAAAC CCATGTTGGTTTTACGCTCAATCTCGGTTACGTAGAAA
GTTTTTAGCCTACCTAT IS110 (1221bp)
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_121655_124879_48_our results
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_121655_124892_48_gold standard
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_126469_128012_23_our results
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_126525_128025_22_gold standard
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_131261_134422_47_our results
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_131313_134493_47_gold standard
Thermoanaerobacter pseudethanolicus ATCC 33223_NC_010321_134731_135597_13_our results
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_134784_135610_12_gold standard
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_137187_138123_14_our results
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_137241_138136_13_gold standard
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_139713_141450_26_our results
Thermoanaerobacter pseudethanolicus ATCC 33223_ _NC_010321_139765_141463_25_gold standard


GTCTTATCTGAACTATGAGGGATGTAAAC ATAGAGGTACCAGGACAAGCTTGGCTCGAAGGAGAAG
GTCTTATCTGAACTATGAGGGATGTAAAC CTTTTCAGTTCAAATAATATCTCGTTCGCAAGGAAA
GTCTTATCTGAACTATGAGGGATGTAAAC ATPase (792bp)-IS (1230bp)- IS (1071bp)
GTCTTATCTGAACTATGAGGGATGTAAAC CTTTTTCTTTTTCAAAGACCTTGCACACAAAAAATC
GTCTTATCTGAACTATGAGGGATGTAAAC CATAGAAAGGCGGACTTTTTATTACTGGAAGGATGAAG
GTCTTATCTGAACTATGAG IS (909bp)- IS(1230bp) −IS(438bp)
GAGGGATGTAAAC ACAAAAAATTGCAAGGTTCTTTGGAACAAATATTGATGA
GTCTTATCTGAACTATGAGGGATGTAAAC TCAACTTTTCGGTTTCTCTTTCAGCTTGTTTGACTGC
GTCTTATCTGAACTATGAGGGATGTAAAC ATTTCCAATGTTGCACATATAGGCAAATATCCGCTGCTA
GTCTTATCTGAACTATGAG IS(1248bp)_IS(486bp)
IS(987bp)- IS(1230bp) −IS(321bp)
GAGGGATGTAAAC ACATTGTTTGGGTCCTCCGATTTAGAGCCGATTGAGCCG
GTCTTATCTGAACTATGAGGGATGTAAAC TTGCATAACTAGTGGGATCATTAATCGGCTCTATAATA
GTCTTATCTGAACTATGAGGGATGTAAAC AAGGGGGTGTTGGAAAAACTACTATTGCACTTGCAC
GTTTTATCTAAACTATGAGGGATGTAAAC
IS (1131bp) −IS(1221bp)
GTTTACATCCCTCATAGTTCAGATAAGAC TAAAGTTATTAATCCTCTTTCAGCCCACTTGTCGTA
GTTTACATCCCTCATAGTTCAGATAAGAC AGCTTTTTGGAACGACTAATCCAGATAGCCCTTATCA
GTTTACATCCCTC IS110 (1287bp)
CTCATAGTTCAGATAAGAC TATCAACATTGAACCAAAAGCTGTTATGTTGGTCATCA
GTTTACATCCCTCATAGTTCAGATAAGAC TAACGCAGACACGGCAACAAAAGAAAAAGCGTTGATAA
GTTTACATCCTTCATAGTTCAGATAAGAC TGATGATTATTACCAACAACATACAGATATAAAAAA
GTTTACATCCTTCATAGTTCAGATAAGAC
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _344075_347921_58 _our results
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _344075_347921_58 _gold standard
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _351757_353150_21 _our results
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _351757_353160_21 _gold standard
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _356437_357973_23 _our results
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _356504_357983_22 _gold standard
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _539664_541402_26 _our results
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _539731_541402_25 _gold standard
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _2513322_2520034_102 _our results
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _2513322_2519967_101 _gold standard
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _2522005_2524601_39 _our results
Caldicellulosiruptor saccharolyticus DSM 8903_NC_009437 _2521995_2524601_39 _gold standard

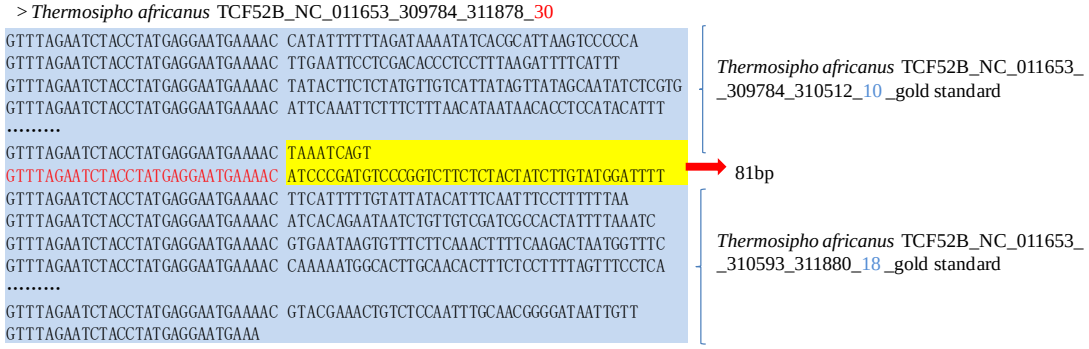

ATAGGTAGGCTAAAAAC CAGCATTACCGCTTGTAAAAGTATTTGTATCTAATTTTAA
TTTCAATCCCTTATAGGTAAGCTAAAAAC AAATTTTTGCAGGGGAGGAGACATAAAACGAAATGAA
GTTTCAATCCCTTATAGGTAAGCTAAAAAC CCACTCCCGCCCAAATTCAAAACCTCATTTTTTCCTT
GTTTCAATCCCTTATAGGTAAGCTAAAAAC IS110 (1220bp)
ATAGGTAGGCTAAAAAC TAGAGAGGGTAAAAGCGGCATGGCAAGTTTTTGAAA
GTTTCAATTCCTTATAGGTAGGCTAAAAAC TTAAAGGAGGAAGATAAAAATGATAACTTCTGTAAA
GTTTCAATTCCTTATAGGTAGGCTAAAAAC GAAAACTAACCCGCAGCCGATGCGGGGAGATAAAAAAA
GTTTCAATTCCTTATAG IS110 (1221bp)
ATAGGTAAGCTAAAAAC AATGAAATTCAAATCTTCAACAGTAAAGATTTTGGC
GTTTCAATTCCTTATAGGTAGGCTAAAAAC CACATACAGGACATTCCCACACCCGCTCCGATAACT
GTTTCAATTCCTTATAGGTAGGCTAAAAAC AACCAAACTTGAAAAAAGCTGCATTTTTCAAGTTTA
GTTTCAATTCCTTATAG IS110 (1221bp)
ATAGGTAGGCTAAAAAC CTTTTCTGACCACCAAAAACCTATGTGGCTGTTCTG
GTTTCAATTCCTTATAGGTAGGCTAAAAAC GATTCTACTGCATTATTACATCTTGTGCTACATCA
GTTTCAATTCCTTATAGGTAGGCTAAAAAC TTTAAAATCGTATAAAGAGCTTATTGAGCTGTATTT
GTTTCAATTCCTTATAGGTAGGCTAAAAAC
Thermoanaerobacter sp. X513_NC_014538_ _2149198_2149615_6_our results
Thermoanaerobacter sp. X513_NC_014538_ _2149254_2149615_5_gold standard
Thermoanaerobacter sp. X513_NC_014538_ _2340512_2341253_11_our results
Thermoanaerobacter sp. X513_NC_014538_ _2340499_2341198_10_gold standard
Thermoanaerobacter sp. X513_NC_014538_ _2342843_2347020_63_our results
Thermoanaerobacter sp. X513_NC_014538_ _2342830_2346967_62_gold standard
Thermoanaerobacter sp. X513_NC_014538_ _2348597_2362246_205_our results
Thermoanaerobacter sp. X513_NC_014538_ _2348597_2362246_205_gold standard
GTTTTTAGCTTACCTATAAGGGATTGAAAC AAGGAAAAAATGAGGTTTTGAATTTGGGCGGGAGTGG
GTTTTTAGCTTACCTATAAGGGATTGAAA TTAAAATTAGATACAAATACTTTTACAAGCGGTAATGCTG
GTTTTTAGCCTACCTAT IS110 (1220bp)
ATAGGTAGGCTAAAAAC TAGAGAGGGTAAAAGCGGCATGGCAAGTTTTTGAAA
GTTTCAATTCCTTATAGGTAGGCTAAAAAC TTAAAGGAGGAAGATAAAAATGATAACTTCTGTAAA
GTTTCAATTCCTTATAGGTAGGCTAAAAAC GAAAACTAACCCGCAGCCGATGCGGGGAGATAAAAAAA
GTTTCAATTCCTTATAG IS110 (1221bp)
ATAGGTAAGCTAAAAAC AATGAAATTCAAATCTTCAACAGTAAAGATTTTGGC
GTTTCAATTCCTTATAGGTAGGCTAAAAAC CACATACAGGACATTCCCACACCCGCTCCGATAACT
GTTTCAATTCCTTATAGGTAGGCTAAAAAC AACCAAACTTGAAAAAAGCTGCATTTTTCAAGTTTA
GTTTCAATTCCTTATAG IS110 (1221bp)
ATAGGTAGGCTAAAAAC CTTTTCTGACCACCAAAAACCTATGTGGCTGTTCTG
GTTTCAATTCCTTATAGGTAGGCTAAAAAC GATTCTACTGCATTATTACATCTTGTGCTACATCA
GTTTCAATTCCTTATAGGTAGGCTAAAAAC TTTAAAATCGTATAAAGAGCTTATTGAGCTGTATTT
GTTTCAATTCCTTATAGGTAGGCTAAAAAC
Thermoanaerobacter sp. X514_ NC_010320_ _764967_765384_6 _our results
Thermoanaerobacter sp. X514 _NC_010320_ _764967_765328_5 _gold standard
Thermoanaerobacter sp. X514_ NC_010320_ _2340417_2341158_11 _our results
Thermoanaerobacter sp. X514_ NC_010320_ _2340404_2341103_10 _gold standard
Thermoanaerobacter sp. X514_ NC_010320_ _2342748_2346925_63 _our results
Thermoanaerobacter sp. X514_ NC_010320_ _2342735_2346872_62 _gold standard
Thermoanaerobacter sp. X514_ NC_010320_ _2348515_2363082_219 _our results
Thermoanaerobacter sp. X514_ NC_010320_ _2348502_2363082_219 _gold standard

Figure S6. 15 CRISPRs broken by undetected DRs inside

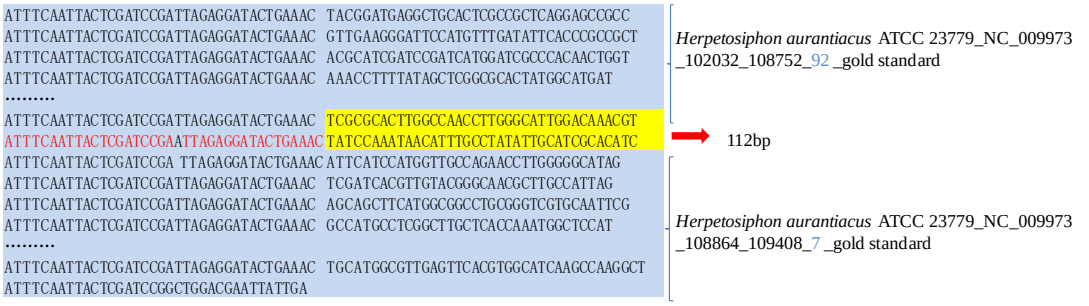
Some DRs were not detected in the database CRISPRdb, so that a long CRISPR may be annotated as two neighboring ones with almost identical DRs. 4 CRISPRs have a full DR copy that were not detected in the database CRISPRdb. The representative example is found in the deltaproteobacterium *Myxococcus fulvus* HW-1. CRISPRdbannotates two neighboring CRISPRs NC_015711_2680594_2682129 and NC_015711_2682223_2687985, with 21 and 80 spacers, respectively. These two CRISPRs have the same DR sequence (GTCGCTCCCCGTGAACGCGGGGAGCGTGGGTTGAAAC), and a 94-bp gap in between, as demonstrated in Figure S6. But there is an identical DR copy in the 94-bp gap, which is not detected in the database CRISPRdb due to an unknown reason.The sequence between this DR copy and the annotated CRISPR NC_015711_2680594_2682129 identically matches to three spacers in the same genome. A CRISPR spacer is supposed to be acquired from foreign invasive elements(Sorek, et al., 2008), and the data suggests that the microbial defense system CRISPR has

generated four spacers to respond to this foreign element.The shared Cas genes further support that NC_015711_2680594_2682129 and NC_015711_2682223_2687985may be joined by the 94-bp gap as one longer CRISPR.Three other similar cases were detected in the bacteria*Caldicellulosiruptor obsidiansis* OB47, *Thermosipho africanus* TCF52B and *Herpetosiphon aurantiacus* ATCC 23779, as shown in the Figure S6.These DRs were missed mainly due to their short lengths slightly below the threshold of spacer/DR ∈[0.6, 2.5].11 other CRISPRs were broken mainly due to an internal partial DR copy that was not detected by the database CRISPRdb, as shown in the Figure S6.

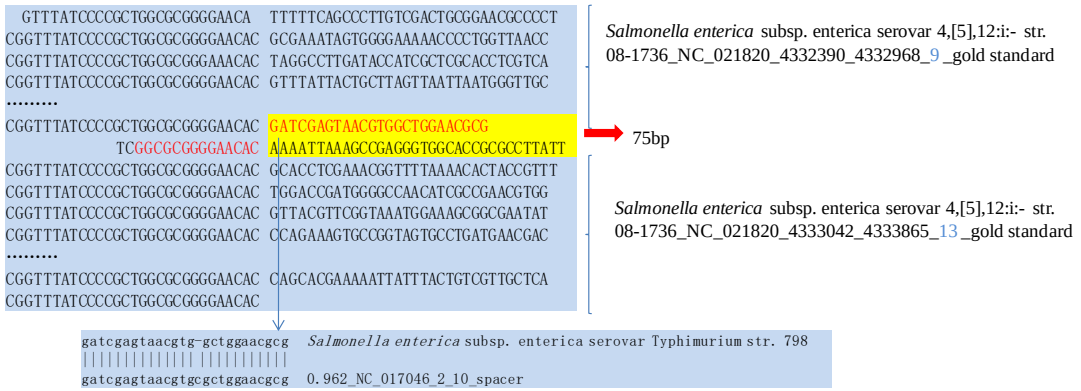
> *Thermosipho africanus* TCF52B_NC_011653_309784_311878_30



> *Herpetosiphon aurantiacus* ATCC 23779_NC_009973_102032_109407_101



> *Salmonella enterica* subsp. enterica serovar 4,[5],12:i:- str. 08-1736_NC_021820_4332392_4333865_24



>*Salmonella enterica* subsp. enterica serovar Heidelberg str. 41578_NC_021810_6276_7936_27

>*Salmonella enterica* subsp. enterica serovar Heidelberg str. B182_NC_017623_3740803_3742463_27

>*Salmonella enterica* subsp. enterica serovar Heidelberg str. SL476_NC_011083_3051219_3052879_27

>*Salmonella enterica* subsp. enterica serovar Typhimurium str. 14028S_NC_016856_3096850_3098323_24

>*Salmonella enterica* subsp. enterica serovar Typhimurium str. D23580_NC_016854_3069600_3071012_23

>*Salmonella enterica* subsp. enterica serovar Heidelberg str. CFSAN002069_NC_021812_3668815_3670469_26

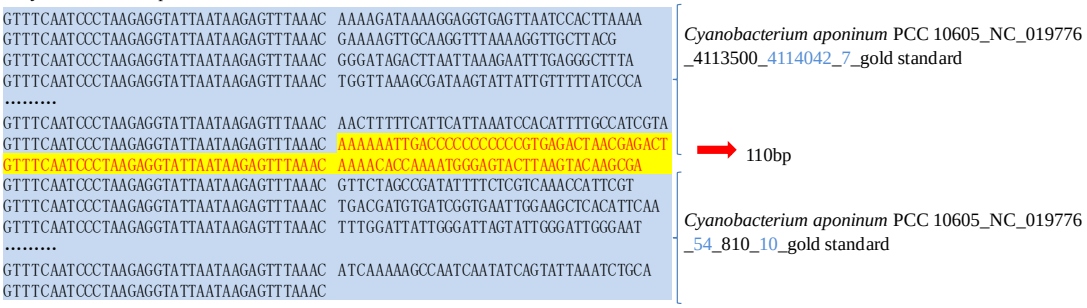
>*Salmonella enterica* subsp. enterica serovar Typhimurium var. 5_str. CFSAN001921-NC_021814_490784_492562_28

>*Salmonella typhimurium* LT2_NC_003197_3076613_3078147_25

Figure S7. 8 CRISPRs broken at the beginning of circular chromosomes

Unlike the eukaryotic counterparts, most of the microbial chromosomesare in the circular shape(Duggin, et al., 2008), but the database CRISPRdb regards a CRISPR spanning the beginning point of a circular chromosome as two.We manually checked the 4,065 CRISPRs annotated in the database CRISPRdb, and detected 8 such cases. Figure S7 shows two CRISPRs NC_022084_2214800_2215162 andNC_022084_29_2779 in the archaea *Thermococcus litoralis* DSM 5473, with 5 and 40 spacers, respectively.The identical DRs and the shared Cas genes suggest that these two closely located CRISPRs may be joined into one by the 38-bp sequence between them. This updated information is important, since the missing spacer may be a key anti-invasion factor. 7 other cases were detected in the database CRISPRdb, as demonstrated in the Figure S7.

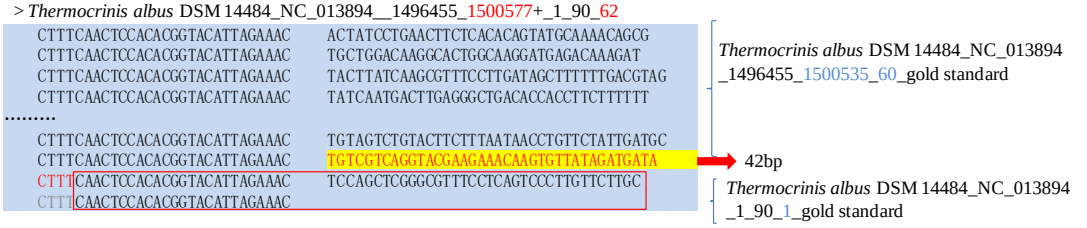
> *Cyanobacterium aponinum* PCC 10605_NC_019776_4113500_4114099+1_810_19



> *Thermococcus sp.* CL1_NC_018015_1949366_1950313+1_682_24

> *Candidatus Puniceispirillum marinum* IMCC1322_NC_014010_2753369_2753527+1_1792_29

> *Thermocrinis albus* DSM 14484_NC_013894__1496455_1500577+_1_90_62



> *Fervidicoccus fontis* Kam940_NC_017461_ 1319183_1319206 +1_794_13

Description of the updated CRISPR definition

The three previous CRISPR annotation programs do not have a consensus agreement on the range of a DR length. In the default settings, CRT, CRISPRFinder and PILER-CR assume that a DR is at least 19, 23 and 16 bps, respectively.But the Cyanobacterium *Microcystis aeruginosa* NIES 843 and the Firmicutes *Thermacetogenium phaeum* DSM 12270 have a CRISPR with the minimum DR lengths 14 and 15, respectively. The program CRT requires a DR to be at least 19 bps in length, which will miss CRISPRs with a short 17-bp DR in the 7 *Firmicutes* and an *Actinobacteria* genomes.The maximum DR length observed in the curated CRISPR annotations is 55 bps. So the program PILER-CR's default value for this feature 64 bps is not strictly supported by the observations. CRT requires the maximum CRISPR DR to be at most 38 bps, which will not

recognize CRISPRs in the 30 bacterial genomes. CRISPRFinder has the same setting with caCRISPR for the maximum DR length 55 bps.

This study proposes the range of a spacer length in two measurements, *i.e*. 9-95 bps and 0.3-2.5 DRs. The program CRT assumes a spacer to be at least 19 bps in length, which will miss CRISPRs in the four Archaea Crenarchaeota genomes and 21 Bacterial genomes (14/21~66.67% are Proteobacteria).CRISPRs in the 191 and 49 prokaryotic genomes will not be recognized by the programs CRT and PILER-CR due to their assumptions of the maximum spacer lengths 48 and 64 bps, respectively. CRISPRFinder has the same requirement as caCRISPR for the maximum spacer length as 2.5 DRs, but its assumption of a minimum spacer length 0.6 DR will miss a CRISPR with the minimum spacer/DR ratio 0.594 and CRISPRs in the 15 bacterial genomes. So the data suggests that the spacer length in two measurements will provide higher specificity and cover all the known CRISPRs.

References

Chen, Y., *et al.* (2008) A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens Rf4, *Genetics*, **179**, 2291-2297.

Deveau, H., Garneau, J.E. and Moineau, S. (2010) CRISPR/Cas system and its role in phage-bacteria interactions, *Annual review of microbiology*, **64**, 475-493.

Duggin, I.G., *et al.* (2008) The replication fork trap and termination of chromosome replication, *Molecular microbiology*, **70**, 1323-1333.

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

Grissa, I., Vergnaud, G. and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats, *Nucleic acids research*, **35**, W52-57.

Lawrence, C.M. and White, M.F. (2011) Recognition of archaeal CRISPR RNA: No P in the alindromic repeat?, *Structure*, **19**, 142-144.

Liu, T. and Huang, J. (2014) Quality control of homologous recombination, *Cellular and molecular life sciences : CMLS*, **71**, 3779-3797.

Siguier, P., Filee, J. and Chandler, M. (2006) Insertion sequences in prokaryotic genomes, *Current opinion in microbiology*, **9**, 526-531.

Sorek, R., Kunin, V. and Hugenholtz, P. (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea, *Nature reviews. Microbiology*, **6**, 181-186.

Wang, X., *et al.* (2009) Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization, *Genome research*, **19**, 1026-1032.
Wheeler, D.L., *et al*. (2008) Database resources of the National Center for Biotechnology Information, *Nucleic acids research*, **36**, D13-21.
Zhou, F., Olman, V. and Xu, Y. (2008) Insertion Sequences show diverse recent activities in Cyanobacteria and Archaea, *BMC genomics*, **9**, 36.