Journal of Advances in Microbiology
1(3): 1-13, 2016; Article no.JAMB.31364

SCIENCEDOMAIN *international*
www.sciencedomain.org

Direct Repeats and Spacer Diversities as Found in *Lactobacillus pentosus* KCA1 CRISPR loci and *Cas*9 Structural Signature

Kingsley C. Anukam[1,2*]

[1]Department of Medical Laboratory Science, University of Benin, P.M.B. 1154, Benin City, Edo State, Nigeria.
[2]Uzobiogene Genomics, London, Ontario, Canada.

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMB/2016/31364
Editor(s):
(1) Ana Claudia Correia Coelho, Department of Veterinary Sciences, University of Trás-os-Montes and Alto Douro, Portugal.
Reviewers:
(1) Mahmoud M. M. Zaky, Port-Said University, Egypt.
(2) Fakruddin, Institute of Food Science and Technology (IFST), Bangladesh Council of Scientic and Industrial Research (BCSIR), Dhaka, Bangladesh.
(3) Dennis A. Bazylinski, University of Nevada at Las Vegas, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history/17955

Original Research Article

Received 2nd January 2017
Accepted 6th February 2017
Published 24th February 2017

ABSTRACT

Aim: To provide an insight on Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) direct repeats and spacer organization present in *Lactobacillus pentosus* KCA1. Little or no information exist on CRISPR and CRISPR-associated (*Cas*) genes in *Lactobacilli* as a vista of new technique for potential genome editing.
Methodology: The genome sequence of *Lactobacillus pentosus* KCA1 was obtained from gene bank and bioinformatic tools such as CRISPR finder was used to analyze direct repeats and spacer diversities. CRISPR Target tool was used for spacer target prediction. The iterative threading assembly refinement (I-TASSER) algorithm was used for the prediction of the *Cas*9 3D structure, and the binding site by this integrated algorithm. CueMol2 version 2.2.2.366 was used to visualize the binding sites.
Results: The genome of *Lactobacillus pentosus* KCA1 encodes six CRISPR arrays with direct repeat (DR) length consensus exhibiting repeat polymorphisms ending with adenine-guanine (AG) nucleotide except CRISPR1 that has mainstream DR length ending with adenine-cytosine (AC) and

**Corresponding author: E-mail: kingsley.anukam@uniben.edu, kanukam@gmail.com;*

terminal repeats ending with adenine-thymine (AT). The domain of KCA1_Cas9 has 153 amino acid residues belonging to HNH Cas9-type, located between positions 786-938. KCA1_0112 Cas9 has one chain A architecture and three putative ligand binding sites residues located at 10 (Asp), 777 (Glu) and 992 (Asp) positions of the 4086 base pairs.
Conclusion: Spacer analysis revealed that *L. pentosus* KCA1 may have been exposed to several mobile genetic elements as the spacers matched sequences from bacteriophages and plasmids. Further studies are needed to explore the structural architecture of KCA1 Cas9 as a potential part of CRISPR genome editing functionality.

Keywords: CRISPR; Cas9; genome editing; Lactobacillus pentosus; repeat regions; spacer diversity.

1. INTRODUCTION

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and CRISPR-associated proteins have evolved as immunity repertoire widely distributed in bacteria and archaea [1] against invading viruses and mobile genetic elements such as bacteriophages, viruses and plasmids. The CRISPR array consists of strings of direct repeats sequences, typically 21-48 base pairs [2] that are interspaced with short distinctive spacer sequences derived from foreign DNA. Adjacent to the CRISPR loci, are cas gene cassettes [3] that function with generated CRISPR RNAs (crRNAs) sequel to transcription and maturation of the CRISPR locus. These systems act as reconnaissance for the detection and destruction of foreign DNA by RNA guided interference.

There are three known CRISPR–Cas system types (I, II and III) and each of these uses distinct molecular mechanisms for obstructing the integration of the foreign DNA elements. Type I and type III systems use a large complex of Cas proteins for crRNA-guided targeting, whereas the type II system requires only the Cas9 protein for RNA‑guided DNA recognition and cleavage, a property that has proved to be extremely useful for genome engineering applications [4].

In the last decade, significant efforts have been made in recognizing that the sequences of the CRISPR spacers were similar to sequences from mobile genetic elements including but not limited to plasmids and viruses [5,6]. In a study that involved phage-challenge experiment, Barrangou et al. [7] reported that CRISPR loci acquire fragments of invading DNA and that these new spacers result in sequence-specific resistance to the phage that is infecting the bacterium. There is a correlation between CRISPR spacers and phage susceptibility and exposure [8].

The CRISPR-Cas9 system introduces double-strand DNA breaks at a specific locus in the genome by using a complex of the Cas9 nuclease with either a chimeric single guide RNA (sgRNA) or two short RNAs (a CRISPR RNA (crRNA) and a trans-activating RNA (tracrRNA) [9].

The discovery of CRISPR-Cas created a simple two-component system in which changes to the 20-nucleotide guide sequence of the sgRNA can programme Cas9 to target any DNA sequence of interest as long as it is adjacent to a protospacer adjacent motiff (PAM). The CRISPR–Cas9 system only requires a change in the guide RNA sequence to alter target specificity [10].

There are distinct polymorphic CRISPRs that have been discovered in several organisms notably *Methanocaldococcus jannaschii* with 18 CRISPR loci [11], and *Streptococcus thermophilus* having 3 CRISPR loci [8]. Recent studies have employed the polymorphic characteristics as a genotyping tool for some species of *Lactobacilli* [12]. However, little information is available on the nature of the CRISPR-Cas repertoire in *Lactobacillus pentosus* KCA1. The objective of this study was to provide an insight on CRISPR direct repeats and spacer organization present in *Lactobacillus pentosus* KCA1, a recently sequenced bacterium isolated from the vagina of a healthy Igbo-African women [13].

2. MATERIALS AND METHODS

The genome sequence of *Lactobacillus pentosus* KCA1 was obtained from the gene bank through the National Center for Biotechnology Information's website (http://www.ncbi.nlm.nih.gov/nuccore/NZ_CM001538.1) with accession number NZ_CM001538 NZ_AKAO01000000. The six CRISPR arrays or the repeat regions were downloaded from

http://www.ncbi.nlm.nih.gov/nuccore/393256311?report=graph. In order to confirm the identification of the CRISPR arrays, the whole genome comprising of 3,426,323 base pairs was downloaded. The CRISPR database CRISPRdb [2] and CRISPRFinder were used to identify CRISPR loci in the published *Lactobacillus pentosus* KCA1 genome. CRISPR finder online software was used to identify the repeat regions, mostly the CRISPR start and end positions, direct repeat (DR) length and direct length consensus and number of spacers. Basic local alignment sequence tool (BLAST) algorithm [14] was used to compare and contrast the sequences of CRISPR repeats, CRISPR spacers and Cas9 genes against other bacteria that are closely related.

The spacer target prediction was done by using the CRISPRTarget online tool. (http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis) [15].

The iterative threading assembly refinement (I-TASSER) algorithm [16], which is a four stage integrated platform for automated protein structure and function prediction based on the sequence-to-structure-to-function paradigm was used. The amino acid sequence of *L. pentosus* KCA1_0112 (Cas9) was submitted to the swiss-model server for prediction of the 3D structure, similar structures in PDB, function and the binding site by this integrated algorithm. The Swiss-model server (http://swissmodel.expasy.org) compares with non-redundant set of protein database (PDB) structures [17]. CueMol2 version 2.2.2.366 was used to visualize the binding sites (http://www.cuemol.org).

3. RESULTS

The genome of *Lactobacillus pentosus* KCA1 encodes six CRISPR arrays in the chromosome (Fig. 1).

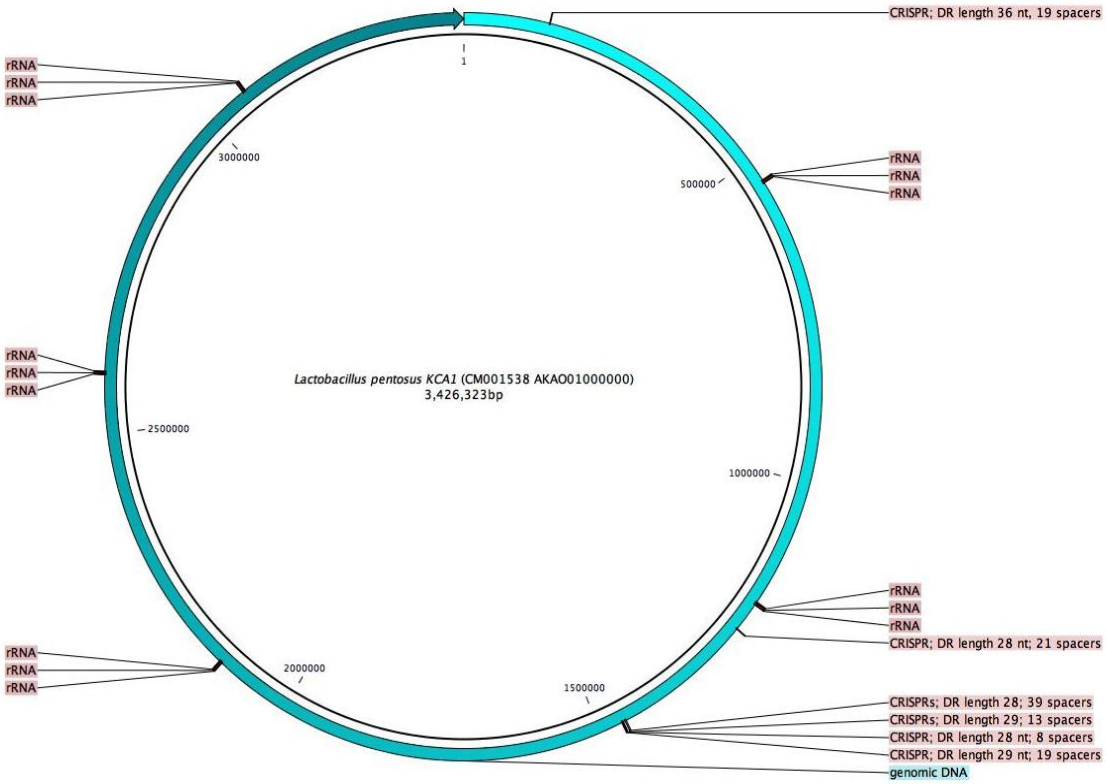

Fig. 1. The location of the six CRISPR loci in the *L. pentosus* KCA1 chromosome and the location the five rRNA loci. CLC sequence viewer version 7.7 was used

The leader sequence preceding CRISPR1 is composed of AT-rich 24 nucleotides "TTAGTGGATTATGTGAGAATGCCG".

The first CRISPR1 is located between 131563-132851 and has a length of 1289 base-pairs constituting DR length consensus of 36 nucleotides (nt), and 19 spacers. Preceding this CRISPR are four CRISPR-associated proteins: KCA1_0112 (CRISPR-associated protein, *Cas*9), KCA1_0113 (CRISPR-associated protein *Cas*1), KCA1_0114 (CRISPR-associated protein *Cas*2), KCA1_0115 (CRISPR type II-A-associated protein *Csn*2) as shown in Fig. 2.

The second CRISPR (1306 base-pairs) is found between 1239838-1241143 with DR length 28 nt and 21 spacers. The DNA polymerase III, alpha subunit preceded this CRISPR, followed by two ribosomal subunits (rpmI -LSU ribosomal protein L35p, rplT-LSU ribosomal protein L20p). The third CRISPR is located between 1456695-1459056 and has 2362 base-pairs with DR length 28nt and 39 spacers. This CRISPR3 is followed by two *cas* proteins: KCA1_1291 (CRISPR-associated protein *Cas*1), KCA1_1292 (CRISPR-associated protein *Cas*2). CRISPR4 (1461724-1462548) has a length of 824bp and DR length consensus of 29nt with 13 spacers. Immediately following CRISPR4 is a sequence of unknown function

["CAACGGCGCAACCATTGATTT-n (99) GGAGCAATCATCTGGCGTCGTAATGGCTTTC G"].

CRISPR5 (1462702-1463217) appears to be the smallest in terms of number of nucleotide base pairs (515bp) with 28 DR length consensus and 8 spacers. Between CRISPR5 and CRISPR6 is another sequence of unknown function.

["GACGTATCAGCTGTGGGCTGGGT-n(101)-CGACTTCATT"]. The last CRISPR6 is found between 1463351-1464538 and is composed of 1187 bp with 29 DR length and 19 spacers. Downstream CRISPR6 is the six *Cas* cassettes comprising cas3, Cse1, Cse2, Cas7, Cas5 and Cas6.

Interestingly results from CRISPR blast revealed that CRISPR1 has 100% identity to CRISPR_id NC_020229_1 of *Lactobacillus plantarum* ZJ31 with an E-value of 4.00E-13. CRISPR2 and CRISPR3 have 92.86% identity to CRISPR_id NC_009953_5 of *Salinispora arenicola* CNS-205. Similarly, CRISPR4, CRISPR5 and CRISPR6 have over 92% identity to CRISPR_id NC_010610_2 of *Lactobacillus fermentum* IFO-3956 (Table 1).

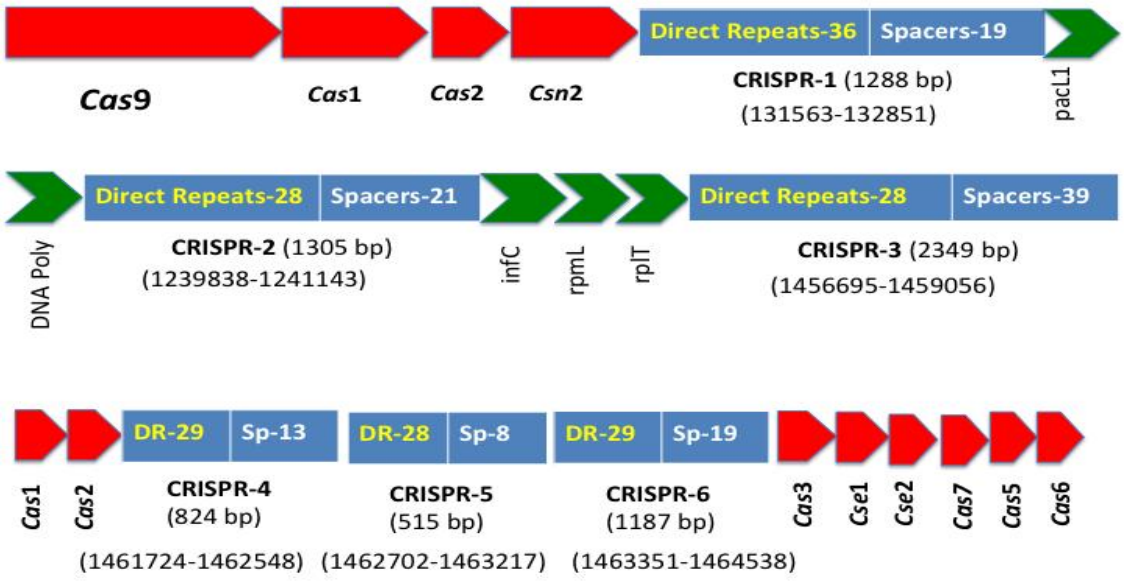

Fig. 2. *Lactobacillus pentosus* KCA1 CRISPR organization showing the location of direct repeats, spacers and the associated *Cas* genes and ribosomal proteins

Table 2 shows the frequency of CRISPR repeats polymorphisms indicating that CRISPR1 has 95% of mainstream repeats and 5% of terminal repeat variants at 12 locations of the DR length consensus. Mainstream or typical repeats of CRISPR2 and CRISPR3 are 100% identical while the initiation repeats of CRISPR2, CRISPR3 and CRISPR4 have initiation variants. CRISPR4 has more variations in the DR repeats with only 57.1% representing mainstream repeats. CRISPR5 is the only CRISPR locus (100%) without any variants or polymorphism throughout the DR repeats. The CRISPR's DR length consensus ended with adenine-guanine (AG) nucleotide except CRISPR1 that has mainstream DR length ending with adenine-cytosine (AC) and terminal repeats ending with adenine-thymine (AT).

3.1 CRISPR Spacers Analysis

Spacer polymorphisms were observed in relation to the spacer size. A total of 118 spacers were identified with the CRISPR spacer finder software. CRISPR-1 has 19 spacers, of which 18 spacers were composed of 30 base pairs and 1 spacer has 29 bp as shown in Fig. 3. CRISPR-3 has 38 spacer polymorphisms than other CRISPRs comprising 35 spacers with 33 bp, 2 spacers with 34 bp and 1 spacer with 35 bp. CRISPR-4 has 13 spacers with 10 having 32 bp and 3 with 33 bp while CRISPR-6 has 19 spacers with 32 bp respectively.

3.2 Spacer Target Prediction

Attempts were made to identify similarity of the spacer sequences to putative phages, plasmids and viruses that may have been exposed to the host genome. All the 118 spacers were analyzed with the CRISPRTarget tool (Supplementary Table 1). Of the 19 spacers present in CRISPR1, 3 spacers matched to sequences of *Lactobacillus* bacteriophage phig1e, while 1 spacer was similar to sequences of *Acetobacter pasteurianus* plasmid pAC258-29. CRISPR2 with 21 spacers, spacer (no. 3, 13) matched to *Lactobacillus* bacteriophage phig1e, spacer (no. 15, and 17) had similar sequence to bacteriophage B103, *Bacillus* phage Nf, and Bacillus phage MG-B1.

CRISPR3 with 38 spacers out of which 4 spacers (nos. 29, 30, 36, 37) had matches similar to sequences of *Lactobacillus* bacteriophage phig1e, spacer (no.31) to *L. plantarum* bacteriophage phiJL-1, spacer (no.25) to *Lactobacillus plantarum* 16 plasmid Lp16I, spacer (no.19) to *Lactobacillus johnsonii* prophage Lj771 and *Lactobacillus* phage phi jlb1, spacer (no.15) to *Lactobacillus* phage ATCC 8014-B1, *Pediococcus* phage clP1, and *Lactobacillus plantarum* bacteriophage phiJL-1.

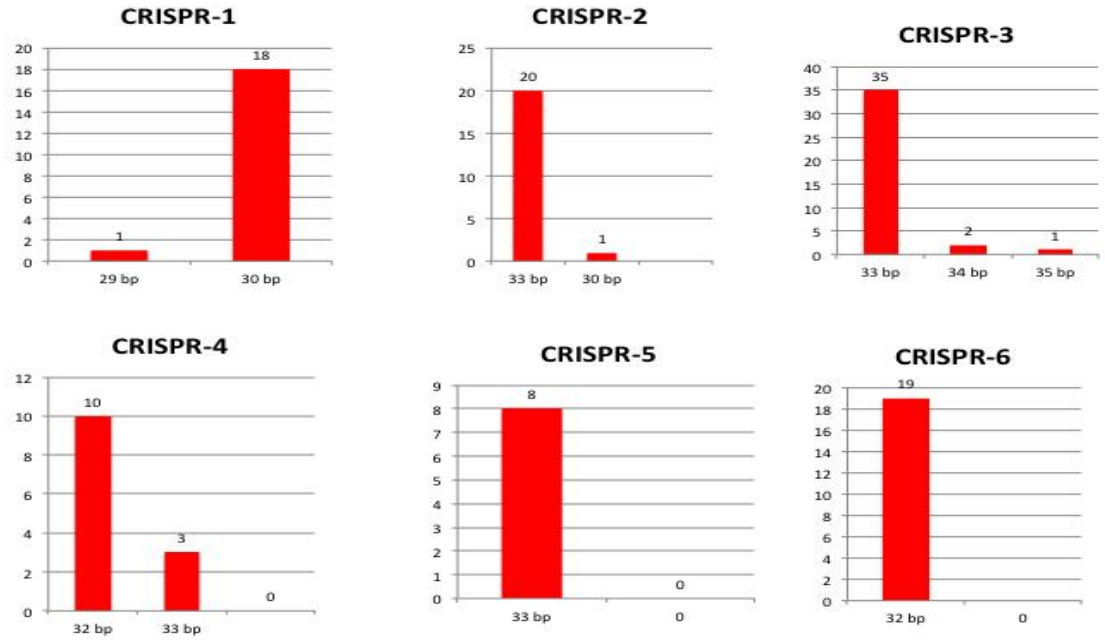

Fig. 3. The spacer polymorphisms in the six CRISPR arrays

Of the 19 spacers present in CRISPR-6, spacer number 3 matched to 4 *Campylobacter* phages (CJIE4-5, CJIE4-4, CJIE4-1, CJIE4-3), spacer (no.4) to *Burkholderia sp*. RPE67 plasmid p1 DNA, spacer (no.6, 14) to *Lactobacillus plantarum* bacteriophage phiJL-1, spacer (no.12) to *Lactobacillus* bacteriophage phig1e, spacer (no. 17) to *Lactobacillus* bacteriophage phig1e and *Streptococcus* phage SMP respectively, spacer (no. 18) to *Lactobacillus* bacteriophage phig1e and *Sinorhizobium meliloti* Rm41 plasmid pRM41A. Spacer no. 11 of CRISPR4 had similarity match to four *Bacillus thuringiensis* strain YC-10 plasmid pYC4, *B. thuringiensis* serovar galleriae HD-29 plasmid pBMB47, *B. thuringiensis* serovar kurstaki HD 1 plasmid unnamed7, *Bacillus thuringiensis* serovar kurstaki str. HD-1 plasmid pBMB46 and to *Bacillus* phage Waukesha92 and *Bacillus* phage phiS58. In the same manner, spacer (no.5) of CRISPR5 matches to *Lactobacillus* bacteriophage phig1e and to *Lactobacillus* bacteriophage phig1e DNA for Rorf162, Holin, Lysin, and Rorf175 genes (supplementary text material showing CRISPR Target analysis indicating the 5' end crRNA flanking protospacer, 3' end crRNA flanking protospacer and 5' end of the sequence spacer).

The KCA1_0112 cas9 is composed of 4086 base pairs, which translates into 1361 amino acids with a mass of 157, 139 Dalton. The domain of KCA1_0112 cas9 has 153 amino acid residues belonging to HNH Cas9-type which is known to cleave the target DNA complementary to crRNA and is located between positions 786-938 as shown in Fig. 4.

The SWISS-MODEL 3-D tool indicated a 36.17% sequence similarity to 4zt0.1.A protein template with normalized QMEAN4 score and the position of the KCA1 Cas9 model in red star (Fig. 5).

The iterative assembly (I-TASSER) for the KCA1_0112 Cas9 protein revealed a confidence score (C-score) of 0.79 and an estimated template modeling score (TM-score) of 0.82±0.08 and estimated root mean square deviation (RMSD) = 7.8±4.4Å relative to the template protein (*4cmqA*). The KCA1_0112 Cas9 has three potential ligand binding sites residues located at 10 (Asp), 777 (Glu) and 992 (Asp) positions of the 4086bp. (Fig. 6).

```
        10                  20                  30                  40                  50
MNEPYGVGLD IGTNSVGWTV VDMNGRVRKV KGKTALGARL FKEGATAEDR
        60                  70                  80                  90                 100
RGFRTTRRRL KRVKWRLRLL REFFDQPISK IDPNFFARRK YSDISPRDPN
       110                 120                 130                 140                 150
YNGLAKTLFN DRTDKEFYDD YATIYHLRDK LMTSNRKFDI REIYLAIHHI
       160                 170                 180                 190                 200
VKYRGNFLRT GPASQYGSAS LHLATSFQKL NDLFAQSEET LNLKLVTDEA
       210                 220                 230                 240                 250
LLQQIQQILV RTDLSRSEQQ RQIWPLMAVL TGASAAEKKR QKNVVVELSK
       260                 270                 280                 290                 300
ALVGLKAKMN VVTLTEVDAA VVKDWTFTME ESQDKLPEIE EQLSEVGQQI {Coiled coil
       310                 320                 330                 340                 350 (279-299) Length 21}
MDEVIQLYAS VNLAQLIPAG KRFSQHMVEK YKHHEKNLEL LKAYIHSQSD
       360                 370                 380                 390                 400
SKRGREIRAT YDRYIDGVDS KPVTQEMFYK DLMKYVEADA TSNHLAAEIK
       410                 420                 430                 440                 450
DEIDSEQFMP KLRTKQNGSI PYQVQQYELD QIISHQKKYY PWLGEENPVA
       460                 470                 480                 490                 500
ERRGKFPYKL DELVGFRVPY YVGPLITKED QQATSGAGFA WMVRKADGPI
       510                 520                 530                 540                 550
TPWNFDQKVD RIASATAFIQ RMQTTDTYLI GEDVLPARSL IYQRFMVLNE
       560                 570                 580                 590                 600
LNNMRVEDRK LAPQQKQRLY NQVFKQHQHV SVKNIQQNLM DAGEYRKTPQ
       610                 620                 630                 640                 650
ITGLADPKGF NSSLSTYHDF KKILLEAIAD EHKRADIEKI ILWSTTFEDS
       660                 670                 680                 690                 700
AIFKQKLEEV AWLTDAQRKQ LSGLRYRGWG QLSHKLLTAF KDDKGRSIMD
       710                 720                 730                 740                 750
GLWETSDNFM QLRKQPAIET QIKEANQANL TDADIQDTIN ELYTSPQNKK
       760                 770                 780                 790                 800
AIREVMLVLD DIKNAMHGQT PSWIFVEAAR GGGVAGRRTQ SRSSQIVEAY {Domain region
       810                 820                 830                 840                 850 (786-938) length 153
KGTAKEIVSE KVQHELNEKI KAKADFNTRL VLYFLQNGRD LYTNEAINID (HNH Cas9-type)}
       860                 870                 880                 890                 900
RLSEYDIDHI LPQSLVKDDS LDNRVLTSAR INREKNDTFA SEKFGRKMGA
       910                 920                 930                 940                 950
QWRELHRNGL MTQRKLKHLL MRPDEISKHA TDFINRQLVE TRQVIKLVEE
       960                 970                 980                 990                1000
LISSEYPAAS IVAVKANLTH QFRQTFNFPK LREVNDYHHA FDASLTAFIG
      1010                1020                1030                1040                1050
MYLLKQYPKL ERFFVYGKFA KQPINLTRFN IIRKLAVAEK PIANIETGEI
      1060                1070                1080                1090                1100
LWDKTADIKY FEKLYNYKRL LVTHEVRENY GAMFKQTLFK ASYNKSKTLV
      1110                1120                1130                1140                1150
PKKNHMETSV YGGYSNQETA YLAIVRIPFK SGFKFIVVGI PTRMVAKIKH
      1160                1170                1180                1190                1200
YQSLGATLNQ ATHKVIEPKF TKISRKTKQT VISDYEVVLP KVYLDQVVRD
      1210                1220                1230                1240                1250
QVKGQMYRFS LGSDKEYHNV QELYLPLSIQ QAFVGHYDES DDQRSNDLVK
      1260                1270                1280                1290                1300
VYDAALKQLQ RYFPLHLSRN FDQVASQAHQ SFEGLKNNVQ TSDKQLGKKE
      1310                1320                1330                1340                1350
VLNSLFVGLH ANATRSNLSV LGMSKDFGRI KSNGITLTDQ AEIIYQSPTG
      1360
LFERKVALKD L
```

Fig. 4. The KCA1 Cas9 nucleotide sequence showing the HNH-type Cas9 domain in green colour

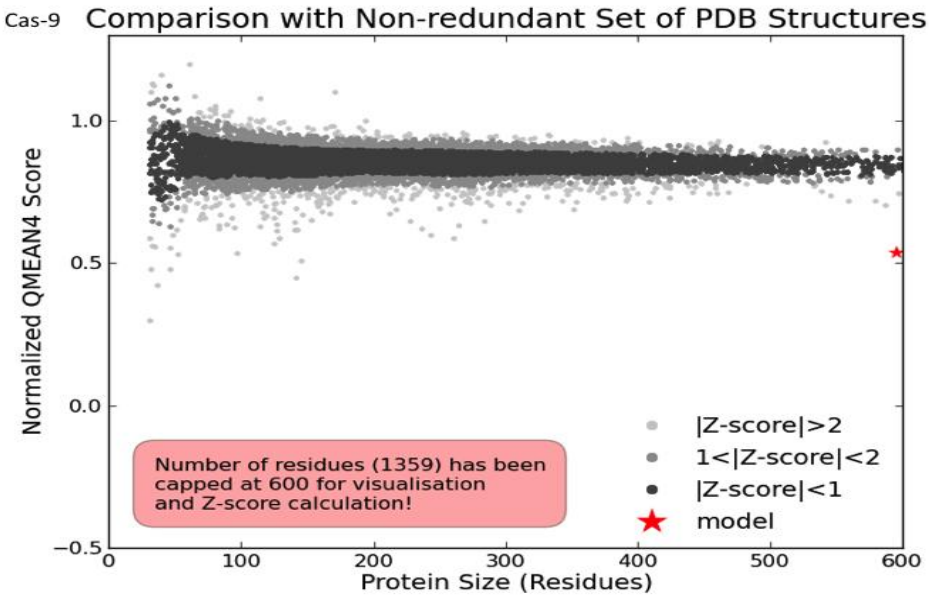

Fig. 5. Comparison of the KCA1 Cas9 protein with the non-redundant set of PDB database structures showing normalized QMean4 score and the position of the KCA1 cas9 model in red star

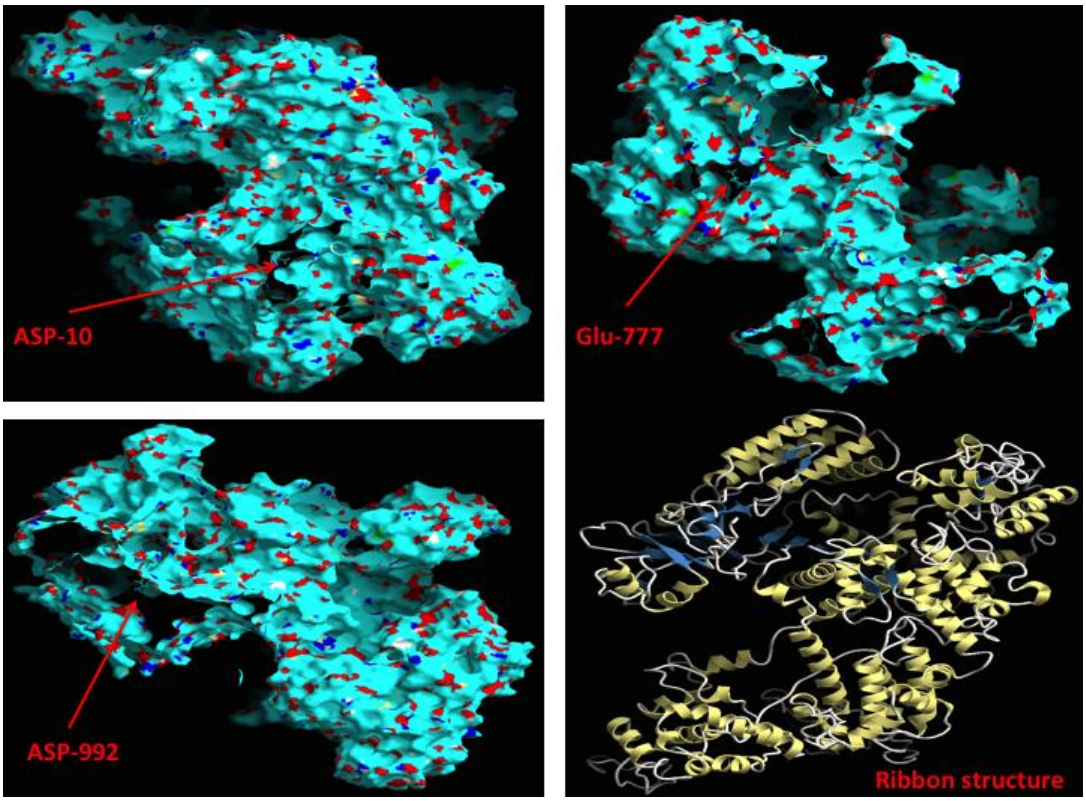

Fig. 6. Ligand binding site as determined by the I-TASSER and visualized with CueMol2 showing surface rendering structures for the Asp-10, Glu-777 and Asp-992 positions and the ribbon structure of the entire KCA1 Cas9

Clustal w phylogenetic alignment tree of cas9 proteins from 10 *Lactobacillus* species including KCA1_0112 cas9 shows that *Lactobacillus pentosus* KCA1 is closely related with 98.8% identity to *Lactobacillus plantarum* EGD-AQ4, 86.6% identity to *Lactobacillus pentosus* DSM 20314, and 86.6% identity to *Lactobacillus pentosus* MP-10 (Fig. 7).

The amino acid composition of KCA1 Cas9 indicates absence of cysteine similar to Cas9 of *L. plantarum* EGD-AQ4. Another distinguishing feature of KCA1 Cas9 is extra acquisition of tyrosine (57 amino acid-tyrosine) when compared with other tested organisms as shown in Table 3.

Comparative analysis using CueMol2 software shows that KCA1_Cas9 has one chain A (Fig. 8), while *Streptococcus pyogenes* Cas9 (4zt0.pdb) has 4 chains (A, B, C, D).

4. DISCUSSION

This is the first description of *Lactobacillus pentosus* KCA1, a vaginal isolate, harboring complete six CRISPR arrays and a type II CRISPR-Cas scheme typified by signature *Cas*9 endonuclease gene, with extensive diversity in spacer repertoire. Previous studies have shown that type II CRISPR- *Cas* systems are evolutionarily widespread and functionally essential in the genus *Lactobacilli* [18]. These Type II systems were detected in 36% of the genus -*Lactobacillus* and associated genera, however type II occur in only 5% of all bacterial genomes analysed to date [19]. However, there are no reports on the CRISPR-*Cas*9 in *Lactobacillus pentosus* species. KCA1 CRISPR 1 locus is a highly conserved 36-nucleotide CRISPR sequence with 100% identity to CRISPR_id NC_020229_1 of *Lactobacillus plantarum* ZJ31, which is a related species. Interestingly, the arrangement of the type II CRISPR architecture in *L. pentosus* KCA1 is similar to other lactic acid bacteria containing in addition to Cas 9, the universal *Cas* 1 and *Cas* 2 genes and the corresponding *Csn* 2 that is distinctive to type II-A systems. It appears that the location of CRISPR1 in KCA1 occurred in the first half of the chromosome with CRISPR-3, -4, -5 and -6 clustering together and having a total of 79 spacers. Notably, CRISPR-3 with the highest 39 spacers was preceded upstream by ribosomal protein rpmL and rplT, thus reflecting the critical role of the CRISPR RNA biogenesis and interference for sequence–specific recognition. Remarkably, 6 *Cas* genes were found downstream CRISPR 6 and the functional roles of the unique organization of these CRISPRs are yet to be determined.

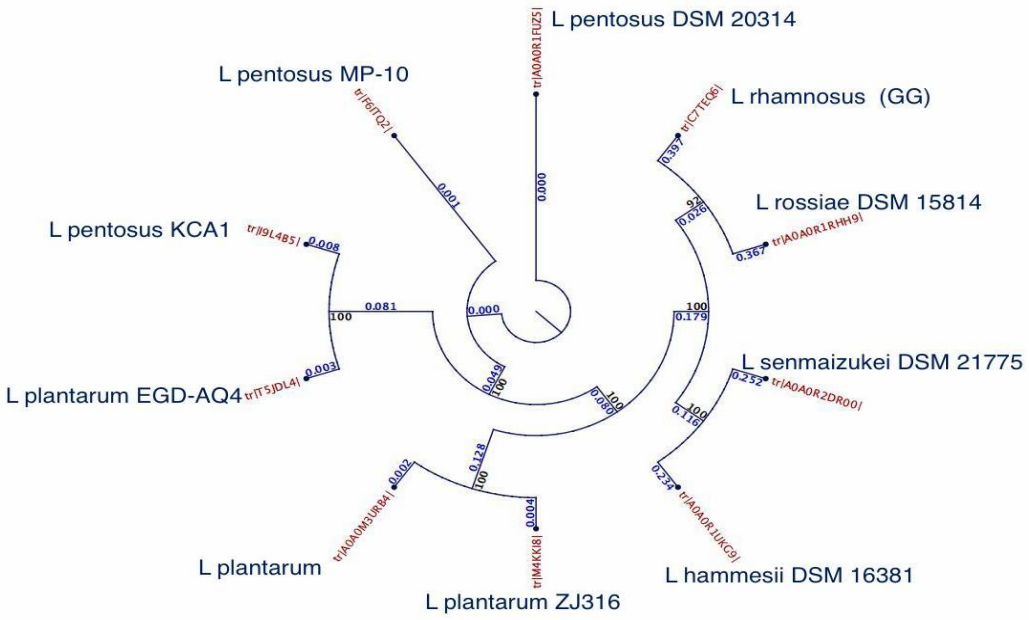

Fig. 7. Phylogenetic tree was created with CLC sequence viewer 7.7 showing circular cladogram layout tree of *Cas*9. Bootstrap values (%) are shown at the branch node in black colour while the branch lengths are shown in blue colour

Table 1. KCA1_CRISPR showing the location, CRISPR length, E-value and similarity to CRISPR identity number

	Location	CRISPR length	Direct Repeat length consensus	No. Spacers	Identity	E-value	CRISPR_id	Organism
CRISPR 1	131563-132851	1288	GTCTTGAATAGTAGTCATATCAAACAGGTTTAGAAC(36)	19	1	4.00E-13	NC_020229_1	*L. plantarum* ZJ31
CRISPR 2	1239838-1241143	1305	GGATCACCCCCGCATACGCGGGGAACAG (28)	21	0.9286	2.00E-04	NC_009953_5	*Salinispora arenicola* CNS-205
CRISPR 3	1456695-1459056	2349	GGATCACCCCCGCATACGCGGGGAACAG (28)	38	0.928571429	4.00E-04	NC_009953_5	*Salinispora arenicola* CNS-205
CRISPR 4	1461724-1462548	824	AGGATCACCCCCGCATACACGGGGAATAG (29)	13	0.9286	1.00E-04	NC_010610_2	*Lactobacillus fermentum* IFO 3956
CRISPR 5	1462702-1463217	515	GGATCACCCCCGCATACACGGGGAATAG (28)	8	0.925925926	3.00E-04	NC_010610_2	*Lactobacillus fermentum* IFO 3956
CRISPR 6	1463351-1464538	1187	AGGATCACCCCCGCATACACGGGGAATAG (29)	19	0.9286	2.00E-04	NC_010610_2	*Lactobacillus fermentum* IFO 3956

Table 2. Analysis of *Lactobacillus pentosus* KCA1 CRISPR repeat sequences showing frequency of repeat polymorphisms. Underlined nucleotides show the deviation from the mainstream repeats

	Type	Repeat sequence	No. of Repeats	Frequency (%)
CRISPR 1	Mainstream repeats	GTCTTGAATAGTAGTCATATCAAACAGGTTTAGAAC	19	95
	Terminal repeats	GTCTTGAATAGTAGTTATTAAATTCGGCTATTAAAT	1	5
CRISPR 2	Initiation	GAATCACCCCCGCATACGCGGGGAACAG	1	4.55
	Mainstream repeats	GGATCACCCCCGCATACGCGGGGAACAG	21	95.45
CRISPR 3	Initiation	AAATCACCCTCGTATGCACGGGGAACAG	1	2.5
	Mainstream repeats	GGATCACCCCCGCATACGCGGGGAACAG	38	97.5
CRISPR 4	Initiation	AAGACTGCCCCCGTATGCACGGGGAATAG	1	7.1
	Mainstream repeats	AGGATCACCCCCGCATACACGGGGAATAG	8	57.1
	Variant repeats	AGGATCACCCCCGCACACACGGGGAACAG	1	7.1
		GGATCACCCCCCGCATACACGGGGAATAG	1	7.1
		GGGATCACCCCCGCATACACGGGGAATAG	3	21.6
CRISPR 5	Mainstream repeats	GGATCACCCCCGCATACACGGGGAATAG	9	100
CRISPR 6	Mainstream repeats	AGGATCACCCCCGCATACACGGGGAATAG	16	80
	Variant repeats	GGGATCACCCCCGCATACACGGGGAATAG	4	20

Table 3. CRISPR-associated endonuclease Cas9 protein statistics

Sequence information										
Protein entry from Uniprot	**I9L4B5**	**T5JDL4**	**A0A0R1FUZ5**	**F6ITQ2**	**A0A0M3URB4**	**M4KKI8**	**A0A0R1UKG9**	**A0A0R2DR00**	**A0A0R1RHH9**	**C7TEQ6**
Length	1361	1362	1358	1358	1358	1358	1362	1345	1378	1363
Name	I9L4B5	T5JDL4	A0A0R1FUZ5	F6ITQ2	A0A0M3URB4	M4KKI8	A0A0R1UKG9	A0A0R2DR00	A0A0R1RHH9	C7TEQ6
Organism	*L. pentosus* KCA1	*L. plantarum* EGD-AQ4	*L. pentosus* DSM 20314	*L. pentosus* MP-10	*L. plantarum*	*L. plantarum* ZJ316	*L. hammesii* DSM 16381	*L. senmaizukei* DSM 21775	*L. rossiae* DSM 15814	*L. rhamnosus* GG)
Weight	157.137	157.256	156.566	156.578	155.858	155.973	157.862	156.151	158.812	156.595
Isoelectric point	9.5	9.49	9.47	9.47	9.68	9.7	9.12	8.78	9.19	9.17
Aliphatic index	84.982	84.919	85.398	85.685	89.632	89.919	86.997	86.439	82.743	87.572
Counts of amino acids										
Amino acid	**I9L4B5**	**T5JDL4**	**A0A0R1FUZ5**	**F6ITQ2**	**A0A0M3URB4**	**M4KKI8**	**A0A0R1UKG9**	**A0A0R2DR00**	**A0A0R1RHH9**	**C7TEQ6**
Alanine (A)	95	95	96	96	101	102	83	89	88	105
Cysteine (C)	0	0	1	1	0	0	2	1	4	1
Aspartic Acid (D)	82	83	79	79	89	89	97	101	98	96
Glutamic Acid (E)	85	86	88	88	70	70	89	86	86	86
Phenylalanine (F)	59	59	60	60	59	60	57	57	69	63
Glycine (G)	65	65	69	69	73	72	69	64	72	71
Histidine (H)	33	33	30	30	20	21	38	38	32	32
Isoleucine (I)	80	80	81	82	86	85	89	82	93	92
Lysine (K)	117	117	115	115	114	114	116	98	124	119
Leucine (L)	129	129	133	133	137	138	140	139	124	144
Methionine (M)	29	29	28	28	26	26	24	28	27	21
Asparagine (N)	63	62	67	67	68	68	56	53	80	69
Proline (P)	41	42	42	42	40	40	47	44	40	42
Glutamine (Q)	88	88	88	88	87	87	82	81	70	72
Arginine (R)	86	87	85	85	90	91	88	97	81	82
Serine (S)	78	78	62	62	54	52	65	64	75	65
Threonine (T)	75	74	88	87	92	91	79	82	80	80
Valine (V)	85	85	79	79	85	86	72	73	71	58
Tryptophan (W)	14	14	14	14	14	14	16	19	16	15
Tyrosine (Y)	57	56	53	53	53	52	53	49	48	50

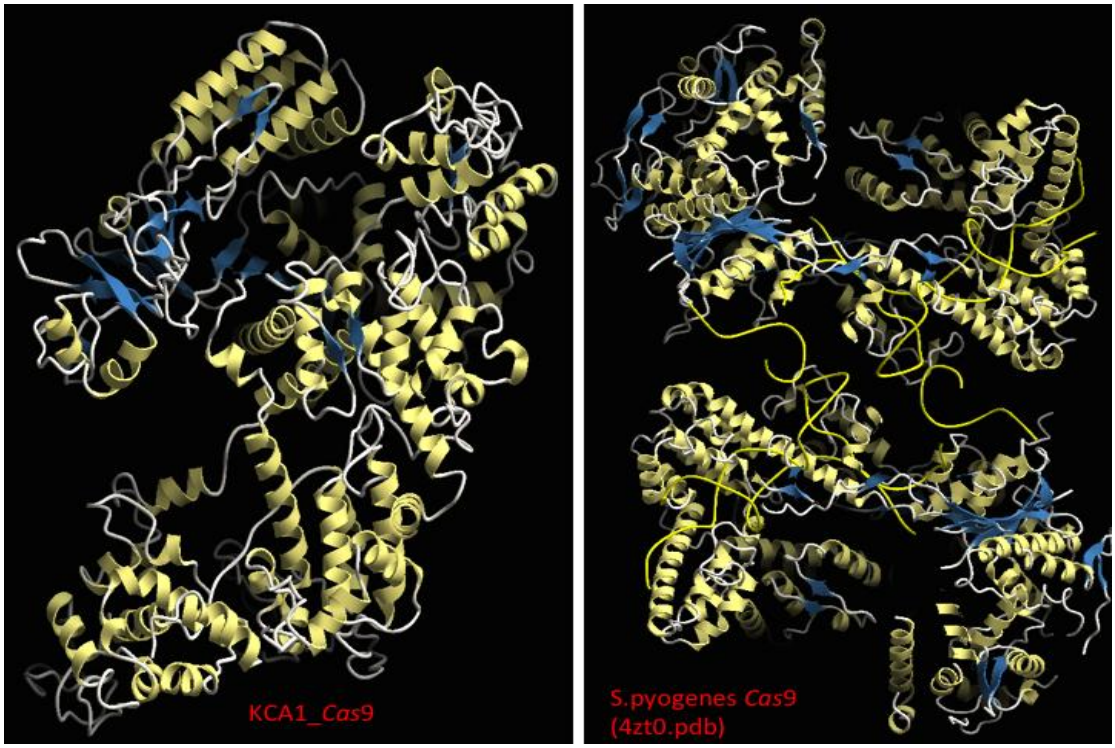

Fig. 8. Comparative ribbon structure of *Lactobacillus pentosus* KCA1 Cas9 and *Streptococcus pyogenes* cas9 (4zt0.pdb)

CRISPR locus is typically defined by the sequence of the repeat. Remarkably, the mainstream repeat sequence of CRISPR-1 was distinct from other CRISPRs with only 28% identity to CRISPR-2 and CRISPR-3 (Table 2). Observing closely, CRISPR-1 is dominated by high sequence degeneracy at the 3' of the terminal repeat, which appears to be in line with the orientation of the neighboring *Cas*9 gene. Conversely, CRISPR-4, CRSPR-5 and CRISPR-6 appear to be enriched with highly conserved mainstream repeat sequence with less polymorphism. It is important to note that the *Lactobacillus pentosus* KCA1 CRISPR mainstream repeat sequence terminals are unique, indicating specific signature associated with KCA1 CRISPRs. In contrast, Sorek et al., [20] reported that several CRISPR repeats from other lactic acid bacteria have a GAAA(C/G) 3' terminus, while *Streptococcus thermophilus* 1 (Sthe1) and *Streptococcus thermophilus* 3 (Sthe3) families identified in the upper suprafamily had a A(A/C)AAC 3' terminus sequence as reported by Horvath et al. [21].

Spacer analysis revealed that *L. pentosus* KCA1 may have been exposed to several mobile genetic elements as the spacers matched sequences from bacteriophages and plasmids. Previous studies have shown that immune function of CRISPR loci has a correlation between phage sensitivity and non-existence of spacers corresponding to the sequence of that particular phage [6]. It is important to note that most of the spacers that matched the sequence of the bacteriophages are phages that infected *Lactobacillus* species of vaginal origin. Some studies done on South African women without bacterial vaginosis revealed abundance of *Lactobacillus*-specific bacteriophages [22]. Similarly, studies from the human microbiome project provided substantial evidence that an abundance of *Lactobacillus*-targeting phage sequences are present in the vagina [23]. However, another notable observation in the spacer sequences that matches to plasmids but did not have the corresponding match to *Lactobacillus* species, suggests that the CRISPR array of *Lactobacillus pentosus* KCA1 has the capability of distinguishing bacteriophage and plasmids. The observed spacer sequences with similarity to proto-spacers of plasmid sequences may have provided immunity or resistance to plasmid elements and it is also possible that the

presence of CRISPR spacers may have been responsible for the absence of plasmids in the genome of *Lactobacillus pentosus* KCA1 [13].

The domain region of KCA1 *Cas*9 contained one tryptophan residue at position 902, which may be involved in binding to the proto-spacer adjacent motifs (PAM). A recent study has shown that two tryptophan containing flexible loops in the C-terminal domain of the *Streptococcus pyogenes* Cas9 nuclease lobe are involved in PAM recognition, and mutation of these residues affects both the binding and cleavage of target DNA [24].

5. CONCLUSIONS

Spacer analysis revealed that *L. pentosus* KCA1 may have been exposed to several mobile genetic elements as the spacers matched sequences from bacteriophages and plasmids. The simplicity of *L. pentosus* KCA1 CRISPR-Cas9 system suggests that it could be used in combination with RNA guide sequences for DNA targeting and cleavage and it has the potential to open a new prospect for next-generation Cas9-mediated genome editing. Further studies are needed to ascertain the structural and functional repertoire of *Lactobacillus pentosus* KCA1 Cas9.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Computational Biology. 2005;1:474-483.
2. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics. 2007;8:172–182.
3. Jansen R, Van Embden JDA, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology. 2002; 43:1565–1575.
4. Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nature Biotechnology. 2008;31(9):833-838.
5. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of Molecular Evolution. 2005;60: 174–182.
6. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005;151:2551–2561.
7. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315:1709–1712.
8. Horvath P, Romero DA, Coûté-Monvoisin A, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR Loci in *Streptococcus thermophilus*. Journal of Bacteriology. 2008;190(4):1401-1412.
9. Cheng AW, Wang H, Yang H, Shi L, Katz Y, Theunissen TW, et al. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Research. 2013;23(10):1163-1171.
10. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507:62–67.
11. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG. et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science. 1996;273:1058–1073.
12. Briner AE, Barrangou R. *Lactobacillus buchneri* genotyping on the basis of Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Locus diversity. Applied and Environmental Microbiology. 2014;80(3):994-1001.
13. Anukam KC, Macklaim JM, Gloor GB, Reid G, Boekhorst J, Renckens B, et al. Genome sequence of *Lactobacillus pentosus* KCA1: Vaginal isolate from a healthy premenopausal woman. PLoS One. 2013;8(3):e59239.
14. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped

BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1997;25:3389–3402.

15. Biswas A, Gagnon JN, Brouns SJ, Fineran PC, Brown CM. CRISPR Target: Bioinformatic prediction and analysis of crRNA targets. RNA Biology. 2013;10(5): 817–827.
16. Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. Nature Protocols. 2010;5:725-738.
17. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. Bioinformatics. 2006; 22:195-201.
18. Sun Z, Harris HMB, McCann A, Guo C, Argimon S, Zhang W, et al. Expanding the biotechnology potential of *Lactobacilli* through comparative genomics of 213 strains and associated genera. Nature Communications. 2015;6:8322.
19. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Research. 2014;42:6091–6105.
20. Sorek R, Kunin V, Hugenholtz P. CRISPR-a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Review Microbiology. 2008;6:181–186.
21. Horvath P, Coûté-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, Barrangou R. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. International Journal of Food Microbiology. 2009; 131(1):62-70.
22. Damelin LH, Paximadis M, Mavri-Damelin D, Birkhead M, Lewis DA, Tiemessen CT. Identification of predominant culturable vaginal *Lactobacillus* species and associated bacteriophages from women with and without vaginal discharge syndrome in South Africa. Journal of Medical Microbiology. 2011;60(2):180-183.
23. Rho M, Wu YW, Tang H, Doak TG, Ye Y. Diverse CRISPRs evolving in human microbiomes. PLoS Genetics. 2012;8: e1002441.
24. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science. 2014;343(6176):1247997.

© 2016 Anukam; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/17955