

# **A comprehensive curation demonstrates the dynamic evolutionary patterns of prokaryotic CRISPRs**

Guoqin Mai<sup>1,\*</sup>, Ruiquan Ge<sup>1,2,\*</sup>, Guoquan Sun<sup>1</sup>, Qinghan Meng<sup>1,2</sup>, Fengfeng Zhou<sup>1,#</sup>

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

<sup>2</sup>Shenzhen 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.healthinformatics.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 April 14, 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* PAI 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* PA15 missed one complete DR in their flanking regions too, as in the Figure S2.

> *Gluconacetobacter diazotrophicus* PA15\_NC\_010125\_2181482\_2182177\_10\_our results

AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	CTGGCCGGTAAATTGCGTGACGGCGGGTTC
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	ATCGCATGACCTTTGGTTTCGACCGGGTAT
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	TCGAAGACCGCGCTGGACGACATGGGAAG
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	CGATTCGATACCTTGC GCGTGCGCACTGG
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GCAAGATAACGGCCTTCGGCCACACGAAGA
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GCAGATTTGATACCGGCAACGACGGTCTT
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GAGCGGAGGCAGCTTGTGGCCAATATGGC
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GATCGCGCATGATGCGCGGATCAACGCTCA
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	ATCGAGCGCGGCGGACGTCGTTGTCGCC
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	CATGGTGTGAGCTTGC TCGGCGGTTCTC
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	




> *Gluconacetobacter diazotrophicus* PA15\_NC\_010125\_2181482\_2182111\_9\_gold standard

AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	CTGGCCGGTAAATTGCGTGACGGCGGGTTC
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	ATCGCATGACCTTTGGTTTCGACCGGGTAT
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	TCGAAGACCGCGCTGGACGACATGGGAAG
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	CGATTCGATACCTTGC GCGTGCGCACTGG
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GCAAGATAACGGCCTTCGGCCACACGAAGA
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GCAGATTTGATACCGGCAACGACGGTCTT
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GAGCGGAGGCAGCTTGTGGCCAATATGGC
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	GATCGCGCATGATGCGCGGATCAACGCTCA
AGCCTACCATCGGCAAAATCGGTAGGGAAACCACGGC	ATCGAGCGCGGCGGACGTCGTTGTCGCC

> *Gluconacetobacter diazotrophicus* PA15\_NC\_010125\_62935\_64899\_29\_our results

GTTTTAATCCCCGCTTCCGCGTGGGGAGCGAC	GACAACCTTGC GACTCCTGTCCGACGCGGTCGGC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GGGTGCGATCCTCCGCATGGTGCAGGAGGACATC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CGCCTCGTCGATGATGACGGGCGAATTCATCAGCC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GATCACGCGCACGCCAGCCACTCGTCGTAGG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CAACTGTCCGATATCTCGGCCAATCTCTCCAACG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CGTTTCTCGATACGCGGTCC AATCCCCTCCTTCTT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CAACGTCTCCGACCCTTGACGTGGTCTCAGCCT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GAGCCGGGACGAGCGTGAATCCTGCTGCGCCG
.....	
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CTTCTGTTCCGCATCGCCGGCGCGTCTGGTCCG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	AAGATCCATGTCTTCCAAGCGCATTCCCTAAAT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GGGAAACGAACCGTGATGTCGTGGCGAAAGAAAG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	




> *Gluconacetobacter diazotrophicus* PA15\_NC\_010125\_62935\_64832\_28\_gold standard

GTTTTAATCCCCGCTTCCGCGTGGGGAGCGAC	GACAACCTTGC GACTCCTGTCCGACGCGGTCGGC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GGGTGCGATCCTCCGCATGGTGCAGGAGGACATC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CGCCTCGTCGATGATGACGGGCGAATTCATCAGCC
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GATCACGCGCACGCCAGCCACTCGTCGTAGG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CAACTGTCCGATATCTCGGCCAATCTCTCCAACG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CGTTTCTCGATACGCGGTCC AATCCCCTCCTTCTT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CAACGTCTCCGACCCTTGACGTGGTCTCAGCCT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	GAGCCGGGACGAGCGTGAATCCTGCTGCGCCG
.....	
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	CTTCTGTTCCGCATCGCCGGCGCGTCTGGTCCG
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	AAGATCCATGTCTTCCAAGCGCATTCCCTAAAT
GTTTCAATCCACGCTCCCGCACAGGGAGCGAC	

> *Gluconacetobacter diazotrophicus* PA15\_NC\_010125\_2253747\_2255112\_22\_our results

AGATTTCATCCCTGCATATGCGAGGAAACAC	AAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTTCATCCCCGACGTGCGGGGAAACACG	TCCCCCTGCTCGGCAACCCCAAATCAAGAT
CGGTTTCATCCCCGACGTGCGGGGAAACAC	GCCGCGACGGTGAAGGGGAGGTTGTTGACGC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	GAGTACCCGATTCCCACTCTCATCTAATATC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	CTTTTCTCAGCCGACACTCTCCATGCAAGAAC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	ACGCGGGTCTATGAGATTCCGATCATGCAGGA
CGGTTTCATCCCCGACGTGCGGGGAAACAC	CATTGACCCGGCACCCCTCCCATGTGACGTTT
CGGTTTCATCCCCGACGTGCGGGGAAACAC	TGGCTCGGACCATCTTAGCGGACGACATGGC
.....	
CGGTTTCATCCCCGACGTGCGGGGAAACAC	TTGCGTCAAACGAAGCAATATCAGGCGCAG
CGGTTTCATCCCCGACGTGCGGGGAAACAC	TTGCCTCTAACGAGGCGCAGTATCAGGCACAG
CGGTTTCATCCCCGACGTGCGGGGAAACAC	CCGTCGTTTCGATCGTACGGGATCATCGGC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	



> *Gluconacetobacter diazotrophicus* PA15\_NC\_010125\_2253747\_2255055\_21\_gold standard

AGATTTCATCCCTGCATATGCGAGGAAACAC	AAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTTCATCCCCGACGTGCGGGGAAACACG	TCCCCCTGCTCGGCAACCCCAAATCAAGAT
CGGTTTCATCCCCGACGTGCGGGGAAACAC	GCCGCGACGGTGAAGGGGAGGTTGTTGACGC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	GAGTACCCGATTCCCACTCTCATCTAATATC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	CTTTTCTCAGCCGACACTCTCCATGCAAGAAC
CGGTTTCATCCCCGACGTGCGGGGAAACAC	ACGCGGGTCTATGAGATTCCGATCATGCAGGA
CGGTTTCATCCCCGACGTGCGGGGAAACAC	CATTGACCCGGCACCCCTCCCATGTGACGTTT
CGGTTTCATCCCCGACGTGCGGGGAAACAC	TGGCTCGGACCATCTTAGCGGACGACATGGC
.....	
CGGTTTCATCCCCGACGTGCGGGGAAACAC	TTGCGTCAAACGAAGCAATATCAGGCGCAG
CGGTTTCATCCCCGACGTGCGGGGAAACAC	TTGCCTCTAACGAGGCGCAGTATCAGGCACAG
CGGTTTCATCCCCGACGTGCGGGGAAACAC	

> *Gluconacetobacter diazotrophicus* PA15\_NC\_011365\_460172\_461964\_29\_our results

AGATTCATCCCTGCATATGCGAGGA	ACACAAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTTCATCCCCGCACGTGCGGGAA	CACGTCCCCCTGCTCGGCAACCCAACTCAAGAT
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGCCGCGACGGTGAAGGGGAGGTGTTCGACGC
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCCTGATCCTGCGCCGACGCCGAACCAGTCCG
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGCCGCGACGGTGAAGGGGAGGTGTTCGACGC
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCCTGATCCTGCGCCGACGCCGAACCAGTCCG
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGAGTCACCGATTCCCATCTCATCTAAATATC
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCTTTTCTCAGCCGACACTCTCCATGCAAGAAC
.....	
CGGTTTCATCCCCGCACGTGCGGGAA	ACACAACGTGGAATGGCACCAGATCGAAACGATGCT
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGCCAAGACCTGCGCCTGGAGCGCCTGCTTGTG
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCGTGGGGAATATCCTGACACACATGGGCGT
CGGTTTCATCCCCGCACGTGCGGGAA	

> *Gluconacetobacter diazotrophicus* PA15\_NC\_011365\_460172\_461907\_28\_gold standard

AGATTCATCCCTGCATATGCGAGGA	ACACAAGGCGGCCTCGTCCAGGATCACCAGACCCTG
CGGTTTCATCCCCGCACGTGCGGGAA	CACGTCCCCCTGCTCGGCAACCCAACTCAAGAT
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGCCGCGACGGTGAAGGGGAGGTGTTCGACGC
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCCTGATCCTGCGCCGACGCCGAACCAGTCCG
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGCCGCGACGGTGAAGGGGAGGTGTTCGACGC
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCCTGATCCTGCGCCGACGCCGAACCAGTCCG
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGAGTCACCGATTCCCATCTCATCTAAATATC
CGGTTTCATCCCCGCACGTGCGGGAA	ACACCTTTTCTCAGCCGACACTCTCCATGCAAGAAC
.....	
CGGTTTCATCCCCGCACGTGCGGGAA	ACACAACGTGGAATGGCACCAGATCGAAACGATGCT
CGGTTTCATCCCCGCACGTGCGGGAA	ACACGCCAAGACCTGCGCCTGGAGCGCCTGCTTGTG
CGGTTTCATCCCCGCACGTGCGGGAA	

> *Gluconacetobacter diazotrophicus* PA15\_NC\_011365\_388303\_388602\_4\_our results

AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	TCAGCCGGCTGGGCATTGCCCTACGAGAGG
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	AGGGGAGCGCTACCTTCGACAAGGGCCT
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	CCTCGTTGCTCGCGCGGGTGCAGGGCTG
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	TCCGCAGTTCACGCGTACCAGCATCGCGG
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	

> *Gluconacetobacter diazotrophicus* PA15\_NC\_011365\_388303\_388536\_3\_gold standard

AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	TCAGCCGGCTGGGCATTGCCCTACGAGAGG
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	AGGGGAGCGCTACCTTCGACAAGGGCCT
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	CCTCGTTGCTCGCGCGGGTGCAGGGCTG
AGCCTACCATCGGCAAAATCGGTAGGGAAACACGGC	

GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	TTTCCGGATTAGAATAATAAGGAACTTCCGGTACGGG
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	GGTATACTAAGCGTACCGTTCGTCTCAACATCGACTA
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	ACACCATATGCTCCGGGCAGAGGAGCATGTTCACT
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	CCAGTCTCTCCTCCTCAGGAGGACTGGCTTTTCAAA
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	CACAATAAATCCCCCTTTGAAAAACGAGTTCAAGC
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	TATGTAGCTGCATATCTACACCTCCACGGATATAAGAG
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	TCTTCTGGCTCTGGCACGGCGAGAATATCTCGCCGT
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	TCCAAATTACCAAGACAAACATGCTTGTCTATGGGTG
.....	
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	GCCCATATCTCTTCGTGATGCTCCGGTCTCTGCTG
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	AGCAGCATCTTTTTCACCTCATGATATCATTATGAC
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	TTTCTATACTGGGGATGTTGACTTCTTGCCTTCCTT
GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	

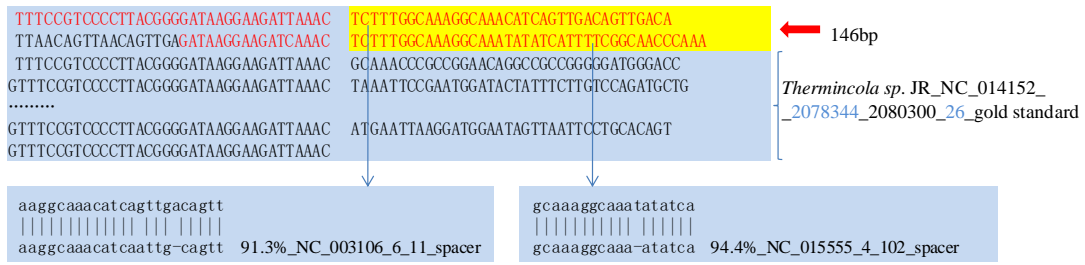
> *Methanosaeta thermophila* PT\_NC\_008553\_670839\_677982\_97\_our results

> *Methanosaeta thermophila* PT\_NC\_008553\_670839\_677982\_96\_gold standard

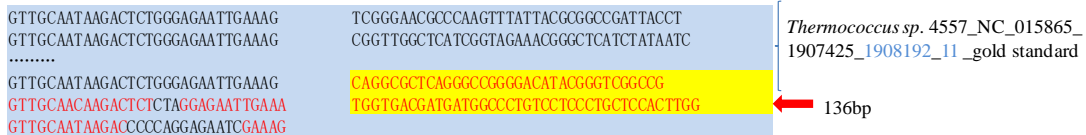
### Figure S3. 4 cases with two new DRs

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.

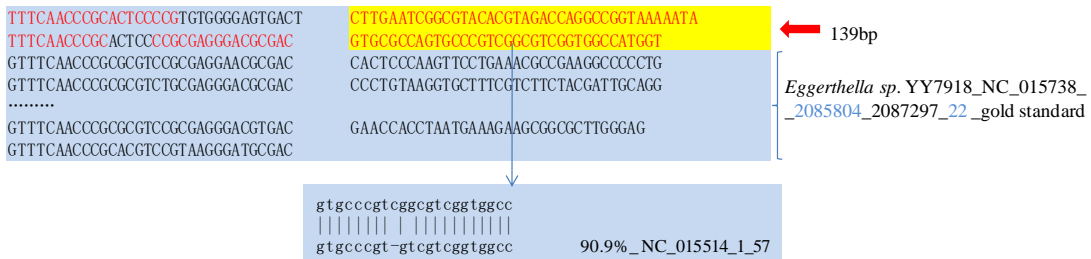
> *Thermincola* sp. JR\_NC\_014152\_2078201\_2080300\_28\_our results



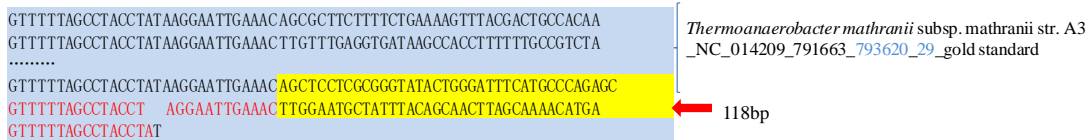
> *Thermococcus* sp. 4557\_NC\_015865\_1907425\_1908328\_13\_our results



> *Eggerthella* sp. YY7918\_NC\_015738\_2085666\_2087297\_24\_our results



> *Thermoanaerobacter mathranii* subsp. mathranii str. A3\_NC\_014209\_791663\_793738\_31\_our results



### 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	TCCATGAGCATCACCAATATCCCAAGGATTAAATG
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAGAG	TCGTCTGATTCCCGCTATTACTCAGTCGGGGAAGACTAG
GTTTCAATCCCGTTGCCGGGATTCATTGTAATGAAAG	TCATACAAAGCATCACTTATGTAATGTTCTGAAAA
GTTTCAATCCCGTTGCCGGGATTCATTCTAATGAAAG	GAGGGCTTTTCAGGCTCCGTTTGTGCGTGGCGATCG
GTTTAAATCCCGTTGCCGGGATTCATTGACTGAAAG	CAGCCCATCTTTGGCTGTAAAGGGTATTTGGAGAA
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	CTAAATAAACCAATCAGAGGGTTAATCATGACTTAT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	CGCTTGTGCGTGGCGATCCACCGTCGATCGTTTGGT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	GAGCAATAAAAAGGAAGGGTGGTTAGTTCGATATA
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	<b>CGATCGCTCCGTCGAATCTTATCT</b>
<b>GTTGCCGGGATTCATTGACTGAAAG</b>	<b>CTCAACCGGCCAAATCTCACCAGCTCGATTAGT</b>
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	TCTATCGGGGGAATCTGGGAGGTCTGAAACAGTGG
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	CTGGGCGACAATCAGCGGATATGGTTTAAAGTGGCT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	TTCACTTGTCAACGCTCAAGACTCACCCCGACT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	

>*Oscillatoria acuminata* PCC 6304\_NC\_019693\_6234891\_6235861\_12\_gold standard

GTTTCAATCCCGTTGCCGGGATTCATTATAATGAAAG	TCCATGAGCATCACCAATATCCCAAGGATTAAATG
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAGAG	TCGTCTGATTCCCGCTATTACTCAGTCGGGGAAGACTAG
GTTTCAATCCCGTTGCCGGGATTCATTGTAATGAAAG	TCATACAAAGCATCACTTATGTAATGTTCTGAAAA
GTTTCAATCCCGTTGCCGGGATTCATTCTAATGAAAG	GAGGGCTTTTCAGGCTCCGTTTGTGCGTGGCGATCG
GTTTAAATCCCGTTGCCGGGATTCATTGACTGAAAG	CAGCCCATCTTTGGCTGTAAAGGGTATTTGGAGAA
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	CTAAATAAACCAATCAGAGGGTTAATCATGACTTAT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	CGCTTGTGCGTGGCGATCCACCGTCGATCGTTTGGT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	GAGCAATAAAAAGGAAGGGTGGTTAGTTCGATATA
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	<b>CGATCGCTCCGTCGAATCTTATCTGTTGCCGGGATTCATTGACTGAAAGTCAACCGGCCAAATCTCACCAGCTCGATTAGT</b>
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	TCTATCGGGGGAATCTGGGAGGTCTGAAACAGTGG
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	CTGGGCGACAATCAGCGGATATGGTTTAAAGTGGCT
GTTTCAATCCCGTTGCCGGGATTCATTGACTGAAAG	TTCACTTGTCAACGCTCAAGACTCACCCCGACT

>*Chlorobium phaeobacteroides* DSM 266\_NC\_008639\_1625359\_1633049\_115\_our results

GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AGGCTATCCGTATCTGGCTTTCACGCTGTAAGCT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CTGTCTGATGCGCGATGATTCCCGTAAATAG
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AAGGCCAAATCGACCATCGACGGGACGGTCT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	GTGATGTCGGTTCGGCAGCAGCGCTCAAAAA
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CTGCATCCACGGCAGTTCAGGGGAGGGATTCGGTTAT
GTTTCAATCCACGGCTCCGCGCAGGGGGCGGAC	<b>GGTGTCACTTACGGTTGTGTTTTTGTGCGGAATGGA</b>
<b>TTCATCCACGGGCCCGCAGGGGGCGGACTAC</b>	<b>CATTGAATACTCAAGATGGAACCTCAACTACTC</b>
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CATGCTTGCATTACCTGATCGCAAGCCACCCAA
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AAAAAGGTGATGTGATCACACAACCTGCCAAGCAT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	TCGGATTGTACTCTTGCACAAGTGGGTGATCGGCA
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AGTTGTCCAGGCTTGCACGACCTCAATCAACGG
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AGGCTATCCGTATCTGGCTTTCACGCTGTAAGCT
.....	
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	GCATGGTCGAGAAAGAGTTCCTGAATCCGACCTGT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CAATGCTTCTGTAAAGTTGAGCAAGCGCAACTACTG
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CACAGCCCGCCTTATCAGCAGGCGTGTTCGAGGC
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	

>*Chlorobium phaeobacteroides* DSM 266\_NC\_008639\_1625359\_1633049\_114\_gold standard

GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AGGCTATCCGTATCTGGCTTTCACGCTGTAAGCT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CTGTCTGATGCGCGATGATTCCCGTAAATAG
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AAGGCCAAATCGACCATCGACGGGACGGTCT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	GTGATGTCGGTTCGGCAGCAGCGCTCAAAAA
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CTGCATCCACGGCAGTTCAGGGGAGGGATTCGGTTAT
GTTTCAATCCACGGCTCCGCGCAGGGGGCGGAC	<b>GGTGTCACTTACGGTTGTGTTTTTGTGCGGAATGGA</b>
<b>TTCATCCACGGGCCCGCAGGGGGCGGACTAC</b>	<b>CATTGAATACTCAAGATGGAACCTCAACTACTC</b>
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CATGCTTGCATTACCTGATCGCAAGCCACCCAA
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AAAAAGGTGATGTGATCACACAACCTGCCAAGCAT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	TCGGATTGTACTCTTGCACAAGTGGGTGATCGGCA
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AGTTGTCCAGGCTTGCACGACCTCAATCAACGG
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CTCTGTATCAAGGATGACAGTGAAGTACATC
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	AGGCTATCCGTATCTGGCTTTCACGCTGTAAGCT
.....	
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	GCATGGTCGAGAAAGAGTTCCTGAATCCGACCTGT
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CAATGCTTCTGTAAAGTTGAGCAAGCGCAACTACTG
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	CACAGCCCGCCTTATCAGCAGGCGTGTTCGAGGC
GTTTCAATCCACGGGCCCGCAGGGGGCGGAC	



>Frankia sp. Ccl3\_NC\_007777\_3904715\_3905896\_16\_our results

```
CTTGGGAACCTGCCGGGGCGGGCGCCCGGGCGTGT  CTTCTGGGTCGGGATATGGTGCTTATCGTGACTC
GTTGTGATCCTCGCGCTGGCCCTGGGGCAGATCCAGC  GCCCAGTTCGGGCATCTTCGC ←
GTTGGGATCCTCGCCGAGGGCGGTCCCTGGGGCTGC  TGGTGGCGAGTGGGCTGCGCCAGCGTCAGGTT
GTTGTGATCCTGCCGAGGGCGGTCCCTGGGGCTGC  GACGACCTGGTGCTGATGTCACCAACGGCACCGCC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CGTCCGCCGACAGCAAAGACCTACACGCCACCCACC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  GTGTCCGCCCTCCGTGAGCTGCTGCTCCGTGGTT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CTGGGACACGGCCGGATCCTGCCAGCGCAACGCC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  GATGGTGACGACCTGGTGCTGATGCAACCAAGC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  TCGATGGTGCCTGGGCTGGTCAACGCCCTGGAT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  AAGTGTGCAACATGTGCTGACTGCATCCGGGT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CCCC GCCGGGCTCACCTGCGCCGCTGGTCTAT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  TCGGGAACACCAAGTGGTGGAGCCAGCGGGTA
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  TCGCCCCCGCTATCTGTCCCGCTTTCTCGGGA
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  GACCGCTGCTTGAGCCGACGGTCAACATGACCC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CCGCTCCGGCTGGATGCGCAGATGTCGCCCTTGGG
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CTGTAGGGGACGTTGTAATGCTTCCGGCCGGAG
```

>Frankia sp. Ccl3\_NC\_007777\_3904716\_3905896\_15\_gold standard

```
CTTGGGAACCTGCCGGGGCGGGCGCCCGGGCGTGT  TTTCTGGGTCGGGATATGGTGCTTATCGTGACTCAGTTGATCCTCGCGCTGCCCTCGCGGAGATCCAGCCAGTTGGCGCATCTGC
GTTGTGATCCTCGCGCTGGCCCTGGGGCAGATCCAGC  TGGTGGCGAGTGGGCTGCGCCAGCGTCAGGTT
GTTGTGATCCTCGCCGAGGGCGGTCCCTGGGGCTGC  GACGACCTGGTGCTGATGTCACCAACGGCACCGCC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CGTCCGCCGACAGCAAAGACCTACACGCCACCCACC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  GTGTCCGCCCTCCGTGAGCTGCTGCTCCGTGGTT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CTGGGACACGGCCGGATCCTGCCAGCGCAACGCC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  GATGGTGACGACCTGGTGCTGATGCAACCAAGC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  TCGATGGTGCCTGGGCTGGTCAACGCCCTGGAT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  AAGTGTGCAACATGTGCTGACTGCATCCGGGT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CCCC GCCGGGCTCACCTGCGCCGCTGGTCTAT
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  TCGGGAACACCAAGTGGTGGAGCCAGCGGGTA
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  TCGCCCCCGCTATCTGTCCCGCTTTCTCGGGA
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  GACCGCTGCTTGAGCCGACGGTCAACATGACCC
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CCGCTCCGGCTGGATGCGCAGATGTCGCCCTTGGG
GTTGTGATCCTCGCCGAGGGCGATCCCTGGGGCTGC  CTGTAGGGGACGTTGTAATGCTTCCGGCCGGAG
```

## 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 *Thermoanaerobacter italicus* 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.

CACTATTTTCAGGATAGGTAGGCTAAAAA GTTTCAATTCCTCATAGGTAGGCTAAAAAC .....	Transposase (1080bp)- Transposase (426bp) AAACGAGGTAGGTATCAAGGCGGGTGTATCTGCCGCTGATAA CCCAATCCGAAATAGAGGTAGGAGAAATGCAGGAGT	<i>Thermoanaerobacter italicus</i> Ab9_NC_013921_2330986_2331339_5_our results <i>Thermoanaerobacter italicus</i> Ab9_NC_013921_2330986_2331286_4_gold standard
GTTTCAATTCCTCATAGGTAGGCTAAAAAC GTTTCAATTCCTCATAGGTAGGCTAAAAAC .....	CATCATTGAGCGGATTTTGGAGTCTCGGTTTATTTA Transposase (1221bp) ATAGGTAGGCTAAAAAC CCCGCTTGCACTAGGATATATTTTCGGCGAGGCAGCGGT GAAGATGCAATAAGAGTAGCAGATAAAATGCTGAA	
GTTTCAATTCCTCATAGGTAGGCTAAAAAC GTTTCAATTCCTCATAGGTAGGCTAAAAAC	TCTTCTATTCTGTTTAACTATAGAATAAT	<i>Thermoanaerobacter italicus</i> Ab9_NC_013921_2332916_2334204_19_our results <i>Thermoanaerobacter italicus</i> Ab9_NC_013921_2332916_2334204_19_gold standard

GTTTCAATTAATCTTAAACCTATTAGGATTGAAAC .....	TTCAGAATTGACACAAATTATTGCTATAAAAAATGATAGC	<i>Microcystis aeruginosa</i> NIES-843_NC_010296_2814769_2822874_112_our results <i>Microcystis aeruginosa</i> NIES-843_NC_010296_2814769_2822884_112_gold standard
GTTTCAATTAATCTTAAACCTATTAGGATTGAAAC GTTTCAATTAATCTTAAACCTATT	TTGTAATTGAGCTTCGTGTACGCCAATTTAATC MITE(180bp) CTATTAGGGATTGAAAC TCCCTGCTTGGTAAATTAAGTATGCCCCATAAGC	
GTTTCAATTAATCTTAAACCTATTAGGATTGAAAC .....	CAGCCGGAATGTAGCAGACAGCCTGGCAGACCAA	<i>Microcystis aeruginosa</i> NIES-843_NC_010296_2823052_2824590_21_our results <i>Microcystis aeruginosa</i> NIES-843_NC_010296_2823103_2824595_20_gold standard
GTTTCAATTAATCTTAAACCTATTAGGATTGAAAC GTTTCAATTAATCTTAAACCTATTAGGGA	TTCTTTGATGATGATGTAACCGGAAGCAATC transposase (513bp)-transposase (675bp)- MITE(167bp) TTAGGGATTGAAAC AGCAATATTTTGGCATCGGATTGGTCAAACT	<i>Microcystis aeruginosa</i> NIES-843_NC_010296_2826180_2829164_41_our results <i>Microcystis aeruginosa</i> NIES-843_NC_010296_2826228_2829166_40_gold standard
GTTTCAATTAATCTTAAACCTATTAGGATTGAAAC .....	CCTAAGACAGTGCCTTTGAGCGATAAGGCGATTATC	
GTTTCAATTAATCTTAAACCTATTAGGATTGAAAC GTTTCAATTAATCTTAAACCTATTAGGATTGAA	CTCACTTAGCTAGTGTTCGCCCTCTCGTAAAAAGCAA	

GGGACTTAGGCATTTTCGGGCAGCAAAAAAGGCTCAAACCATACGGGACAAGGGTTAACCTCGATTATGATTTTCCTTGATCTATCTGGGTTTCAGCGATTTTG  
GGCTTCGGAAGTAAATCCCAAGCAGGGATTGG GGTGAGGCTTTTCGATTTGTTTTGTCTAAGTCCCA (180bp, appears 19 times)

TTGTTACTCGAAAAAGCTCGGAAGTATAAGAGTTTGTGCTGCCTATTTTTCGGAAAAATAGGTTGTAATGGGAGACTTTTGTTCCTAGACAAAAATAACTACT  
TTTTCAAAAAACACTTTTCTATTTTTTTGTTGGGATTTTATTCTCTGGAAGTCCCT(167bp, appears 8 times)

GTCAGAACGACTTCCTGATGAAGAAGGATTAAAGC .....	GCCGCCGCGCTTGTATGCCGCGCATGTACTG	<i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_1466441_1467521_15_our results <i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_1466441_1467533_15_gold standard
GTCAGAACGACTTCCTGATGAAGAAGGATTAAAGC GTCAGAACGACTTCCTGATGAAGA	GGCTGGCCGCGTCCATGCAATCCATGTCACGGC IS66 (927bp)-phage DNA methylase (641bp) GATGAAGAAGGATTAAAGC CTTCATCAGAAATGGCGCTCGTGCATCGGTGGG	
GTCAGAACGACTTCCTGATGAAGAAGGATTAAAGC .....	TGCCCTGACGCGACTGTGAGGTTTCGAC	<i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_1469911_1471052_16_our results <i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_1469966_1471052_15_gold standard
GTCAGAACGACTTCCTGATGAAGAAGGATTAAAGC GTCAGAACGACTTCCTGATGAAGAAGGATTAAAGC .....	TTCGCAATTTCTGAGGATCATTCGTGTTCTTCGT	
TTTCTGAGCTGCCTGCGCGGCAGCGAAC .....	CAGGGATAGGTGATGTTCTATTACGGTTATG	<i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_2520192_2523887_61_our results <i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_2520192_2523887_61_gold standard
TTTCTGAGCTGCCTGCGCGGCAGCGAAC TTTCTGAGCTGCCTGCGCGGCAGCGAAC .....	ATCATCGAGGAAGCCGCAAGGTGCGCAGCCA ISPsy4 (222bp)- istB(171bp) CGCTCAGATATTCGCGGCGGTGATCTGGAT	<i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_2928156_2928605_7_our results <i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_2928217_2928605_6_gold standard
TTTCTGAGCTGCCTGCGCGGCAGCGAAC TTTCTGAGCTGCCTGCGCGGCAGCGAAC TTTCTGAGCTGCCTGCGCGGCAGCGAAC .....	CTGGCCCAAGCCGACGACCCGCGCTGCTACGA IS110(1383bp) GGCGGCAGTCCGGTGGTGGCATTGATGATGA	
TTTCTGAGCTGCCTGCGCGGCAGCGAAC TTTCTGAGCTGCCTACGCGGCAGCGAAC TTTCTGAGCTGCCTGCGCGGCAGCGAAC	GCTGCGGGAACCTCATGAGCATGCCTGGCGA GGGGCGATCAGTCCGGGGTCGAGGATCGA	<i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_2931057_2932046_16_our results <i>Thioalkalivibrio nitratireducens</i> DSM 14787_NC_019902_2931057_2932046_16_gold standard

GTTTCAATTCCATAGGTACGCTGAGAAC .....	GCATTGGCTTTCATCTCCCTCGGCATGCCGGC	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2439391_2440133_11_our results
GTTTCAATTCCATAGGTACGCTGATAAC GTTTCAATTCCATAG	GCTCTGGATGCTTGGAGAAATGCAAGGAACGGAATA IS110 (2800bp)	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2439391_2440146_11_gold standard
AGGTACGCTGAGAAC GTTTCAATTCCATAGGTACGCTGAGAAC .....	ACATAATATTCCTTGTATGCATAAACCCAGCCGTCAGGA TCTGCACATTGGGGCATAAAGCGACGCCCCAAAAACC	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2443573_2445992_36_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATATA	TTCTCGGCTGGCAGGCTGATTTAGAGGAGGTGTTCTAA IS1634(1701bp)	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2443641_2446006_35_gold standard
TCATAGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC .....	CGGTAGATAGTGCCGACGAGGCTCTT CGATAATTGTTCAAACCTTGCATTCAATTCGTTCTCT	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2447892_2448160_4_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATAG	GAAATATACCATGCGGGGGTGTTCAGTCCGGTTCA IS110 (1215bp)	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2447944_2448173_3_gold standard
AGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC .....	TAAAACTCGTATTCCTTCCAGTCTTTTCTTTGGGCA ACGGAACCACTCCTGCCAATATGCTTCTTTTCGTTTC	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2449719_2451060_20_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC .....	CCCGACAATACCCTCATGATACCGTCTCGGTTCTTC	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2449704_2451060_20_gold standard
GTTTCAATTCCATAGGTACGCTGAAAAAC .....	TTTGACCACCACTCATCTTAGCAATGCTAACGGGA	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2453564_2455390_27_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATAG	ACATTTTCAATTTTATGTAGATTGAAATAICGTTCTG MITE (300bp)	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2453564_2455337_26_gold standard
AGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC .....	CGGTGTGTGATGTATGAAAAGATGACGATGACCAGGGC ATGTGCCGAGGTGTGCCGAACGCGCAGATATCC	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2455691_2458143_37_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATATA	GCATGGACTGTGAGCGGATGTACGTGCTCTGGGAGTA IS1634(1701bp)	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2455676_2458157_37_gold standard
TCATAGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC .....	GCCGAGAAAATGTTCTCGGTGGCCTGGGATTACGCT GATGCCCTCCTTCTTGGATTGAACGGCAAGAAGTG	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2460043_2463456_51_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATAG	CGGACCCTTCAACCCAGGAGGAGGTTTTGACCTG MITE (300bp)	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2460106_2463402_49_gold standard
AGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC .....	GACTGGCAGAGATTCGGGATGTCGTGAGGCAGTATC	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2463757_2464503_11_our results
GTTTCAATTCCATAGGTACGCTGAAAAAC GTTTCAATTCCATAGGTACGCTGAAAAAC	GAGGGAGAATGCCGATGGTGGGTCGAGTCCGCGG	<i>Thermacetogenium phaeum</i> DSM 12270_NC_018870_2463742_2464503_11_gold standard

GTCGTAATCCCTTCAA	MITE(49bp)	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
AAATCAGGTCATTAATCCAAT	TGAAGATTAATGACCCGAAGCAATCCAGATTTAGAAGA	_1428433_1430685_30_ourresults
GTCGTAATCCCTTCAAATCAGGTCATTAATCCAAT	CATAGGCTGATTAAGGGCATGTAAGCATGGAGCTTGT	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
.....		_1428495_1430685_29_gold standard
GTCGTAATCCCTTCAAATCAGGTCATTAATCCAAT	GGAAGGGCAACAACCCCTCCGGAGGTTTTAAAATGA	
GTCGTAATCCCTTCAAATCAGGTCATTAATCCAAT		
.....	IS4 (1368bp)	
GTCGTAATCCCTTCAA	MITE(49bp)	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
AAATCAGGTCATTAATCCAAT	AGTATATCTTAACTTAACAATGGATTAGAAGACTTTT	_1709402_1709566_2_ourresults
GTCGTAATCCCTTCAAATCAGGTCATTAATCCAAT	TCGGTGCTGCGTATTATGCAACTGCTCCACGGAAGGG	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
GTCGTAATCCCTTCTAATCAGGTCATTC		_1709462_1709573_1_gold standard
.....		
GTTGTAATCCCTTCAA	MITE(47bp)	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
AAATCAGGTCATGGATTCCAAT	GGTATAGATAATAAAAAATGTCATGTTTATGATATA	_1709819_1710226_5_ourresults
GTCGTAATCCCTTCTAATCAGGTCATTAATCCAAT	AATAAACTATCCAAGAACTTTTGAATGAGATAGAA	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
.....		_1709819_1710226_5_gold standard
GTCGTAATCCCTTCAAGTCAGGTCATGAATCCAAT	AAATTCGGACTGAGCAATAGGATGGCAAAAAA	
GTCGTAATCCCTTCTAATCAGGTCATTC		
.....		
GTTGTAATCCCTTCAA	MITE(47bp)	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
AAATCAGGTCATGGATTCCAAT	GTTGAAGCACTCGGCGCAGTCTATCTTTGAGCAGA	_1710472_1710885_5_ourresults
GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT	AGAGAGATTTCTCCATTTCCCTAATTTAATACAGTTGAT	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
.....		_1710472_1710885_5_gold standard
GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT	TACACAATAAAACAGATACTTGTATTGCTTTTCTTA	
GTCGTAATCCCTTCTAATCAGGTCATTC		
.....		
GTTGTAATCCCTTCAA	MITE(47bp)	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
AAATCAGGTCATGGATTCCAAT	CTGAAGCGAATCAATTATATAAAAAATCTTTCTCTTCT	_1711131_1712652_20_ourresults
GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT	AGTGATCGGGGTGGTACGTCGGTTTGATATCGAAA	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
.....		_1711131_1712652_20_gold standard
GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT	GCCGGAAGGGAGCACACCCCTCCGGAGGTTACAAAA	
GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT		
.....		
GTTGTAATCCCTTCAA	MITE(47bp)	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
AAGTCAGGTCATGAATCCAAT	TGGGAAATATATAGATACCTGCTCTCCGGGTTGA	_1734298_1735756_19_ourresults
GTCGTAATCCCTTCAAGTCAGGTCATGAATCCAAT	TTAAAAAGAAATCCCAAGAAAATAGAGGAATCCTCAAG	<i>Flexistipes sinusarabici</i> DSM 4947_NC_015672_
.....		_1734298_1735756_19_gold standard
GCAATCCCTTCAAGTCAGGTCATGAATCCAAT	ATCTTCGCTCCGTGGGATATCCCGAGGACTTGAAAAA	
GTCGTAATCCCTTCAAGTCAGGTCATGAATCCAAT		

ACATTTGGCGAATATCTCATCGCAAATGTTTCTGTTTCAGAGCTTGCTGG (49bp, appears 3 times)

ACCCAGCCATCATATAATCGGCTGGGTTTCTGGTCCGAGCTTGCTGG (47bp, appears 4 times)

GTTTCAATTACCGTAGGTACTCAGAAC TTTCAAAGCCTTAAAGTACTATCAGAAC .....	<b>TCATATAGGTGCATTCCTTCATTGCGACATAATCGCAAA</b> TTGGACAAGCTAAAGCTAGGGCTTTACGGCGCATCCTA	<i>Thermobacillus composti</i> KWC4_NC_019897_3270701_3271073_5_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3270768_3271074_4_gold standard
GTTTCAATTCCTTATAGGTACTCAGAAC TTCAATTCCTTATAGGTACTCAGAAC GTTTCAATTCCTCATAGGTACGATCAAAAC .....	CGATACGGGCGATGTCATAGTACTTCGCCAGATGAGA <b>IS110(1105bp)</b> TTCACACTTCGGCATGTCGGCCCTCGGACTTACGAC	<i>Thermobacillus composti</i> KWC4_NC_019897_3272406_3273774_20_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3272406_3273720_19_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC <b>GTTTCAATTCCTCATAG</b> ATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC .....	<b>TATAGTGCTCGGGTTCAAGGACGCCGGAATATACCGG</b> <b>IS110(1227bp)</b> GGAGCCTAAACCATGCGTCAATATCGGGACGAAT GAGCAGCCGGCGCCCAATCCGATGCACAGACGACG	<i>Thermobacillus composti</i> KWC4_NC_019897_3275324_3277594_34_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3275311_3277541_33_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC <b>GTTTCAATTCCTCATAG</b> ATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC .....	<b>ACTCCTGCGAGCGCTCGAGGCGCCGACGGCATT</b> <b>IS110(1227bp)</b> CATCAGCGCGCCGACACCGGCCACGTCGACACCG GCAACCGGAACCGGTGCGCGCCGCGCGTGTGT	<i>Thermobacillus composti</i> KWC4_NC_019897_3279144_3280208_16_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3279131_3280155_15_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC <b>GTTTCAATTCCTCATAG</b> ATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC .....	<b>CGGGACGATCTATGTCATGTTCTCCGGCGTGAAGT</b> <b>IS110(1227bp)</b> AAAGGAATGGGACAGACATCGTCTGTTCCTGAAATTA CATGTCGAGCTCCACCATGAGAGCCGATCCGCCGCGC	<i>Thermobacillus composti</i> KWC4_NC_019897_3281758_3283032_19_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3281745_3282977_18_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC <b>GTTTCAATTCCTCATAG</b> ATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC .....	<b>CTGTATCGTATATTCAGTTCCTGATCCTCGGAAGCATT</b> <b>IS110(1228bp)</b> GTTACGGCTGTGCTCCACTAGCCCTTCCACGCTTC GTGATACGTCCGGAATACTCGTACCGTCATAATCCGT	<i>Thermobacillus composti</i> KWC4_NC_019897_3284583_3285001_6_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3284570_3285001_6_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC .....	GTGAGCCATCAGAAATATCTCGTTCCTATCCGT <b>IS(1293bp)-IS(1278bp)-IS110(1224bp)</b> GATCGAAAAGCGAGAAACGATTGAGTTGAACGTGACGA GAAGGAGTTTACCGTCGCCGTCATAACGCTGGAGT	<i>Thermobacillus composti</i> KWC4_NC_019897_3298570_3299110_8_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3298557_3299057_7_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC <b>GTTTCAATTCCTCATAG</b> ATAGGTAAGATCAA GTTTCAATTCCTCATAGGTACGATCAAAAC .....	<b>TGAGGTGGTGTGAATTATAGAAGCCTGAAAACAGG</b> <b>IS110(1227bp)</b> AAACCTACATCCGATGGACTACACTGGTACAGATCGTAC GATGCCCTGAAGTTCCAGACGGCCTGATACGCCA	<i>Thermobacillus composti</i> KWC4_NC_019897_3300646_3300996_5_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3300647_3300943_4_gold standard
GTTTCAATTCCTCATAGGTACGATCAAAAC <b>GTTTCAATTCCTCATAG</b> ATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC GTTTCAATTCCTCATAGGTACGATCAAAAC	<b>TCAGTTTGGCCATGTATGTGATCTTCGGCATTTCGT</b> <b>IS110(1225bp)</b> ACCGATGCCGAGAACATGGAGCTAAGAATCCGTATG TTTGCATGTACATGGCAGCGCGGAACAGAGATCG GGCCACGATCTCACCCACGCTGCTCACCGGTGT	<i>Thermobacillus composti</i> KWC4_NC_019897_3302548_3302761_3_our results <i>Thermobacillus composti</i> KWC4_NC_019897_3302535_3302761_3_gold standard

GTTTTAGCCTACCTATAAGGAATTGAAAC .....	AGCTGAAAGACATTTTAAAGTAGCAGGGCAAGGTACAG <b>IS110(1221bp)</b>	<i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_121660_124897_48_our results <i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_121660_124897_48_gold standard
GTTTTAGCCTACCTATAAGGAATTGAAAC GTTTTAGCCTACCTAT <b>CTATAAGGAATTGAAAC</b>	AGTATTAATAAATTAAGGAGTTTAAAAATGGAAAA <b>CTAATITGCTTTAAGCTTGTGCTACAAGATAGACCTTG</b>	<i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_126473_128016_23_our results <i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_126529_128029_22_gold standard
GTTTTAGCCTACCTATAAGGAATTGAAAC .....	CGCCGAGGTCTTCTCCTGTCGGATATAATGCGACAGG AGAGGAACAATTTGAATGATGAAATAGGACAAGA <b>IS1634(2879bp)</b> <b>AGGTCAAACCTGGGGCAAAGACTTACGACGGGAAA</b>	<i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_131265_133542_34_our results <i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_131317_133555_33_gold standard
GTTTTAGCCTACCTATAAGGAATTGAAAC GTTTTAGCCTACCTAT <b>CTATAAGGAATTGAAAC</b>	TGCCATCACTATCATACTATCTCCATCCACTCTTA TGGAATAGCAAAGCAAGTGCAGAACTGGATAAAAAA <b>IS110(1221bp)</b> <b>CCTATCAGCTCCGGTGGCACCCAAACGCAATAGC</b>	<i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_135132_136869_26_our results <i>Thermoanaerobacter brockii</i> subsp. finnii Ako-1_NC_014964_135184_136882_25_gold standard
GTTTTAGCCTACCTATAAGGAATTGAAAC .....	GATAGCCTTCAGCTATTAACTTTCCGCTTCCAATTG CCATGTTGTTTACGCTCAATCTCGTTACGTAGAAA <b>IS110(1221bp)</b>	

GTTTTAGCCTACCTATAAGGAATTGAAAC	AGCTGAAAGACATTTTAAAGTAGCAGGGCAAGGTACAG	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
.....		_NC_010321_121655_124879_48_our results
GTTTTAGCCTACCTATAAGGAATTGAAAC	AGTATTAATAAAATTAAGGAGGTTTAAAAATGGAAAA	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTAT	IS110 (1221bp)	_NC_010321_121655_124892_48_gold standard
CTATAAGGAATTGAAAC	CTAATTTGGCTTTAAGCTTGTGTGCTACAAGATAGACCTTG	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTATAAGGAATTGAAAC	CGCCGAGGCTCTCTCCTGTCGCGATATAATGGCGACAGG	_NC_010321_126469_128012_23_our results
.....		<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTATAAGGAATTGAAAC	AGAGGAACAATTGGAAATGTATGAAATAGGACAAAGA	_NC_010321_126525_128025_22_gold standard
GTTTTAGCCTACCTAT	IS110 (1221bp) – IS110 (1221bp)	
CTATAAGGAATTGAAAC	AGGTCAAAAGCTGGGGCAAGACTACGACGGGAAA	
GTTTTAGCCTACCTATAAGGAATTGAAAC	TGCCATCACTATCATACTATCTCCATCCACTCTTA	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
.....		_NC_010321_131261_134422_47_our results
GTTTTAGCCTACCTATAAGGAATTGAAAC	CACTACGGCATGCGGTAGCTCCTGATCATGTTCCG	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTATAAGGAATTGAAAC	MITE (309bp)	_NC_010321_131313_134493_47_gold standard
CTATAAGGAATTGAAAC	GTTTTTTGTACATTTCACTATCAAGGATTTTTCTAT	<i>Thermoanaerobacter pseudethanolicus</i> ATCC
GTTTTAGCCTACCTATAAGGAATTGAAAC	TGCCGCTGTGCTTGCAGTATTCTACGTCGCCGA	33223_NC_010321_134731_135597_13_our results
.....		<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTATAAGGAATTGAAAC	CCGCCACATACATCCAAACAGGAAGTACCAGCATGG	_NC_010321_134784_135610_12_gold standard
GTTTTAGCCTACCTAT	IS110 (1221bp)	
CTATAAGGAATTGAAAC	AGAAGAATATTGACAGTTTAACTGTCTCAAAGATAGC	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTATAAGGAATTGAAAC	ACTGTCCCTCAATTCCCTGTTTCCCTCCTTAAGCTT	_NC_010321_137187_138123_14_our results
.....		<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTATAAGGAATTGAAAC	TGGATAGCAAAGGCAAGTGCAGAACTGGATAAAAAGA	_NC_010321_137241_138136_13_gold standard
GTTTTAGCCTACCTAT	IS110 (1221bp)	
CTATAAGGAATTGAAAC	CCTATCAGCTCCGGTGGCACCACCAAAGCAGTGG	
GTTTTAGCCTACCTATAAGGAATTGAAAC	GATAGCCTTCAGCTATTAACTTTCCGCTTCCAATTG	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
.....		_NC_010321_139713_141450_26_our results
GTTTTAGCCTACCTATAAGGAATTGAAAC	CCATGTTGGTTTACGCTCAATCTGGTTACGTAGAAA	<i>Thermoanaerobacter pseudethanolicus</i> ATCC 33223_
GTTTTAGCCTACCTAT	IS110 (1221bp)	_NC_010321_139765_141463_25_gold standard

GTCTTATCTGAACTATGAGGGATGTAAC	ATAGAGGTACAGGACAAGCTTGGCTCGAAGGAGAAG	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
.....		_344075_347921_58_our results
GTCTTATCTGAACTATGAGGGATGTAAC	CTTTTCAGTCTAAATAATATCTGCTTCGCAAGGAAA	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GTCTTATCTGAACTATGAGGGATGTAAC	ATPase (792bp)-IS (1230bp)-IS (1071bp)	_344075_347921_58_gold standard
GTCTTATCTGAACTATGAGGGATGTAAC	CTTTTCCTTTTCAAAGACCTTGCACACAAAAAATC	
.....		<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GTCTTATCTGAACTATGAGGGATGTAAC	CATAGAAAGCGGACTTTTTACTGGAAGGATGAAG	_351757_353150_21_our results
GTCTTATCTGAACTATGAG	IS (909bp)-IS (1230bp) – IS (438bp)	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GAGGGATGTAAC	ACAAAAAATGCAAGGTCTTGGAAACAATATGATGA	_351757_353160_21_gold standard
GTCTTATCTGAACTATGAGGGATGTAAC	TCAACTTTCGGTTTCTCTTTCAGCTTGTGTGACTGC	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
.....		_356437_357973_23_our results
GTCTTATCTGAACTATGAGGGATGTAAC	ATTTCCAAATGTCACATATAGGCAAAATACCCGCTGTA	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GTCTTATCTGAACTATGAG	IS (1248bp) – IS (486bp)	_356504_357983_22_gold standard
.....		<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GAGGGATGTAAC	IS (987bp)-IS (1230bp) – IS (321bp)	
GTCTTATCTGAACTATGAGGGATGTAAC	ACAT TGT TGGGTCCTCGATTTAGAGCCGATGAGCCG	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
.....		_539664_541402_26_our results
GTCTTATCTGAACTATGAGGGATGTAAC	TTGCATACTAGTGGGATCAATAATCGGCTCTATAATA	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
.....		_539731_541402_25_gold standard
GTCTTATCTGAACTATGAGGGATGTAAC	AAGGGGTTGGAAAACTACTATTGCCACTTGAC	
GTCTTATCTGAACTATGAGGGATGTAAC	IS (1131bp) – IS (1221bp)	
.....		<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GTTTACATCCTCATAGTTCAGATAAGAC	TAAAGTATTAATCCTCTTTCAGCCCACTGTGCTA	_2513322_2520034_102_our results
.....		<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
GTTTACATCCTCATAGTTCAGATAAGAC	AGCTTTTGGAAAGCTAATCCAGATAGCCCTTATCA	_2513322_2519967_101_gold standard
GTCTATAGTTCAGATAAGAC	IS110 (1287bp)	
GTTTACATCCTCATAGTTCAGATAAGAC	TATCAACATTGAACCAAAGCTGTTATGTGGTCATCA	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
.....		_2522005_2524601_39_our results
GTTTACATCCTCATAGTTCAGATAAGAC	TAACGCAGACAGCGCAACAAAGAAAGCGTGTGATAA	<i>Caldicellulosinuptor saccharolyticus</i> DSM 8903_NC_009437
.....		_2521995_2524601_39_gold standard
GTTTACATCCTCATAGTTCAGATAAGAC	TGATGATTATTACCAACAACATACAGATAAAAAA	

ATAGGTAGGCTAAAAAC TTTCAATCCCTTATAGGTAAGCTAAAAAC .....	<b>CAGCATTACCGCTGTAAAAGTATTGTATCTAATTTAA</b> AAATTTTGCAGGGGAGGAGACATAAAACGAAATGAA	<i>Thermoanaerobacter</i> sp. X513_NC_014538_2149198_2149615_6_ourresults
GTTTCAATCCCTTATAGGTAAGCTAAAAAC GTTTCAATCCCTTATAGGTAAGCTAAAAAC ATAGGTAGGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	CCACTCCCGCCAAATTCAAAACCTCATTTTTCCTT <b>IS110 (1220bp)</b> TAGAGAGGGTAAAAGCGGCATGCCAAGTTTGTAAA TTAAAGGAGGAAGATAAAAATGATAACTTCTGTAAA	<i>Thermoanaerobacter</i> sp. X513_NC_014538_2149254_2149615_5_gold standard
GTTTCAATTCCTTATAGGTAGGCTAAAAAC <b>GTTTCAATTCCTTATAG</b> ATAGGTAAGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	<b>GAAAACCTAACCCGACCGATGCGGGGAGATAAAAAAA</b> <b>IS110 (1221bp)</b> AATGAAATTCAAATCTTCAACAGTAAAGATTTGGC CACATACAGGACATCCACACCCGCTCCGATAACT	<i>Thermoanaerobacter</i> sp. X513_NC_014538_2340512_2341253_11_ourresults <i>Thermoanaerobacter</i> sp. X513_NC_014538_2340499_2341198_10_gold standard
GTTTCAATTCCTTATAGGTAGGCTAAAAAC <b>GTTTCAATTCCTTATAG</b> ATAGGTAGGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	<b>AACCAAACCTTGAAAAAGCTGCATTTTCAAGTTTA</b> <b>IS110 (1221bp)</b> CTTTTCTGACCACCAAAAACCTATGTGGCTGTTCTG GATTCCTACTGCATTATACATCTGTGTGCTACATCA	<i>Thermoanaerobacter</i> sp. X513_NC_014538_2342830_2346967_62_gold standard
GTTTCAATTCCTTATAGGTAGGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	TTTAAAAATCGTATAAAGAGCTTATTGAGCTGTATT	<i>Thermoanaerobacter</i> sp. X513_NC_014538_2348597_2362246_205_ourresults <i>Thermoanaerobacter</i> sp. X513_NC_014538_2348597_2362246_205_gold standard
GTTTTAGCTTACCTATAAGGGATTGAAAC .....	AAGGAAAAATGAGGTTTTGAATTTGGCGGGAGTGG	<i>Thermoanaerobacter</i> sp. X514_NC_010320_764967_765384_6_ourresults
GTTTTAGCTTACCTATAAGGGATTGAAAC <b>GTTTTAGCCTACCTAT</b> ATAGGTAGGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	<b>TTAAAATTAGATACAAAATACTTTTACAAGCGGTAATGCTG</b> <b>IS110 (1220bp)</b> TAGAGAGGGTAAAAGCGGCATGGCAAGTTTGTAAA TTAAAGGAGGAAGATAAAAATGATAACTTCTGTAAA	<i>Thermoanaerobacter</i> sp. X514_NC_010320_764967_765328_5_gold standard <i>Thermoanaerobacter</i> sp. X514_NC_010320_2340417_2341158_11_ourresults
GTTTCAATTCCTTATAGGTAGGCTAAAAAC <b>GTTTCAATTCCTTATAG</b> ATAGGTAAGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	<b>GAAAACCTAACCCGACCGATGCGGGGAGATAAAAAAA</b> <b>IS110 (1221bp)</b> AATGAAATTCAAATCTTCAACAGTAAAGATTTGGC CACATACAGGACATCCACACCCGCTCCGATAACT	<i>Thermoanaerobacter</i> sp. X514_NC_010320_2340404_2341103_10_gold standard
GTTTCAATTCCTTATAGGTAGGCTAAAAAC <b>GTTTCAATTCCTTATAG</b> ATAGGTAGGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	<b>AACCAAACCTTGAAAAAGCTGCATTTTCAAGTTTA</b> <b>IS110 (1221bp)</b> CTTTTCTGACCACCAAAAACCTATGTGGCTGTTCTG GATTCCTACTGCATTATTACATCTGTGTGCTACATCA	<i>Thermoanaerobacter</i> sp. X514_NC_010320_2342748_2346925_63_ourresults <i>Thermoanaerobacter</i> sp. X514_NC_010320_2342735_2346872_62_gold standard
GTTTCAATTCCTTATAGGTAGGCTAAAAAC GTTTCAATTCCTTATAGGTAGGCTAAAAAC .....	TTTAAAAATCGTATAAAGAGCTTATTGAGCTGTATT	<i>Thermoanaerobacter</i> sp. X514_NC_010320_2348515_2363082_219_ourresults <i>Thermoanaerobacter</i> sp. X514_NC_010320_2348502_2363082_219_gold standard

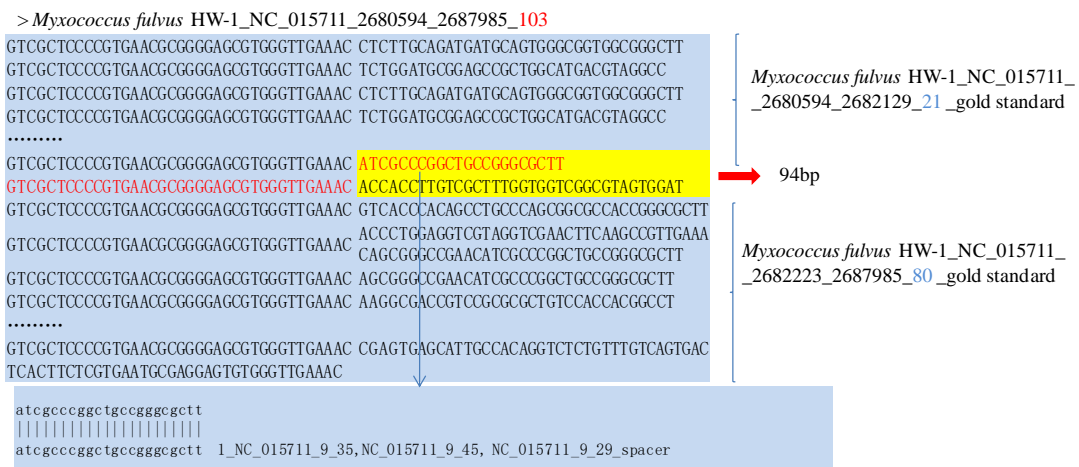
	GTTTCGGTCCCTCTCGGGTTTTGGGTCTGACGAC	
Mycobacterium africanum GM041182	CRISPR_30_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_29	CRISPR_29-IS6110(1355bp)- CRISPR_29
Mycobacterium bovis AF2122/97	CRISPR_17_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_24	CRISPR_16-IS6110(1355bp)- CRISPR_24
Mycobacterium bovis BCG str. Korea 1168P	CRISPR_19_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_29	CRISPR_18-IS6110(1355bp)- CRISPR_29
Mycobacterium bovis BCG str. Mexico	CRISPR_19_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_29	CRISPR_18-IS6110(1355bp)- CRISPR_29
Mycobacterium bovis BCG str. Pasteur 1173P2	CRISPR_19_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_29	CRISPR_18-IS6110(1355bp)- CRISPR_29
Mycobacterium bovis BCG str. Tokyo 172	CRISPR_19_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_29	CRISPR_18-IS6110(1355bp)- CRISPR_29
Mycobacterium tuberculosis CAS/N1TR204	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_23
Mycobacterium tuberculosis CDC1551	CRISPR_19_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_15	CRISPR_18-IS6110(1355bp)- CRISPR_15
Mycobacterium tuberculosis EAI5	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_23
Mycobacterium tuberculosis EAI5/N1TR206	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_23
Mycobacterium tuberculosis str. Haarlem/N1TR202	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_23
Mycobacterium tuberculosis H37Ra	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_23
Mycobacterium tuberculosis H37Rv	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_23
Mycobacterium tuberculosis str. Beijing/N1TR203	CRISPR_17_GTTTCGGTCCCTCTCGGGTTTTGGG-IS6110(1355bp)- -GGGTTTTGGGTCTGACGAC_CRISPR_23	CRISPR_17-IS6110(1355bp)- CRISPR_22
Mycobacterium canettii CIPT 140060008	ISL3(2846bp)-GGTTTTGGGTCTGACGAC_CRISPR_26	ISL3(2846bp)- CRISPR_25
Mycobacterium tuberculosis F11	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- CRISPR_18	CRISPR_17-IS6110(1355bp)- CRISPR_18
Mycobacterium tuberculosis KZN 1435	CRISPR_17_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- CRISPR_21	CRISPR_16-IS6110(1355bp)- CRISPR_21
Mycobacterium tuberculosis KZN 4207	CRISPR_17_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- CRISPR_21	CRISPR_16-IS6110(1355bp)- CRISPR_21
Mycobacterium tuberculosis KZN 605	CRISPR_17_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- CRISPR_21	CRISPR_16-IS6110(1355bp)- CRISPR_21
Mycobacterium tuberculosis CDC5079	CRISPR_15_GTTTCGGTCCCTCTC-IS6110(1355bp)	CRISPR_14-IS6110(1355bp)
Mycobacterium tuberculosis CDC5180	CRISPR_15_GTTTCGGTCCCTCTC-IS6110(1355bp)	CRISPR_14-IS6110(1355bp)
Mycobacterium tuberculosis CTRI-2	CRISPR_18_GTTTCGGTCCCTCTCGGGT-IS6110(1355bp)- CRISPR_6_GTTTCGGTCCCTCTC- -IS6110(1355bp)-CTCGGGTTTTGGGTCTGACGAC-CRISPR_15 CRISPR_17-IS6110(1355bp)- CRISPR_5 -	
Mycobacterium tuberculosis str. Haarlem	CRISPR_11_GTTTCGGTCCCTCTCGGGTTTTGGGTCTGAC-IS6110(1355bp)- CRISPR_6_GTTTCGGTCCCTCTCGGGT- -IS6110(1355bp)-GGGTTTTGGGTCTGACGAC-CRISPR_24 CRISPR_11-IS6110(1355bp)- CRISPR_5 -	
IS6110(1355bp)- CRISPR_24		

**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\_2687985 may be joined by the 94-bp gap as one longer CRISPR. Three other similar cases were detected in the bacteria *Caldicellulosiruptor obsidiansis* OB47, *Thermosiphon 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  $\in$  [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

<pre> GTTTAGAATCTACCTATGAGGAATGAAAAC CATATTTTTAGATAAAAATACACGCATTAAGTCCCCCA GTTTAGAATCTACCTATGAGGAATGAAAAC TTGAATTCCTCGACACCCCTCTTAAGATTTTCATTT GTTTAGAATCTACCTATGAGGAATGAAAAC TATACTTCTCTATGTTGCAATTAAGTTATAGCAATATCTCGTG GTTTAGAATCTACCTATGAGGAATGAAAAC ATTCAAATCTTCTTTAACATAATAACACCTCCATACATTT ..... GTTTAGAATCTACCTATGAGGAATGAAAAC TAAATCAGT GTTTAGAATCTACCTATGAGGAATGAAAAC ATCCCGATGTCGCGGCTTCTCTACTATCTGTATGGATTTT GTTTAGAATCTACCTATGAGGAATGAAAAC TTCATTTTGTATTATACATTTCAATTCCTTTTTTAA GTTTAGAATCTACCTATGAGGAATGAAAAC ATCACAGAATAATCTGTGTGTCGACGCACTATTTTAAATC GTTTAGAATCTACCTATGAGGAATGAAAAC GTGAATAAGTGTTCCTTCAAACCTTTCAAGACTAATGGTTTC GTTTAGAATCTACCTATGAGGAATGAAAAC CAAAAATGGCACTTGCAACACTTCTCTTTAGTTTCTCCTCA ..... GTTTAGAATCTACCTATGAGGAATGAAAAC GTACGAACTGTCTCAATTGCAACGGGATAATGTGT GTTTAGAATCTACCTATGAGGAATGAAAAC </pre>	<p>81bp</p>	<p><i>Thermosipho africanus</i> TCF52B_NC_011653_309784_310512_10_gold standard</p> <p><i>Thermosipho africanus</i> TCF52B_NC_011653_310593_311880_18_gold standard</p>
---	-------------	---

> *Herpetosiphon aurantiacus* ATCC 23779\_NC\_009973\_102032\_109407\_101

<pre> ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC TACGGATGAGGCTGCACTCGCGCTCAGGACCGCC ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC GTTGAAGGGATTCATGTTGATAITCACCCGCGCT ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC ACGCATCGATCCGATCATGGATCGCCACAACCTGGT ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC AAACCTTTTATAGCTCGCGCACTATGGCATGAT ..... ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC TCGCGCACTTGGCCAACTTGGGCAITGGACAAACGT ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC TATCCAAATAACATTTGCTATAITGCATCGCACATC ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC ATTCATCCATGGTTGCCAGAACCTTGGGGCATAG ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC TCGATCACGTTGACGGGCAACGCTTGCATTAG ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC AGCAGCTTCATGGCGGCTGCGGGTCTGCAATTCCG ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC GCCATGCCGCTGCTCAACAAATGGCTCCAT ..... ATTTCAATFACTCGATCCGATTAGAGGATACTGAAAC TGCATGGCGTTGAGTTCACGTGGCATCAAGCCAAGGCT ATTTCAATFACTCGATCCGCTGGACGAATATTGA </pre>	<p>112bp</p>	<p><i>Herpetosiphon aurantiacus</i> ATCC 23779_NC_009973_102032_108752_92_gold standard</p> <p><i>Herpetosiphon aurantiacus</i> ATCC 23779_NC_009973_108864_109408_7_gold standard</p>
--	--------------	--

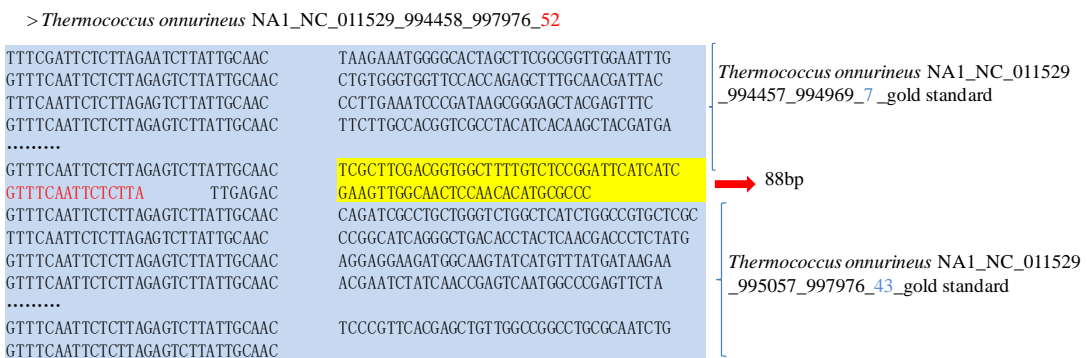
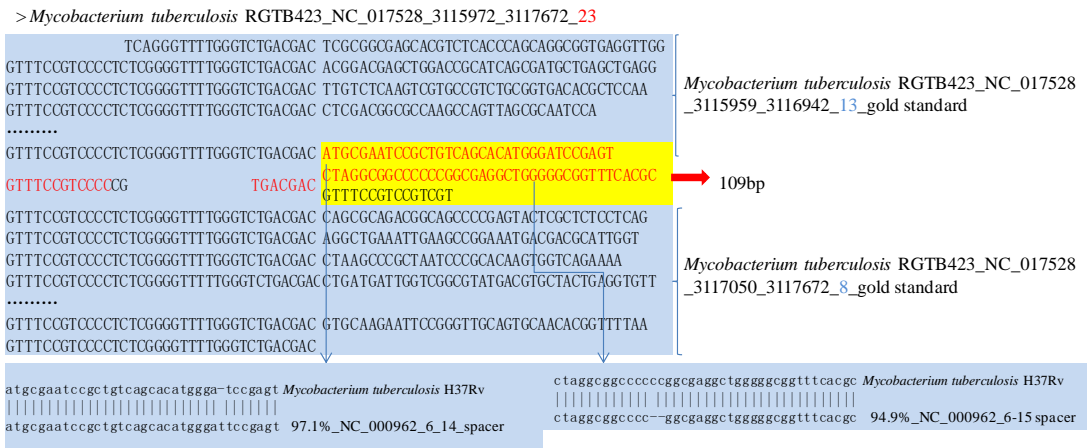
> *Salmonella enterica* subsp. enterica serovar 4,[5],12:i:- str. 08-1736\_NC\_021820\_4332392\_4333865\_24

<pre> GTTTATCCCCTGCGCGGGGAACA TTTTTCAGCCCTTGTGACTGCGAACGCCCTT CGGTTTATCCCCTGCGCGGGGAACAC GCGAAATAGTGGGAAAAACCCCTGGTTAACC CGGTTTATCCCCTGCGCGGGGAACAC TAGGCCTTGATACCATCGCTGCACTGCTCA CGGTTTATCCCCTGCGCGGGGAACAC GTTTATTACTGCTAGTTAATTAATGGGTGTC ..... CGGTTTATCCCCTGCGCGGGGAACAC GATCGAGTAACGTGGCTGGAACGG TCGGCGCGGGGAACAC AAAATTAAAGCCGAGGGTGGCACCAGCCCTTATT CGGTTTATCCCCTGCGCGGGGAACAC GCACCTCGAAACGGTTTTAAACACTACCCTTT CGGTTTATCCCCTGCGCGGGGAACAC TGGACCGATGGGGCCAAACATCGCCGAACGTGG CGGTTTATCCCCTGCGCGGGGAACAC GTTACGTTGCGTAAATGAAAGCGCGCAATAT CGGTTTATCCCCTGCGCGGGGAACAC CCAGAAAGTCCGCGTAGTGCCTGATGAACGAC ..... CGGTTTATCCCCTGCGCGGGGAACAC CAGCACGAAAAATTATTACTGTGCTGTCTCA CGGTTTATCCCCTGCGCGGGGAACAC </pre>	<p>75bp</p>	<p><i>Salmonella enterica</i> subsp. enterica serovar 4,[5],12:i:- str. 08-1736_NC_021820_4332390_4332968_9_gold standard</p> <p><i>Salmonella enterica</i> subsp. enterica serovar 4,[5],12:i:- str. 08-1736_NC_021820_4333042_4333865_13_gold standard</p>
--	-------------	--

<pre> gatcgagtaacgtg-gctggaacg-g       gatcgagtaacgtg-gctggaacg-g </pre>	<p><i>Salmonella enterica</i> subsp. enterica serovar Typhimurium str. 798</p> <p>0.962_NC_017046_2_10_spacer</p>
--	---

- > *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 chromosomes are 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 and NC\_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

GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	AAAAGATAAAAAGGAGGTGAGTAAATCCACTTAAAA	} <i>Cyanobacterium aponinum</i> PCC 10605_NC_019776_4113500_4114042_7_gold standard
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	GAAAAGTTGCAAGGTTTAAAAGGTGCTTACG	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	GGGATAGACTTAATTAAGAATTGAGGGCTTTA	} <i>Cyanobacterium aponinum</i> PCC 10605_NC_019776_54_810_10_gold standard
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	TGGTTAAAGCGATAAGTATTGTTTTATCCCA	
.....	.....	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	AACTTTTTTCATTCAATAATCCACATTTTGCCATCGTA	} 110bp
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	AAAAAATTGACCCCGGAGACTAACGAGACT	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	AAAAACCCAAAATGGGAGTACTTAAAGTACAAAGCGA	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	GTTCTAGCCGATATTTCTGTCAAACCAATTCGT	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	TGACGATGTGATCGGTAATTGGAAGTCACATTCAA	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	TTTGATTATTGGGATTAGTATTGGGATTGGGAAT	
.....	.....	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	ATCAAAAAGCCAAATCAATATCAGTATTAATCTGCA	
GTTTCAATCCCTAAGAGGTATTAATAAGAGTTTAAAC	.....	

> *Thermococcus* sp. CL1\_NC\_018015\_1949366\_1950313+1\_682\_24

GTTGCAATAAGACTCTGGGAAATTGAAAC	GCGGGCACTTCATCACCAGCCGACCCAATCACCAT	} <i>Thermococcus</i> sp. CL1_NC_018015_1949366_1950270_13_gold standard
GTTGCAATAAGACTCTGGGAAATTGAAAC	CACAGGGTATAAAGTCCGTGCTGATGGCGTAAACGTG	
GTTGCAATAAGACTCTGGGAAATTGAAAC	GGGGCGGGTGCITGATCTCTATCATCATACGTAGGT	} 90bp
GTTGCAATAAGACTCTGGGAAATTGAAAC	GAAATACTGAAAGCTTAACTGCCTGCTTTCTGC	
.....	.....	
GTTGCAATAAGACTCTGGGAAATTGAAAC	TAGGGTGGGAATTGTAATCGACCCGAGTAGGGGAGAA	} 90bp
GTTGCAATAAGACTCTGGGAAATTGAAAC	GTGATATGGCAAGCCTCTTACT	
GTTGCAATAAGACTCTGGGAAATTGAAAC	TTGGGGTTTTGCAAGGTGGGGAAAGTGTGCCCAT	
GTTGCAATAAGACTCTGGGAAATTGAAAC	ATTAGTATCACGGCAGGCTTATCGCCCTGCTACTA	} <i>Thermococcus</i> sp. CL1_NC_018015_48_682_9_gold standard
GTTGCAATAAGACTCTGGGAAATTGAAAC	TAGAGTATCACITGCTCACGCCATTGGCTTAGCCCTCA	
GTTGCAATAAGACTCTGGGAAATTGAAAC	TAAGCATTTTTGCCTGTGTCGTATGTAGTCTTAA	
.....	.....	
GTTGCAATAAGACTCTGGGAAATTGAAAG	CTTCTCTTGCAAAAGTCGAGCAAGCTTTGAGATACT	
GTTGCAATAAGACTCTGGGAAAGAGAAAG	.....	

> *Candidatus Puniceispirillum marinum* IMCC1322\_NC\_014010\_2753369\_2753527+1\_1792\_29

AGTATAGCACTCTGGTATTGAGAGCC	TAGAGCAACCTCTTCGGTCTCGTCAAACTTTCTGATGGA	} <i>Candidatus Puniceispirillum marinum</i> IMCC1322_NC_014010_2753369_2753527_2_gold standard
AGTATAGCACTCTGGTATTGAGAGCC	TAGAGCAACCTCTATAAGTCAACAACAATAATCGATGTTG	
AGTATAGCACTCTGGTATTGAGAGCC	TGTGAGGTGTGTTGTGAGATAGTGAGCTGC	} 39bp
AGTATAGCACTCTGGTATTGAGAGCC	CGTGGTGGTCTCAAGCAGATAAACCCCTTCC	
AGTATAGCACTCTGGTATTGAGAGCC	TGACCCGCTTGATATCGCATCACGCCACGCC	} <i>Candidatus Puniceispirillum marinum</i> IMCC1322_NC_014010_40_1792_26_gold standard
AGTATAGCACTCTGGTATTGAGAGCC	GTAACCCATGTGCCCAAGTCCGGGCC	
AGTATAGCACTCTGGTATTGAGAGCC	ATGAGCAGAATTCATTATCACTTTGACTT	
.....	.....	
AGTATAGCACTCTGGTATTGAGAGCC	CGATCCCTCGATGATGGTGTACCCGCCAA	
AGTATAGCACTCTGGTATTGAGAGCC	.....	

> *Thermocrinis albus* DSM 14484\_NC\_013894\_1496455\_1500577+1\_90\_62

CTTTCAACTCCACCGGTACATTAGAAAC	ACTATCCTGAACCTTCACACAGTATGCAAAACAGCG	} <i>Thermocrinis albus</i> DSM 14484_NC_013894_1496455_1500535_60_gold standard
CTTTCAACTCCACCGGTACATTAGAAAC	TGCTGGACAAGGCACTGCAAGGATGAGACAAAGAT	
CTTTCAACTCCACCGGTACATTAGAAAC	TACTTATCAAGCGTTTCTTGTAGCTTTTTTGTGACGTAG	} 42bp
CTTTCAACTCCACCGGTACATTAGAAAC	TATCAATGACTGAGGGCTGACACCCTCTTTTTT	
.....	.....	
CTTTCAACTCCACCGGTACATTAGAAAC	TGTAGTCTGTACTTCTTAATAACCTGTTCTATGTATGC	} 42bp
CTTTCAACTCCACCGGTACATTAGAAAC	TGCTGTAGGTACGAAGAACAAGTGTATAGATGATA	
CTTTCAACTCCACCGGTACATTAGAAAC	TCCAGTCCGGGGTTTCTCAGTCCCTTGTCTTGC	} <i>Thermocrinis albus</i> DSM 14484_NC_013894_1_90_1_gold standard
CTTTCAACTCCACCGGTACATTAGAAAC	.....	

> *Fervidicoccus fontis* Kam940\_NC\_017461\_1319183\_1319206+1\_794\_13

GAATCTCTCAGATAGAATTGAAAG	TCTGAAAAATCATCAACAATCAGCTTTTCCCT	} 60bp
GAATCTCTCAGATAGAATTGAAAG	TGGCATAATTTACTCCCTTGTCTCAGAATCAAGCATT	
GAATCTCTCAGATAGAATTGAAAG	TGGCATAATTTACTCCCTTGTCTCAGAATCAAGCATT	} <i>Fervidicoccus fontis</i> Kam940_NC_017461_37_794_12_gold standard
GAATCTCTCAGATAGAATTGAAAG	TTCACAAATTTAGTCAAAATGCGCTAGGAAACAGAA	
GAATCTCTCAGATAGAATTGAAAG	TCAAACCATGCTTCTTAGTATTTCTGCTGCTTCT	
GAATCTCTCAGATAGAATTGAAAG	CTCTTGTAAATGCTCTAGCAGTGCATCATAGTAGGC	
.....	.....	
GAATCTCTCAGATAGAATTGAAAG	AACTCTCCATGTATTCACAAACTTTCCAAATCCT	
GAATCTCTCAGATAGAATTGAAAG	.....	

> *Picrophilus torridus* DSM 9790\_NC\_005877\_1540318\_1545895+1\_309\_87

<pre> CTTTCAATCCTATTAGGTTATTATTAAAC TTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC ..... CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC CTTTCAATCCTATTAGGTTATTATTAAAC </pre>	<pre> TCCGCCGAATTTCAAACACTTATAGTGGACTAGA TATCTCTCTTTTCAATTCCTTATCGATTTATCGATCC CACGGTGGTGATGCAACGATAAGATCAACATTATCA TCCATCTAAAGACATTCTCTCAAATATACAAAGCC ..... TATAATCCTGCTTCCCTACATATCTAACGAACTCT <b>CATTGATTTTTCATGAAATTCATGAATTCGGATAA</b> CTGGATTTCCTTGGCCATATTGCGGAGACTATATA ATATATCGAATTTCAAACCTTATAGATAGACTAGA GCCAGAGGCATTTCCGCGTGGTCTTAAATGA ACAACAGGCAGAAAACAGAATTAAGACTGGCCATTGAG </pre>	<p><i>Picrophilus torridus</i> DSM 9790_NC_005877_1540318_1545873_82_gold standard</p> <p>→ 36bp</p> <p><i>Picrophilus torridus</i> DSM 9790_NC_005877_15_309_4_gold standard</p>
--	--	---

> *Thermococcus litoralis* DSM 5473\_NC\_022084\_2214800\_2215172+1\_2779\_46

<pre> CTTTCAATCTTTAAAGCTTATTGGAAC CTTTCAATCTTTAAAGCTTATTGGAAC CTTTCAATCTTTAAAGCTTATTGGAAC CTTTCAATCTTTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC ..... CTTTCAATCTTCTAAAGCTTATTGGAAC CTTTCAATCTTCTAAAGCTTATTGGAAC </pre>	<pre> AAAGCGTTATTAAGATGTAGTGCCTACTCTCCAT CTGTTTCATCCTACGCGGACAAGAAATGGATGATC ATAGGAGTGCTTCCAGTCAATCTTCTCGTATGT ATATATATAAATTAATTCGTTTACTGTACAATAAATTA CACCTCAGAATATGTGAAGTTTAGCCCTGTGATGCT <b>ACCGAAAAGAGAACCAAAAGGAGGAGGTGAAATAGA</b> AGCTTGATGGCTTGAATGAAAAACTGCCCTCTTGA TAAACTTGCTGAGATTGCGGCGATTTTACTATA AAAGACAACCTCGTAACAATTGGGTGGTCTGCTGTT TTGAAAAGAACTCAGGGAGCAATCAACCTATACAACGA AGCTTGATGGCTTGAATGAAAAACTGCCCTCTTGA ..... GATGAGAACGGGACGAAAGCTAAGTTACTACGTGGAGAGT </pre>	<p><i>Thermococcus litoralis</i> DSM 5473_NC_022084_2214800_2215162_5_gold standard</p> <p>→ 38bp</p> <p><i>Thermococcus litoralis</i> DSM 5473NC_022084_29_2779_40_gold standard</p>
--	--	---

> *Acidilobus saccharovorans* 345-15\_NC\_014374\_1495994-1496453+1-1050\_23

<pre> GTTTCAATAAAGTTTCACGGTTTCA ACTTCTGGGAGTGAGGCACCTGGATAGCGCCACAGC TTTTCAAGCCATCTTGGTTTC TTCGAGGATAGGGAGAATGCCAAAATATGACAGAGGTGTTACGCTGTCTATGAA GTTTCAACACCATCAGATGGTTTC CTCCACGTCTACGGCAATGAACTGACAGCAAAGCAACGC GTTTCAACACCATCAGATGGTTTC AGAAAATCAGTATGCCAGGTATAGACGTCCCAAACAGATA ..... GTTTCAACACCATTCTTGGTTTC ACTGCCTTCAGTGCCOAGTTGGCGGCATAAACACTG GTTTCAACACCATTCTTGGTTTC <b>CTGGAGAGGCCCCACAGGCCTCCCTGAGGCCAAGACATCCA</b> GTTTCAACACCATTCTTGGTTTC CCATTGGTAAATGAGATGTGGTTACGATCCAGAGCATAAA GTTTCAACACCATTCTTGGTTTC GCCGTGGAAGTAGTCTCGTCTGTCGACCACTCTCTCCA GTTTCAACACCATTCTTGGTTTC TGATTGTAACCTGGGCGCTGACAGGGTACATGTAGGTG GTTTCAACACCATTCTTGGTTTC ACGTGCCCTATGGAACGCTGAAGGTATACCTATCAAGGTCAA ..... GTTTCAACACCATTCTTGGTTTC TCGGAAAAGAGGAAAAGGAAAGGGTAACGTTGTTAA GTTTCAACACCATTCTTGGTTTC </pre>	<p><i>Acidilobus saccharovorans</i> 345-15_NC_014374_1495994_1496411_6_gold standard</p> <p>→ 42bp</p> <p><i>Acidilobus saccharovorans</i> 345-15_NC_014374_1_1050_16_gold standard</p>
--	---

## 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 phaenum* 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.