J Ind Microbiol Biotechnol (2014) 41:1677–1685
DOI 10.1007/s10295-014-1507-3

SIMB
Society for Industrial Microbiology and Biotechnology

Whole genome sequencing reveals a novel CRISPR system in industrial *Clostridium acetobutylicum*

Lixin Peng · Jianxin Pei · Hao Pang · Yuan Guo · Lihua Lin · Ribo Huang

Received: 4 July 2014 / Accepted: 28 August 2014 / Published online: 14 September 2014

Abstract *Clostridium acetobutylicum* is an important organism for biobutanol production. Due to frequent exposure to bacteriophages during fermentation, industrial *C. acetobutylicum* strains require a strong immune response against foreign genetic invaders. In the present study, a novel CRISPR system was reported in a *C. acetobutylicum* GXAS18-1 strain by whole genome sequencing, and several specific characteristics of the CRISPR system were revealed as follows: (1) multiple CRISPR loci were confirmed within the whole bacterial genome, while only one cluster of CRISPR-associated genes (Cas) was found in the current strain; (2) similar leader sequences at the 5' end of the multiple CRISPR loci were identified as promoter elements by promoter prediction, suggesting that these CRISPR loci were under the control of the same transcriptional factor; (3) homology analysis indicated that the present Cas genes shared only low sequence similarity with the published Cas families; and (4) concerning gene similarity and gene cluster order, these Cas genes belonged to the csm family and originated from the euryarchaeota by horizontal gene transfer.

Keywords CRISPR/Cas · *Clostridium acetobutylicum* · Whole genome sequencing

L. Peng · J. Pei · H. Pang · Y. Guo · L. Lin · R. Huang (✉)
State Key Laboratory of Non-Food Biomass and Enzyme Technology, National Engineering Research Center for Non-Food Biorefinery, Guangxi Academy of Sciences, Guangxi Key Laboratory of Biorefinery, Nanning, Guangxi 530007, People's Republic of China
e-mail: pengpepery@hotmail.com

L. Peng
College of Life Science and Technology, Guangxi University, Nanning, Guangxi 530005, People's Republic of China

Introduction

Clostridium acetobutylicum is an organism that is commonly used to produce acetone and butanol by traditional acetone–butanol–ethanol fermentation (ABE) [1]. It can convert various biomasses such as corn, starch and molasses into acetone, butanol and ethanol. Butanol is an especially ideal future second-generation renewable biofuel because of its higher energy density and lower volatility. Thus, ABE fermentation has again recently become the focus of biofuels in studies. To date, many *C. acetobutylicum* strains have been isolated and adopted for different feedstock and fermentation processing. Besides having high solvent yields and high tolerance for butanol, industrial *C. acetobutylicum* strains also require an increased immunity against bacteriophages because of their frequent exposure to phages during fermentation [2].

Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated(Cas) were recently proven to be the inheritable immune system of prokaryotes. They have been widely found in bacteria (~40 %) and archaea (~90 %) [3]. A typical CRISPR structure was well described previously [4, 5]. In brief, they are composed of CRISPR arrays and Cas genes. CRISPR arrays consist of repetitive elements (repeats) interspersed with genome targeting sequences (spacers). These repeats are usually ~20–50-bp long and have conserved RNA structures within the CRISPR arrays [6]. Spacers originate from short fragments of invading foreign DNA. When the CRISPR system is activated, spacers should associate with repeats to produce mature CRISPR RNAs(crRNA) to recognize the target DNA. Cas genes are adjacent and upstream of the CRISPR arrays and show high structural similarity to proteins with an endonuclease. It has a DNA or an RNA-binding function [7] and there is a remarkable diversity among the Cas

Table 1 Four space sequences and their matched genes in nr database

Space no.	Length	Similarity	Matched gene (matched range)	Protein ID	In present genome
1	38	100	Hypothetical protein (316–353)	AEI31626.1 ADZ20164.1 AAK79088.1	Complete deletion
2	35	97	DNA replication protein DnaC (55–89)	AEI34732.1 ADZ20985.1 AAK79895.1	Partial deletion
3	38	100	Hypothetical protein (566–603)	AEI32072.1 ADZ20986.1 AAK79896.1	Partial deletion
4	36	97	Site-specific recombinase (867–902)	AEI32090.1 ADZ21005.1 AAK79913.1	Complete deletion

genes within the CRISPR system [8]. So far, there have been more than 45 Cas families identified in various organisms [8].

The mechanism of CRISPR immunity is similar to the RNA interference (RNAi) pathway in eukaryotes, which rely on short RNA for sequence-specific detection and direct degradation of invading nucleic acids. Moreover, this capability of DNA editing is not limited to prokaryotic organisms. The type II CRISPR system, Cas9 nuclease, can function with custom guide RNA to precise cleavage at sequence-specific genomic loci in human and mouse cells, demonstrating an inheritable immune technology and wide applicability as a genome editing tool [9–11].

In this study, we reported the first case of a CRISPR system in *C. acetobutylicum*. Genetic characteristics described here include a CRISPR system interspersed within the whole genome of *C. acetobutylicum*, homologous leader sequences at 5' terminus of CRISPR loci and low similarity to other members of relevant protein families. The above findings reveal a unique evolution of CRISPR system in *C. acetobutylicum* GXAS18-1, and provide a mechanism to confer phage resistance in commercial *C. acetobutylicum* strain.

Materials and methods

Sequencing and assembly

The *C. acetobutylicum* GXAS18-1 strain was originally isolated from the soil and adapted for butanol fermentation. The genome library construction and sequencing of *C. acetobutylicum* GXAS18-1 were performed on an Illumina Genome Analyzer IIX instrument at Beijing Genomic Institute (BGI, Shenzhen, China). The library with 350 bp inserts yielded ~5.4 million 90 bp paired-end reads and 496 Mb of raw data, which covered 90 % of the C. *acetobutylicum* genome with an average of 100-fold depth of data. To assemble the genome, reads were firstly filtered by the SOAP [12] software to exclude the low-quality reads. They produced 480 Mb of clear data and were assembled into a scaffold sequence using the SOAPdenovo software [12] with an optimal K-mer parameter.

Cas gene finding

Three programs including Glimmer [13], SNAP [14] and GeneMark [15] were trained with the *C. acetobutylicum* ATCC 824 genome data and employed to detect potential ORFs from the sequence assembly. The various results were combined into consensus ORFs and then manually curated. All genes were annotated by Blastp to the nr database with an E-value cutoff of 10–6 and a 60 % residue homology. The potential Cas genes and the unmatched genes were further validated by the Blast2go tool using the pfam, hmm-tigr and interpro databases.

CRISPR finding and prediction of the promoter/RNA structure

CRISPR, including repeat and spacer sequences, was identified by the CRISPRdb database [3] (http://crispr.u-psud.fr/). The non-coding sequences found immediately upstream of each CRISPR repeat were selected as the putative leader sequences. These potential leader sequences were further predicted as the promoter using the BDGP Neural Network Promoter Prediction (Reese, 2001). The repeat sequences of the RNA secondary structure were predicted by taking advantage of the RNAfold program [16] (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

Cas gene family finding

To investigate the Cas families, the Cas profiles of the cluster of orthologous groups of protein (COG) [17] were collected as core gene families. The protein database of the bacterial genomes (ftp://ncbi.nih.gov/genomes/Bacteria/) was searched with a hmmsearch algorithm using the

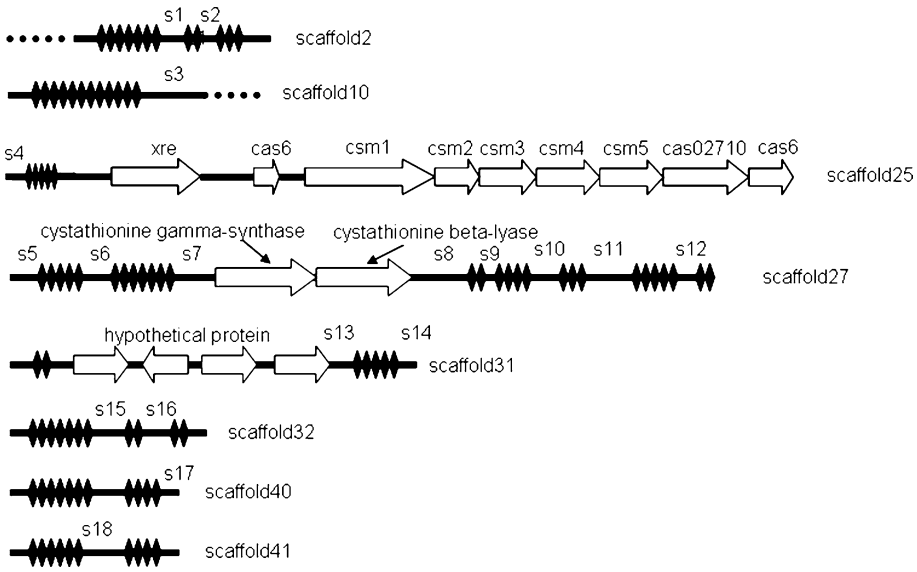

Fig. 1 Location of the multiple CRISPR loci along the assembled genome. *Black diamonds* represent the repeat/spacer units in the CRISPR loci. The *arrow boxes* represent the genes and are labeled with their respective names. The *black lines* represent the non-coding sequences, and the labeled sequences within those non-coding sequences were further explored by homology analysis and promoter prediction

E-value of 6. Cas genes from the pfam and the interpro databases were found to complement by a member of a relevant gene family if they are filtered by an E-value in the hmmsearch algorithm.

Phylogenetic tree comparison

To construct the gene tree, orthologous Cas genes were aligned using the MUSCLE method [18] and were then concatenated together. The gene tree was further constructed using the MEGA [19] software with the NJ algorithm. Moreover, these Cas gene families span across all three domains, and a method, based on concatenation of 31 orthologous genes, was employed to construct a highly resolved species tree [20]. Briefly, 31 gene profiles of COG from the previous report [20] were collected as the core orthology. They were then searched in the protein database of the relevant organisms using the hmmsearch algorithm to extract the best-match orthology. These orthologous genes were then aligned, concatenated and used to construct the final species tree utilizing the NJ algorithm. Both gene tree and species tree were aligned and analyzed using the compare2trees tool [21].

Results

CRISPR in *C. acetobutylicum* genomes

Seven scaffold sequences from the assembled *C. acetobutylicum* genomes were confirmed to contain CRISPR loci (Fig. 1). Except for scaffold10, which harbors a longer CRISPR array (12 repeat/spacer units), CRISPR loci usually contained 8–10 repeat/spacer units scattered along the assembled sequences. Some of them are separated by genes or non-coding sequences (Fig. 1). For example, the CRISPR locus in scaffold27 sequence contains a non-coding sequence and two adjacent genes coding cystathionine gamma-synthase and cystathionine beta-lyase. As a result, this CRISPR locus is divided into five confirmed and three possible repeat-spacer arrays. In scaffold31, the CRISPR locus is interrupted by four consecutive hypothetical genes which lack homology to any sequences found in the non-redundant (NR) database (Fig. 1). Clustered Cas genes are only found at the 3' end of the CRISPR locus within scaffold25, which indicates that this scaffold has the typical structure of the CRISPR system and is composed of the CRISPR locus, the leader at the 5' end of CRISPR and a cluster of eight Cas genes at the 3' end of CRISPR. Interestingly, an additional gene found between the CRISPR array and the cluster of Cas genes was a member of the transcriptional regulator *XRE* family. When a promoter prediction was performed at the leader region, which is located upstream of the 5' end of the non-coding sequences of the *XRE* gene and the cluster of Cas genes, a clear promoter element signal was shown to reside inside these sequences.

Repeat sequences

A total of 94 repeat sequences were collected from all confirmed CRISPR loci. All the repeats were 30-bp long and had a conserved sequence. However, different frequencies of nucleotide substitution were observed along the repeat sequences. Figure 2 presents an alignment of multiple conserved repeat sequences with two sites with high nucleotide polymorphism including a nucleotide transition of T/A (48/46 %) at the 8th position and a transversion of A/G (47 %/30 %) at the 17th position. To further study the effect of the nucleotide substitutions on the RNA secondary structures, a consensus sequence

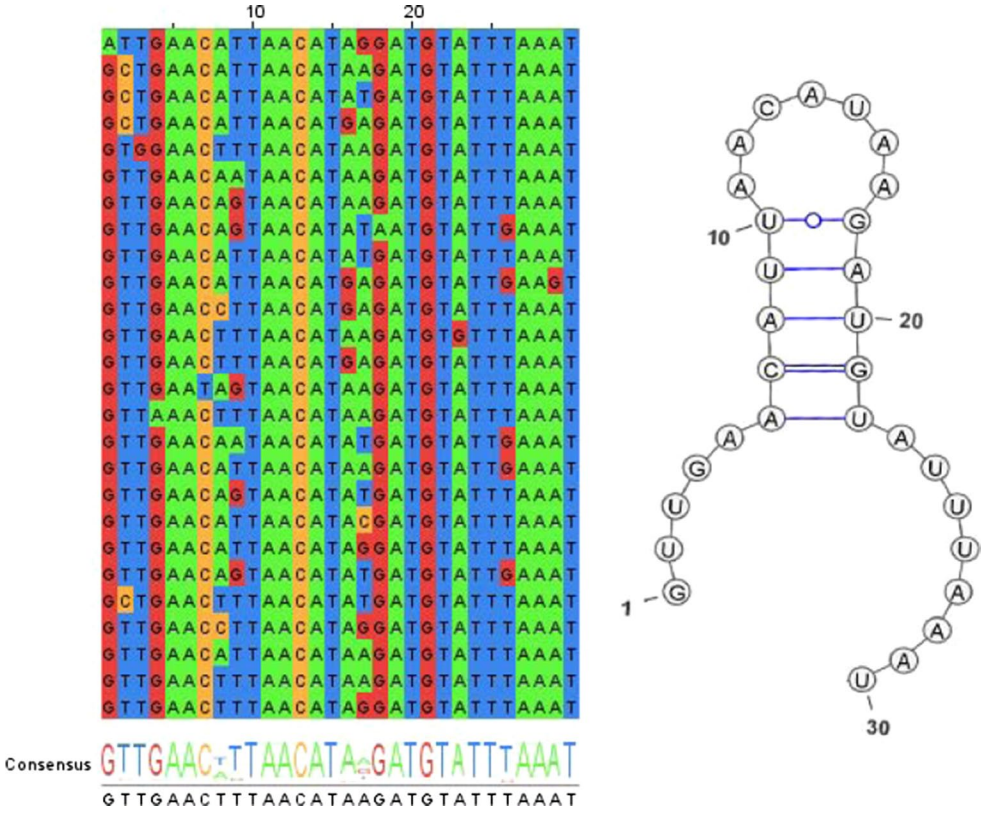

Fig. 2 Alignment of the repeat sequences along their consensus sequence (*left*) and RNA structure (*right*)

of all the repeat sequences was constructed, and then, its secondary structure predicted using the RNAfold software [22]. Results indicated that the sequence can form a stable RNA secondary structure with a 5-bp-long stem loop (Fig. 2). Furthermore, the substitution at the 8th position exists precisely at the stem structure, which results in a U:U pair mismatch resulting in a transition from an A to a T and may decrease the stability of the RNA structure.

According to a previous study by Kunin et al. [6], repeats can be divided into 33 clusters based on sequence similarity. The classification of repeats was also analyzed here using the same method [6]. The analysis indicates that the present repeats can be grouped with cluster 15, which has been found only in a few organisms such as *Syntrophomonas wolfei* Goettingen and *Clostridium tetani* stains. However, the repeats in the present strain lack the conserved 3' terminus sequence of GAAA(C/G) that was found in many other species belonging to cluster 15, but have the sequence of TAAAT.

Space sequences

Eighty space sequences with lengths between 33 and 43 bp were detected in the confirmed CRISPR loci of *C. acetobutylicum*. Only two loci were verified to contain the same repeat/spacer units among them, and all the spacers produced 73 unique spacer sequences. Due to the fact that spacers originate from partially homogenous invading sequences, we analyzed all spacer sequences by searching for them in the bacteriophage genome database and the nr database. The results indicated that the present spacer sequences showed no sequence homology in the current phage genome database. Moreover, four spacer sequences matched with sequences found in the nr database, all of which originated from the complete genome of three other *C. acetobutylicum* strains. However, four spacer sequences lacked any self gene targets in the present *C. acetobutylicum* strain [23] due to a complete or partial deletion of these four genes (Table 1). Two genes are fully deleted in the present assembled genome: one of them encodes a recombinase, while the other encodes a hypothetical protein without an annotated function. However, two genes displayed partial deletions: one encodes a DNA replication protein that is missing the spacer's homologous sequence at the 5' terminus, and the other encodes a hypothetical protein that lacks the spacer's homologous sequence at the 3' terminus.

Leader sequences

Twenty-five non-coding sequences flanking the CRISPR arrays were aligned by ClustalW to explore for potential leader/promoter regions. Among them, eighteen sequences were found to contain a conserved fragment of 200 nucleotides that had on average a 70 % AT content (Fig. 3). The

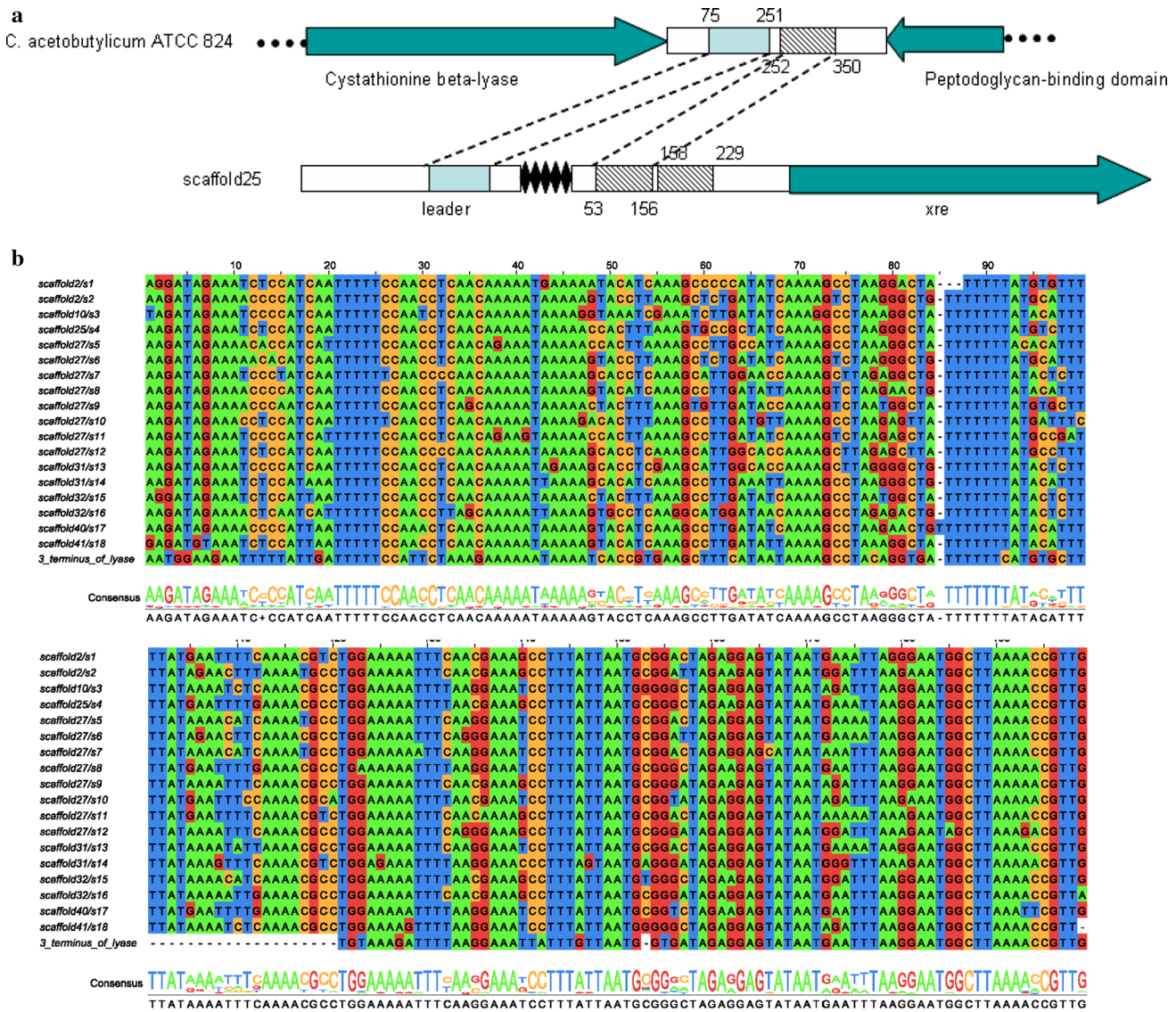

Fig. 3 **a** The result of homology analysis between one CRISPR locus and the non-coding sequence of *C. acetobutylicum* ATCC824. Genes are shown as *arrows*, repeat/spacer units in the CRISPR loci are illustrated as *black diamonds*, and the non-coding sequences including the leader/promoter are depicted as a *box*. **b** Alignment of the 18 non-coding sequences at the 5' terminus of the multiple CRISPR loci and one homologous sequence from *C. acetobutylicum* ATCC824. The locations of these non-coding sequences were previously labeled in Fig. 1

CRISPR arrays were further explored by promoter predictions (see "Materials and Methods"), and the results showed a clear promoter signal with a high score value of 0.99 (the range from 0 to 1). Because these non-coding sequences serve as promoters for the CRISPR loci [24, 25], we suppose that the multiple CRISPR loci are likely under the regulation of the same transcription factor when they are activated. The conserved sequences were also explored in the nr database by blastn for any homology with other sequences. It was found that these conserved sequences displayed a unique sequence similar to that of the three previously published *C. acetobutylicum* genomes. As shown in Fig. 3a, these unique non-coding sequences are located between the cystathionine beta-lyase gene and the peptidoglycan-binding gene and align with the conserved sequences, except for a gap of 20 nucleotides. In addition, a downstream fragment of 100 nucleotides is also homologous to the non-coding sequence that is found between the CRISPR array and the *XRE* gene in scaffold25.

Cas gene

Eight Cas genes were identified by blastp and interproscan in the present *C. acetobutylicum* genome. These included two cas6 genes, csm1, csm2, csm3, csm4, csm5 genes and a gene that is a member of the cas02710 family. The GC

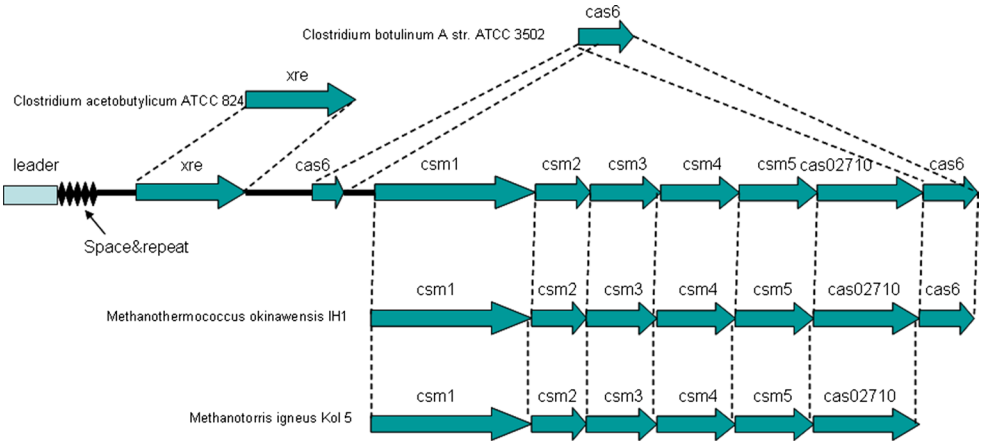

Fig. 4 Homology analysis of the CRISPR system in the *C. acetobutylicum* genome used in this study with the cas6 gene from the *C. botulinum A str.* ATCC 3502, the *XRE* gene from *C. acetobutylicum* ATCC824 and the two CRISPR systems from *Methanothermococcus okinawensis* H1 and *Methanotorris igneus* Kol 5 strains

content of these genes is 25.2 % which is significantly lower than 31.2 % of genome overall ($p < 0.01$). However, they shared only a low degree of similarity at the amino acid level with the relevant homologous sequences, ranging from 40 to 70 %. As described above, they are clustered together and belong to the same operon. Except for the cas02710 (PF09670) gene family that is only found in a few species, csm1–csm5, members of the repeat-associated mysterious protein (RAMP) family and the cas6 gene are widely detected in bacteria and archaea (Fig. 4).

The Cas genes of *C. acetobutylicum* GXAS18-1 lack homologous sequences with the *Clostridium* genus genomes, suggesting that these genes might be captured from another genus by horizontal gene transfer (HGT). A gene tree of Cas was created and compared to the species tree so as to explore any potential HGT relationship (refer to Materials and Methods), and the results are shown in Fig. 5. In contrast to the species tree, where *C. acetobutylicum* clusters with the *Clostridia* and the *Bacilli* genera, the cluster of Cas genes of this present genome showed the closest genetic connection to the genera of *Methanotorris igneus* Kol 5, *Methanococcus arolicus* Nankai-3, *Methanothermococcus okinawensis* H1 and *Methanocaldococcus* (*M.*) such as *M. vulcanius* M7, all of which belong to the euryarchaeota, especially the *Methanococcus arolicus* Nankai-3 and the *Methanothermococcus okinawensis* H1 strains, which contain 6 Cas genes and are in the same gene cluster order csm1-csm2-csm3-csm4-csm5-cas02710 as that seen for the Cas genes of the strain studied in this report. The above findings provided strong evidence that the Cas genes in the present strain were transferred from the euryarchaeota.

Discussion

CRISPR is a genetic immune system that was recently discovered. CRISPR maintains the genetic memory by acquiring new repeat-spacer units to prevent invading DNA such as transformation from environmental DNA, conjugation by plasmids and transduction by bacteriophages [26]. The CRISPR system was thought to be involved in the arms race between host and rapidly evolving phages or environment changes [27, 28]. This arms race conferred the CRISPR system with a high level of diversity in the sequences of the repeat-spacer, the leader and the Cas genes, even within closely related strains. Thus, it allows for the CRISPR loci to be widely used in high-resolution genotyping and forensic medicine, such as in the spoligotyping technique. The elucidation of the CRISPR mechanism has become increasingly important for industrial bacteria utilized in the food or biofuel industries [5].

In this study, we are the first to report a CRISPR system in *C. acetobutylicum*. With three other published genomes of *C. acetobutylicum*, the ratio of CRISPR carriers in *C. acetobutylicum* strains is 25 % that lower than CRISPR carriers in of bacterial (~40 %) and archaea (90 %) compared to previous studies [5, 26, 29, 30]. An explanation might be the number bias of genome sequencing [8]. More CRISPR systems in *C. acetobutylicum* strains may be found when the number of *C. acetobutylicum* genomes was accumulated. The other reason might be that the reported strains were established under laboratory conditions for a long time. Having grown in an environment that lacks exposure to bacteriophages has resulted in the loss of the CRISPR system in the bacteria during rapidly adaptive evolution [26]. Such phenomena were also observed in a recent study on *Mycoplasma gallisepticum*, in which fast evolution and the loss of the CRISPR system were detected after host shifts. However, the associated absence of any CRISPR signals in other *C. acetobutylicum* strains and the evidence from the phylogenetic analysis show that the CRISPR system in the present strain is probably acquired from different genera such as euryarchaeota by HGT as a result of the present strain being exposed to a different environment. This finding was also supported by the evidence from the genes found in the genome of the present strain (unpublished

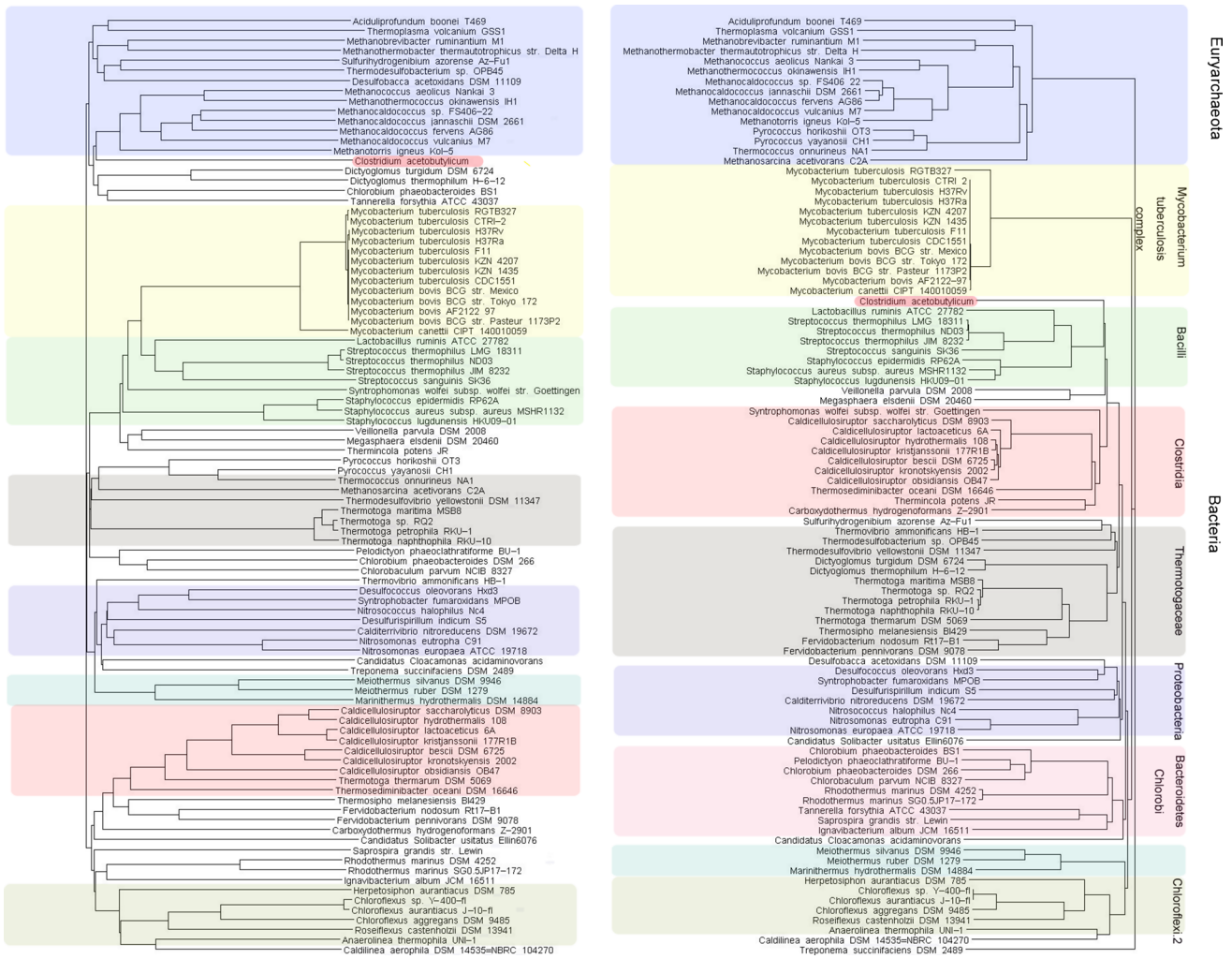

Fig. 5 Comparison of gene tree and species clustering tree. On the *left* is a gene tree constructed using concatenated Cas genes; and on the *right* is a species tree which is constructed from the concatenation of the 31 conserved orthologous genes

data), in which many genes were annotated as bacteriophage genes but lacked any homologous genes with the other three *C. acetobutylicum* strains.

To date, many CRISPR mechanisms remain elusive. In general, the complete CRISPR system can be divided into three basic functional modules based on previous studies [26, 31], namely insertion of new spacers, expression and processing of CRISPR RNA, and CRISPR interference. As a result of the CRISPR system participating in a co-evolutionary battle with rapidly evolving viruses, this has resulted in complex Cas gene families, repeat clusters derived from frequent HGT and micro-recombination in the repeat/spacer units and the Cas genes [7, 27, 32]. Forty-five different Cas gene families [8] and 33 repeat clusters [6] can be found in current CRISPR/Cas database. In the present strain, except for lacking the cas1 and cas2 genes, all the other Cas genes can be proven as typical of the *Mycob. tuberculosis* subtype-like reference strains: namely *Mycob. tuberculosis* CDC1551 and *Mycob. tuberculosis* H37Rv, which contain the csm1–csm5, the cas02710 and the cas6 genes. Furthermore, the weaker linkage between the cas1-cas2 and the subtype-specific genes was also widely observed in many organism of the *Mycob. tuberculosis* subtype, in which the cas1-cas2 genes may associate with other CRISPR loci or operons [4, 8]. Here, we propose that the absence of the cas1-cas2 gene in *C. acetobutylicum* might result from a weak link between the cas1-cas2 genes and subtype-specific genes, which has resulted in the *C. acetobutylicum* having either acquired the CRISPR/Cas system without the cas1-cas2 genes or lost both the relevant genes during subsequent evolution.

As described above, only 4 out of 80 spacers were found to match known gene sequences, and all were shown to be present in the three published genomes of the *C. acetobutylicum* strains. However, none of the spacers were found to match with phage genome sequences in the current

database. This is consistent with a previous study, in which only 47 matched sequences (<1 %) were confirmed from bacteriophage genes among 88 organisms with 4500 spacers reviewed [33]. Taken together the small-scale sequencing of the phage genome [34], it is possible that spacers lack any homologous sequences to phages in this present study. Moreover, four spacers mentioned above lack self-targeting sequence, as two genes are completely lost and two other genes are only partially deleted at the matched regions of the assembled genome. Stern et al. [23] have studied the self-targeting CRISPR spacers from all known CRISPR and have shown that only 0.4 % spacers were self-targeting CRISPR, while most were frequently associated with the partial or full degradation of the CRISPR/Cas activity. Therefore, they postulated a model for CRIPSR autoimmunity and its possible outcomes. In this model, if the fitness cost from autoimmunity was low, the host could thrive due to the benefits from resisting phage DNA. So as to prevent any negative effects, five possible mechanisms of inactivation of the CRISPR/Cas activity could occur. However, this model only considered the negative fitness cost from self-targeting genes. In the present study, our result indicated a potential function of CRISPR autoimmunity that acts as an RNA-programmable genome editor [35]. The carrier could profit from the CRISPR autoimmunity system that can delete self-targeting genes or domains resulting in partial or full abrogation of gene function. This potential function was also shown in a recent study, in which mature crRNAs can be base paired with trans-activating crRNA to form a two-RNA structure that further directs Cas9 to cleave DNA in a site-specific manner [35]. It can also be used to explain why self-targeting spacers are absent in the reported CRISPR system.

Finally, we identified in the present study is a common leader/promoter flanking one end of the repeat-spacers cluster. Out of 23 non-coding flanking sequences adjacent to the repeats-spacers, 18 were found to share a 200-bp conserved homologous sequence, which was found to be an AT-enriched sequence with a clear promoter signal. Although multiple CRISPR loci were found in many other organisms, such as *Methanocaldococcus jannaschii* and *Clostridium difficile* BI1 [36, 37], this is the first reported case of multiple CRISPR loci regulated by the same transcriptional element. As far as the leader sequence functions as a promoter [38–42], it is implied that multiple CRISPR loci are prone to transcribe simultaneously when they are triggered by an invading phage [43–45]. Furthermore, the CRISPR carriers may take advantage of this strategy as a risk diversification mechanism, in which carriers can increase the efficiency of incorporation of new spacers and decrease the risk of defective mutations at the leader sequence or the diversity of the Cas genes.

Acknowledgments This work was supported by BaGui Scholars Program Foundation, Basic Research Fund of Guangxi Academy of Sciences (NO.10YJ25SW13), Guangxi Natural Science Foundation (2013GXNSFBA019106) and Key Technologies R & D Program of Guangxi (10123007-3).

References

1. Beesch SC (1953) Acetone-butanol fermentation of starches. Appl Microbiol 1:85–95
2. Jones DT, Shirley M, Wu X, Keis S (2000) Bacteriophage infections in the industrial acetone butanol (AB) fermentation process. J Mol Microbiol Biotechnol 2:21–26
3. Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8:172
4. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV (2011) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9:467–477
5. Wiedenheft B, Sternberg SH, Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. Nature 482:331–338
6. Kunin V, Sorek R, Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8:R61
7. Vale PF, Little TJ (2010) CRISPR-mediated phage resistance and the ghost of coevolution past. Proc Biol Sci 277:2097–2103
8. Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 1:e60
9. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823
10. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. Science 339:823–826
11. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153:910–918
12. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967
13. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641
14. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59
15. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 33:6494–6506
16. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL (2008) The Vienna RNA websuite. Nucleic Acids Res 36:70–74
17. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41
18. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

19. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599
20. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287
21. Nye TM, Lio P, Gilks WR (2006) A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22:117–119
22. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–3431
23. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet 26:335–340
24. Phok K, Moisan A, Rinaldi D, Brucato N, Carpousis AJ, Gaspin C, Clouet-d'Orval B (2011) Identification of CRISPR and riboswitch related RNAs among novel noncoding RNAs of the euryarchaeon Pyrococcus abyssi. BMC Genom 12:312
25. Westra ER, Pul U, Heidrich N, Jore MM, Lundgren M, Stratmann T, Wurm R, Raine A, Mescher M, Van Heereveld L, Mastop M, Wagner EG, Schnetz K, Van Der Oost J, Wagner R, Brouns SJ (2010) H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. Mol Microbiol 77:1380–1393
26. Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11:181–190
27. Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64:475–493
28. Vale PF, Little TJ (2010) CRISPR-mediated phage resistance and the ghost of coevolution past. Proc Biol Sci 277:2097–2103
29. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315:1709–1712
30. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol 6:181–186
31. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 34:401–407
32. Takeuchi N, Wolf YI, Makarova KS, Koonin EV (2012) Nature and intensity of selection pressure on CRISPR-associated genes. J Bacteriol 194:1216–1225
33. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60:174–182
34. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510
35. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821
36. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073
37. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeno-Tarraga AM, Wang H, Holden MT, Wright A, Churcher C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J (2006) The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet 38:779–786
38. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321:960–964
39. Hale C, Kleppe K, Terns RM, Terns MP (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. RNA 14:2572–2579
40. Lillestol RK, Shah SA, Brugger K, Redder P, Phan H, Christiansen J, Garrett RA (2009) CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. Mol Microbiol 72:259–272
41. Pul U, Wurm R, Arslan Z, Geissen R, Hofmann N, Wagner R (2010) Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. Mol Microbiol 75:1495–1512
42. Semenova E, Nagornykh M, Pyatnitskiy M, Artamonova II, Severinov K (2009) Analysis of CRISPR system function in plant pathogen *Xanthomonas oryzae*. FEMS Microbiol Lett 296:110–116
43. Agari Y, Kashihara A, Yokoyama S, Kuramitsu S, Shinkai A (2008) Global gene expression mediated by *Thermus thermophilus* SdrP, a CRP/FNR family transcriptional regulator. Mol Microbiol 70:60–75
44. Agari Y, Sakamoto K, Tamakoshi M, Oshima T, Kuramitsu S, Shinkai A (2010) Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. J Mol Biol 395:270–281
45. Shinkai A, Kira S, Nakagawa N, Kashihara A, Kuramitsu S, Yokoyama S (2007) Transcription activation mediated by a cyclic AMP receptor protein from *Thermus thermophilus* HB8. J Bacteriol 189:3891–3901