Knowledge-based discovery for designing CRISPR-CAS systems against invading mobilomes in thermophiles

P. Chellapandi¹ · J. Ranjani¹

Received: 14 February 2015 / Revised: 29 June 2015 / Accepted: 27 July 2015 / Published online: 6 August 2015
© Springer Science+Business Media Dordrecht 2015

Abstract Clustered regularly interspaced short palindromic repeats (CRISPRs) are direct features of the prokaryotic genomes involved in resistance to their bacterial viruses and phages. Herein, we have identified CRISPR loci together with CRISPR-associated sequences (CAS) genes to reveal their immunity against genome invaders in the thermophilic archaea and bacteria. Genomic survey of this study implied that genomic distribution of CRISPR-CAS systems was varied from strain to strain, which was determined by the degree of invading mobiloms. Direct repeats found to be equal in some extent in many thermophiles, but their spacers were differed in each strain. Phylogenetic analyses of CAS superfamily revealed that genes *cmr*, *csH*, *csx11*, HD domain, *devR* were belonged to the subtypes of *cas* gene family. The members in *cas* gene family of thermophiles were functionally diverged within closely related genomes and may contribute to develop several defense strategies. Nevertheless, genome dynamics, geological variation and host defense mechanism were contributed to share their molecular functions across the thermophiles. A thermophilic archaean, *Thermococcus gammatolerans* and thermophilic bacteria, *Petrogla mobilis* and *Thermotoga lettingae* have shown superoperons-like appearance to cluster *cas* genes, which were typically evolved for their defense pathways. A *cmr* operon

was identified with a specific promoter in a thermophilic archaean, *Caldivirga maquilensis*. Overall, we concluded that knowledge-based genomic survey and phylogeny-based functional assignment have suggested for designing a reliable genetic regulatory circuit naturally from CRISPR-CAS systems, acquired defense pathways, to thermophiles in future synthetic biology.

Keywords RNAi · CRISPR · CAS · Thermophiles · Superoperon · Defense mechanism · Genetic circuit · Genome invaders

Introduction

Genomes are potential targets of invasion by molecular parasites such as viruses and transposable elements and organisms. RNA-directed defense mechanisms in the genomes of thermophilic bacteria and archaea have evolved like a prokaryotic genome to cope with the constant threat of genome invaders (Chellapandi and Ranjani 2011; Makarova et al. 2012; Weinberger et al. 2012; Vestergaard et al. 2014). The existence of this pathway in these genomes has been hypothesized to parallel the eukaryotic RNAi pathway known as Clustered regularly interspaced short palindromic repeats-CRISPR-associated sequences (CRISPR-CAS) system or prokaryotic RNAi (pRNAi) (Lillestol et al. 2006; Makarova et al. 2006; 2012; Tomoyasu et al. 2008). The common structural characteristics of CRISPR loci are: (1) multiple short direct repeats (2) non-repetitive spacer sequences between the repeats of similar size (3) a common leader sequence of a few hundred base pairs harboring multiple CRISPR loci and (4) accompanied by *cas* gene families. The principal products of the CRISPR loci are

Electronic supplementary material The online version of this article (doi:10.1007/s11693-015-9176-8) contains supplementary material, which is available to authorized users.

✉ P. Chellapandi
pchellapandi@gmail.com

¹ Molecular Systems Engineering Lab, Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli, Tamil Nadu 620024, India

comprised primarily of invader-targeting sequences that are likely the guides for effector complexes of pRNAi system (Hale et al. 2008; Quax et al. 2013; Cooper and Overstreet 2014). CRISPR-CAS systems confer viral resistance in these genomes remain hypothetical and undefined.

Clustered regularly interspaced short palindromic repeats (CRISPRs) have played an important role in the chromosomal rearrangement, DNA mobilization, repair and regulation and also replicon partitioning (Haft et al. 2005; DeBoy et al. 2006; Makarova et al. 2006; 2012; Cooper and Overstreet 2014). Among 25 gene families in the CAS system, eight gene families (*cas1*, *cas2*, *cas4*, *cas6*, *cas7*) have reported for the nuclease activity, nine families (*cas5*) for Repeat-Associated Mysterious Proteins (RAMP) and two families (*cas3*) for DNA/RNA polymerase/helicase activity (Makarova et al. 2006; Quax et al. 2013; Vestergaard et al. 2014). However, the functions of several other *cas* gene families remain obscure.

Approximately 40 % CRISPR loci are present in prokaryotic genomes sequenced to date (Sorek et al. 2008). Thermophiles have more tendencies for existing CRISPR loci than mesophiles (Nelson et al. 1999). Non-repetitive spacer sequences have suggested as a potential target site for the genome invaders in many archaeal genomes (Brügger et al. 2002). CRISPR loci have shown a distinct genetic and molecular diversity in the strains of a thermophilic archaean, *Methanocaldococcus jannaschii* (Bult et al. 1996) and of thermophilic bacterium, *Streptococcus thermophilus* (Horvath et al. 2008). Some of the CRISPR sequences in *Thermoanaerobacter tengcongensis* and *Thermus thermophilus* HB8 chromosomes have shared with their existing megaplasmids (Jansen et al. 2002; Godde and Bickerton 2006). Spacer sequences have also shown a variation in five strains of *Thermotoga neapolitana* by geographic locale (DeBoy et al. 2006). Several studies have been done on the existence of archaeal CRISPR type III systems mediated targeting of mRNA degradation (Zebec et al. 2014) and dual targeting of RNA and DNA interference (Peng et al. 2015). Since, studying importance of genetic diversity of CRISPR loci and geographic distribution of spacers may reveal typical CRISPR-CAS molecular machinery in thermophiles.

Although CRISPR-CAS related information available in the CRISPR database, a detailed knowledge and annotation would be required to understanding its existence and hypothetical molecular mechanism in recently sequenced thermophiles (Vestergaard et al. 2014). In this perspective, the present study was intended to analyze their molecular diversity of structural components in the CRISPR-CAS systems, and to design and construction of their acquired defense pathways against genome invaders by using knowledge-based genome-wide survey and phylogeny-based functional annotation.

Materials and methods

Annotation of CRISPR components at genome-scale

Complete genome sequences of extremophilic bacteria and archaea were retrieved from NCBI database. The number of TRs and their positions in these genomes were predicted with Tandem Repeats Finder (Benson 1999). Alignment parameter was set as 2 match, 7 mismatch and 7 indels. The detailed information of CRISPRs was collected from CRISPR database (Grissa et al. 2007a) for the genomes of extremophiles. Structural components including spacers, direct repeats (DRs) and CAS in the CRISPR loci were predicted by CRISPR Finder (Grissa et al. 2007b) for those genomes whose information not availed in the CRISPR database. Appropriate location of *cas* genes was explored and identified from their genome maps in the KEGG database.

Phylogeny-based functional annotation of CAS proteins

The function of identified *cas* genes were initially searched from CAS Genes database. The sequence similarity hits were searched for hypothetical proteins and proteins with unknown function against NCBI non-redundant protein database using NCBI-BLASTp similarity searching tool (Altschul et al. 1997). The selected sequences were aligned together and generated multiple sequence alignment with complete deletion of gaps using ClustalX 2.0 software (Thompson et al. 1997). The aligned sequences were iterated at each alignment step and manually inspected to delete the low scoring sequences. Neighbor joining algorithm was used to search homogeneous patterns among all lineages. Phylogenetic trees were constructed by MEGA 4.0 software with 1000 bootstraps replicates (Tamura et al. 2007). Finally, the function of these proteins was annotated from analogous clusters formed in the phylogenetic trees.

Design and assembly of CRISPR-CAS systems

The known and functionally identified genes were listed and compiled in the proposed genetic circuits according to the genomic loci in the genome maps of selected organisms. Genes and promoters located in near to *cas* gene cluster of each genome on forward and reverse directions were detected and assembled together in each gene circuit. After complete genomic and literature surveys, typical organization and ordering of components in CRISPR-CAS systems was finally assembled to these genomes in the genetic regulatory circuits.

Results

Genomic distribution of TRs and CRISPRs in extremophiles

A complete details of predicted CRISPRs information of extremophiles is available in Supplementary. The existence of TRs in thermophilic firmicutes are more than thermophilic actinobacteria, which is rather than thermophilic archaea. However, actinobacteria have more CRISPRs than crenarchaeota and hyperthermophilic bacteria. The presence of CRISPRs in psychrophilic archaea is more than other psychrophiles, which are distributed equally throughout the genomes. Acidophilic crenarchaeota, δ -proteobacteria, γ -proteobacteria and firmicutes have less than sixty TRs with non-uniform distributions, but crenarchaeota and proteobacteria have equally distributed CRISPRs in their genomes. The existence of TRs more in halophiles than alkaliphiles.

Annotation of CRISPR systems

Shewanella frigidimarina and *Acidovorax avenae* subsp. citrulli have more TRs, but no CRISPRs in their genomes (Table 1). *Hydrogenobaculum* sp., *Kosmotoga olearia* and *Nitrosopumilus maritimus* have considerable number of TRs with low number of CRISPRs and *Caldivirga maquilingensis* consists of more CRISPRs in its genome. *K. olearia* genome has shown a prevalent CRISPR system comprising of 84 spacers and then *Thermotoga lettingae* has a typical system with 53 spacers (Table 2). Other genomes have shown only questionable CRISPR systems with 2 spacers. The DRs in CRISPR loci are ranged from 23 to 40 bp in the most of the genomes. Equal length of

DRs and CRISPRs are noted in many of these genomes, which are distributed uniformly. Genomic survey of this study found that among 12 extremophiles, typical CRISPR-CAS systems are established in three thermophilic archaea (*C. maquilingensis*, *Thermococcus gammatolerans*, *Thermococcus sibiricus*), and two thermophilic bacteria (*Petrotoga mobilis*, *T. lettingae*).

Identification of CAS family genes

Endoribonucleases, CRISPR-associated protein, HD-like nuclease, CRISPR-associated autoregulator, RAMP associated protein and CRISPR-associated helicase are found to be common in the CAS system of thermophilic archaea (Table 3 and Supplementary). RecB family exonuclease is existed only in *T. gammatolerans*. Nine *cas*-related and *cmr* genes are identified in *C. maquilingensis*. A cluster of *cas* genes is identified in *T. gammatolerans* and *T. sibiricus*. The function of *cas5* gene is still unknown and other genes encoded nuclease-specific proteins. A CAS system consists of *cas1*–*cas6* cluster except *cas2* in the genome of *P. mobilis* as similar to *T. gammatolerans*. This *cas* gene cluster encodes exonuclease, endoribonucleases, helicase and nuclease. The genome of *T. lettingae* has a *cas* gene family (*cas1*–*cas6*), gene clusters of *cmr* and *csh*, and a *csx11* gene. The function of *csh1*, *csh2* and *csx11* genes is not known. Among eight genes identified, DEATH box helicase (e-value $2.4e^{-05}$), DUF364 (e-value $3.5e^{-28}$) and RAMP superfamily (e-value $4.1e^{-05}$) are functionally annotated significantly, and other genes have shown low functional reliability due to less significant alignment score and e-value (Table 4).

Table 1 Genomic distribution of TRs and CRISPRs in the recently sequenced extremophiles

Genome	Strain	Number		
		TRs	CRISPRs	
		Predicted	Confirmed	Questionable
<i>Caldivirga maquilingensis</i>	IC-167	61	7	2
<i>Deinococcus deserti</i>	VCD115	73	0	1
<i>Hydrogenobaculum</i> sp.	Y04AAS1	113	1	5
<i>Kosmotoga olearia</i>	TBF 19.5.1	112	4	2
<i>Nitrosopumilus maritimus</i>	SCM1	169	1	1
<i>Petrotoga mobilis</i>	SJ95	93	2	3
<i>Thermococcus gammatolerans</i>	EJ3	31	3	0
<i>Thermococcus sibiricus</i>	MM 739	25	1	0
<i>Thermotoga lettingae</i>	TMO	26	2	0
<i>Shewanella frigidimarina</i>	NCIMB 400	312	0	2
<i>Acidovorax avenae</i> subsp. citrulli	AAC00-1	270	0	2
<i>Halobacterium</i> sp.	NRC-1	111	0	3

Table 2 Genomic organization of CRISPR system in the recently sequenced extremophiles

Genome	CRISPR position (nts.)		CRISPR length (nts.)	No. of spacers	DR length (nts.)
	Start	End			
<i>Caldivirga maquilingensis</i>	232508	232870	362	5	25
	288199	288427	228	3	27
	1258730	1259356	626	9	25
	1583088	1583382	294	4	25
	1588509	1589276	767	11	25
	1611612	1612046	434	6	25
	1700150	1700375	225	3	26
<i>Deinococcus deserti</i>	2805977	2806083	106	1	29
<i>Hydrogenobaculum</i> sp.	1037710	1039665	1955	29	30
<i>Kosmotoga olearia</i>	439134	439435	301	4	30
	439675	442253	2578	38	30
	530300	530724	424	6	30
	719736	725263	5527	84	30
<i>Nitrosopumilus maritimus</i>	406483	407331	848	15	24
<i>Petrotoga mobilis</i>	2069951	2070245	294	4	30
	2082391	2085301	2910	43	30
<i>Thermococcus gammatolerans</i>	208373	208937	564	8	30
	1221278	1222718	1440	21	30
	1857022	1857814	792	11	30
<i>Thermococcus sibiricus</i>	1328782	1330412	1630	24	30
<i>Thermotoga lettingae</i>	205528	208415	2887	43	30
	1131044	1134585	3541	53	30
<i>Shewanella frigidimarina</i>	3390806	3390904	98	1	25
	3446947	3447045	98	1	37
<i>Acidovorax avenae</i> subsp. citrulli	889272	889378	106	1	24
	3970633	3970757	124	1	35
<i>Halobacterium</i> sp.	34457	34569	112	1	40
	728064	728221	157	2	32
	1509051	1509145	94	1	23

Phylogeny-based functional annotation of CAS family proteins

In phylogenetic tree (Fig. 1), two major clusters are formed in which the first cluster comprises of cas2 and devR while the second cluster has cas, cmr4, cmr6 and HD domain. As devR is phylogenetically related with cas2, its function may be acquired from cas2 through evolutionary process. The function of cmr6 is distinct between cas5 and cas6. Dead box helicase shares its molecular function from cas1 and cas4. Nevertheless, HD domain phylogenetically resembles to cas1, cas3, cas4, cmr1 and cmr4.

The sequences of csh2 and csh1 are formed a separate cluster in phylogenetic tree wherein csh2 phylogenetically corresponds to cas2 while csh1 resembles to cas4 (Fig. 2). The csx1 belongs to a cluster consisting of cas1, cmr1 and csh2. The cas2, cas3, cas5 and cas6 are formed a cluster

together on which cas5 and cas2 closely related to each other whereas cas3 and cas6 are highly homologous. Since cas6 is evolutionarily related to the predicted proteins with a significant bootstrapping (879–1000), the function of it seems to be reliable to cas6. Accordingly, csh1 and csh2 are functionally equivalent to cas4 and cas2, respectively. Hence, phylogenetic analyses of this study clearly revealed the function of identified key proteins which functions still not known in these thermophiles.

Design and assembly of CRISPR-CAS systems

Generally, cas gene family of these genomes is highly organized in the order of cas5–cas3–cas1–cas2 (Fig. 3). Cas gene cluster seems to appear (in downstream) after a few hundred of basepairs of CRISPR locus. Gene subtypes including cmr4, cmr5, cmr1 and cmr6 are clustered

Table 3 Identification CAS protein family in the genomes of thermophiles using knowledge-based genomic survey

Genome	Accession	Gene	Functional description
Thermophilic archaea			
<i>Caldivirga maquilensis</i>	157920907	DxTHG motif	CRISPR-associated protein
	159042071	APE2256	CRISPR-associated protein
	159042087	<i>Cas3</i>	CRISPR-associated helicase
	159042078	<i>Cmr5</i>	RAMP, cas-like function
	159042088	HD domain	HD like nuclease
	159042084	<i>DevR</i>	CRISPR-associated autoregulator
	159042079	<i>Cmr4</i>	RAMP, cas-like function
	159042077	<i>Cmr1</i>	RAMP, cas-like function
	159042076	<i>Cmr6</i>	RAMP, cas-like function
	239910904	<i>Cas5</i>	—
<i>Thermococcus gammatolerans</i>	239910903	<i>DevR</i>	CRISPR-associated autoregulator
	239910907	<i>Cas4</i>	RecB family exonucleases
	239910905	<i>Cas3</i>	CRISPR-associated helicase
	239911018	<i>Cas6</i>	Endoribonucleases
	239910906	HD domain	HD like nuclease
	239910910	<i>Cas2</i>	Endoribonucleases
	242399469	<i>Cas2</i>	Endoribonucleases
<i>Thermococcus sibiricus</i>	242399467	<i>Cas1</i>	Nuclease
Thermophilic bacteria			
<i>Petrotoga mobilis</i>	160903310	<i>Csh2</i>	—
	160903313	<i>Cas6</i>	Endoribonucleases
	160903309	<i>Cas5</i> (Hmari subtype)	—
	160903308	<i>Cas3</i>	CRISPR-associated helicase
	160903306	<i>Cas1</i>	Nuclease
	160360957	<i>Cas6</i>	Endoribonucleases
	160360951	<i>Cas4</i>	RecB family exonucleases
<i>Thermotoga lettingae</i>	157363073	<i>Csh1</i>	—
	157313677	<i>Cas1</i> (TM1802)	Nuclease
	157364064	<i>Csx11</i>	—
	157363072	<i>Csh2</i>	—
	157363070	<i>Cas3</i>	CRISPR-associated helicase
	157363071	<i>Cas5</i> (Hmari subtype)	—
	157364061	<i>Cmr1</i>	RAMP, cas-like function
	157364058	<i>Cmr4</i>	RAMP, cas-like function
	157363077	<i>Cas6</i>	Endoribonucleases
	157363076	<i>Cas4</i>	RecB family exonucleases
	157363075	<i>Cas1</i>	Nuclease
	157363074	<i>Cas2</i>	Endoribonucleases
	157313681	<i>Cas6</i>	Endoribonucleases

together with APE2245 in *C. maquilensis* and transcribed in forward direction by an overlapping promoter. It appears like a superoperon after that a leader and DRs located in this genome. No specific gene ordering is detected for *cas* gene family and its subtypes in *T. sibiricus*. It has consisted of only *cas1*, *dbh* and *recB* functional

genes and other genes assigned as DUF (Domain with Unknown Function). It indicates the event of horizontal gene transfer that may attributes to acquire such genes from distantly related genomes. *T. gammatolerans* has a cluster of *cas* gene family in the order of *cas6*–*cas5*–*cas3*–*cas4*–*cas1*. The genes encoding for a hypothetical protein and

Table 4 Functional prediction of CAS protein family in the genomes of thermophiles using conserved domain-based searching

Genome	Accession	Gene	Predicted function	Score	e-value
Thermophilic archaea					
<i>Caldivirga maquilensis</i>	159042085	Cmaq_1523	Cas5a	22.3	0.00021
<i>Thermococcus gammatolerans</i>	240103348	TGAM_1291	DUF466	5.1	0.83
<i>Thermococcus sibiricus</i>	242399464	TSIB_1489	DUF522 (DevR)	-77.4	0.28
	242399465	TSIB_1490	DEAH box helicase	3.5	2.4e ⁻⁰⁵
	242399468	TSIB_1493	DUF83 (DevR)	-81.5	0.065
Thermophilic bacteria					
<i>Thermotoga lettingae</i>	157364062	Tlet_1203	RAMP superfamily	97.7	4.1e ⁻²⁶
	157364063	Tlet_1204	Plasmid regulatory protein	14.0	0.00055
<i>Petrotoga mobilis</i>	160903314	Pmob_1885	DUF364	104.5	3.5e ⁻²⁸

Fig. 1 A phylogenetic tree was constructed for functional annotation of the proteins with unknown function using known functional CAS protein subfamily of thermophilic archaea. The *numbers* at the nodes indicate the levels of bootstrap support based on data for 1000 replicates. The *scale bar* indicates 0.05 substitutions per nucleotide position

devR are located between *cas6* and *cas5* and HD endonuclease gene is located between *cas3* and *cas4*.

An important *cas* gene cluster is detected in *P. mobilis* and ordered as *cas6-cas5-cas3-cas4-cas1-cas2*. *Csh2* gene and two genes encoding hypothetical proteins are closely located between *cas6* and *cas5* and their transcriptional organization starts from DUF. Genes including *cas5*, *cas3*, *cas4*, *cas1* and *cas2* are continuously ordered that appears like a superoperon. As similar to *P. mobilis*, a

Fig. 2 A phylogenetic tree was constructed for functional annotation of the proteins with unknown function using known functional CAS protein subfamily of thermophilic bacteria. The *numbers* at the nodes indicate the levels of bootstrap support based on data for 1000 replicates. The *scale bar* indicates 0.05 substitutions per nucleotide position

significant transcriptional unit is found in *T. lettingae*, but its organization and order of *cas* genes are varied. *csh1* and *csh2* genes are located between *cas2* and *cas5*. *Cas* genes are organized in two orders; the first one is *cas6*, *cas4*, *cas1* and *cas2* and second one is *cas5* and *cas3*. It assumes that expression of *cas* gene family of these genomes may be processed by extremophilic short interface RNA (xsiRNA) for forming a xRISC (extremophilic) complex to degrade the target RNAs.

Fig. 3 The proposed molecular machinery for CAS functioning and CRISPR formation in the genomes of selected thermophiles. Section A is representing a genomic organization of CAS gene clusters and associated CRISPR loci whereas section B is showing the

proposed molecular mechanism of xsiRNA processing to cleave the target RNAs. xsiRNA: extremophilic short interfering RNA; xRISC: extremophilic RNA induced silencing complexes

Discussion

Generally, an exchange of protein modules would require compatible repeat sequences and structures (Lange et al. 2013). Repetitive sequences are common in the prokaryotic genomes that can determine in their genome complexity and dynamics. The sequence length and their genomic position of TRs are highly variable from strain to strain, and often unique for a single strain (Belkum et al. 1998; Bachmann et al. 2014). Such sequence- and position-specific variations in TRs are detected in many of extremophiles and few of them have equal distribution of TRs throughout their genomes. It suggested the genetic heterogeneity among extremophiles may lead to phenotypic differences due to differential gene transcription or translation. A majority of DRs in CRISPR loci located in different position of the same genome is highly similar or identical as similar to previous work done by Jansen et al (2002) on thermophiles. Some of the long stretches of short TRs in a replicon have reported to be involved in replicon partitioning of *Haloflex mediterranei* and *H.*

volcanii (Mojica et al. 1995; 2005). It suggests the functional integrity of TRs and CRISPRs may contribute a significant role in the molecular process of their genomes as such.

Bacterial CRISPRs are more structured in general than those from archaea, which are the only element to play a vital role in all phases of immunity. CRISPR-CAS systems are actively involved in evolving adaptive immune systems of thermophiles against invading mobilomes (Chellapandi and Ranjani 2011; Chellapandi 2011; Shah et al. 2013). Many viruses of crenarchaeota coexist in stable relationships with their hosts, suggested that CRISPR systems adopt a regulatory role by exhibiting limited levels of interference (Lillestøl et al. 2006). Both spacer acquisition and interference are dependent on two distinct structural motifs, a spacer acquisition motif for acquisition and target interference motif for interference (Shah et al. 2013). Since the CRISPR spacer sequences have shown homology to phage (77 %), plasmids (16 %) and remaining to chromosomes, the existence of the spacer sequences in these genomes may be acquired from

their invading phages and plasmids (Horvath et al. 2008). Also, the prophages account on their genetic variation across closely related strains and contribute in evolution of adaptive immune systems, which was agreed to earlier works (Marraffini and Sontheimer 2009; Bondy-Denomy and Davidson 2014).

Structural parameters or sequence motifs in the repeats are very essential for the enzymes or structural proteins involved in adaptation, crRNA biogenesis and interference (Alkhnbashi et al. 2014). Most sequence families and structure motifs are associated with either bacterial or archaeal CRISPRs. Lange et al. (2013) described that cas subtypes I-A, I-B, I-D, III-A and III-B are more enriched in thermophiles by using CRISPRmap tree. They also revealed the importance of cleavage sites, patterns of RNA structure motifs and sequence conservation, and the link between evolution of CRISPRs and associated cas subtypes. Using CRISPRstrand server, Alkhnbashi et al. (2014) predicted the repeat orientations of multiple mature RNA species (crRNAs) for the identification and characterization of protospacers on genome invaders and regulatory motifs in the leader sequences of CRISPR arrays. Vestergaard et al. (2014) also classified the cas subtypes based on cas-protein occurrences for archaea. Phylogenetic analysis of this study also revealed that many functions of the identified cas proteins could be derived from distantly related genomes, plasmids and phages that was agreed to earlier findings (Nelson et al. 1999; Zivanovic et al. 2002; Boucher et al. 2003; Marraffini and Sontheimer 2009; Chellapandi and Ranjani 2011; Spilman et al. 2013). Euryarchaea are more diverse phylogenetically than crenarchaea because of they are generally heterotrophic in nature and frequent a wide range of natural environments (Vestergaard et al. 2014). It suggested that many of their natural habitats are relatively rich in bacteria, rendering their CRISPR systems more susceptible to exchange between archaea and bacteria.

Some of the genes located in operons of two or three cas genes, consistent with their being co-functional with CRISPR systems (Vestergaard et al. 2014). Superoperons-like organization was detected to cluster the cas and cmr genes in *P. mobilis*, *T. lettingae*, *T. gammatolerans* and *C. maquilingensis*, respectively. These gene clusters may coordinate a primary DNA repair process in their genomes according to earlier investigations (Makarova et al. 2002; 2006; Sasaki et al. 2004). A co-transcriptional expression of cas genes may operate analogously to eukaryotic RNA interference systems. CAS systems are divided into seven or eight subtypes based on operon organization and gene phylogeny (Bult et al. 1996). Some subtypes of these genes are undergone extensive horizontal gene transfer as very similar cas genes found in distantly related genomes, but differed functionally in accordance to the previous studies

(Makarova et al. 2006; Kunin et al. 2007). Cas1, cas3, and cas4 proteins are homologous to a DNA repair protein, a helicase, and recB exonuclease, respectively. The predicted proteins of this study are directly contributed in DNA a repair and recombination mechanism, which was agreed to earlier annotation (Brügger et al. 2002; Sasaki et al. 2004; Godde and Bickerton 2006; Weinberger et al. 2012; Quax et al. 2013).

Archaeal specific genes (*cas1-cas5*) reported to show a degree of specificity across archaeal or bacterial phyla (Bolotin et al. 2005; Haft et al. 2005; Cooper and Overstreet 2014). Therefore, the cas genes of most thermophilic genomes have appeared in the order of *cas5-cas3-cas1-cas2*. It also reports that a superoperon (*cas6-cas4-cas1-cas2-csh1-csh2-cas5-cas3*) is identified and typically established in the genome of *T. lettingae* as similar to earlier findings (Lillestøl et al. 2006). Very often, superoperons are absent in the chromosomes of *T. acidophilum* and *P. abyssi* and of their plasmids, pNOB8 and pKEF9. Since, a complete organization and ordering of such superoperons and gene clusters with suitable promoters may be served as natural biological parts in designing and construction of genetic regulatory network for these genomes.

Conclusions

A genome-wide distribution of the CRISPRs and TRs among thermophilic archaea and bacteria is described in this study. Since the apparent importance of such existence and organization can contribute to the immunity of these genomes against genome invaders (plasmids and phages), the proposed CRISPR-CAS systems are subject of intense interests. Some of the cas gene subtypes are functionally and phylogenetically related to closely related genomes. The presence of superoperons (cas gene cluster) are also identified in these genomes. A cluster of cmr genes is exclusively found in *C. maquilingensis* genome that is transcribed under the control of a separate promoter in forward direction. Available genomic information and bioinformatics resources thus provide a standpoint to reveal its molecular evolutionary hypothesis to thermophiles and to guide the experimental setup for synthetic biology. We conclude that the proposed CRISPR-CAS systems would be useful for strain typing, engineered defense pathway, silencing of endogenous genes in microbial physiology research either using CRISPR arrays or synthetic genetic circuits. However, gene expression studies of the proposed systems are further required to improve their biological reliability and functional robustness in any microbial cell factory.

Acknowledgments The corresponding author is thankful to the University Grants Commission (32-559/2006), New Delhi, India, for financial assistance to carry out the work.

References

- Alkhnbashi OS, Costa F, Shah SA, Garrett RA, Saunders SJ, Backofen R (2014) CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* 30:i489–i496
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Bachmann NL, Petty NK, Ben Zakour NL, Szubert JM, Savill J, Beatson SA (2014) Genome analysis and CRISPR typing of *Salmonella enterica* serovar Virchow. *BMC Genom* 15:389
- Belkum VA, Scherer S, Alphen VL, Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* 62:275–293
- Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580
- Bolotin A, Quinquis B, Sorokin A, Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151:2551–2561
- Bondy-Denomy J, Davidson AR (2014) To acquire or resist: the complex biological effects of CRISPR-Cas systems. *Trends Microbiol* 22:218–225
- Boucher Y, Douad CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF (2003) Lateral gene transfer and the origins of prokaryotic groups. *Ann Rev Genet* 37:283–328
- Brügger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA (2002) Mobile elements in archaeal genomes. *FEMS Microbiol Lett* 206:131–141
- 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
- Chellapandi P (2011) Molecular evolution of methanogens based on their metabolic facets. *Front Biol* 6:490–503
- Chellapandi P, Ranjani J (2011) Molecular machinery of CRISPR-CAS system—RNA mediated defense pathway in Prokaryotes. *Asian J Biotechnol* 3:329–336
- Cooper EL, Overstreet N (2014) Diversity, evolution, and therapeutic applications of small RNAs in prokaryotic and eukaryotic immune systems. *Phys Life Rev* 11:113–134
- DeBoy RT, Mongodin EF, Emerson JB, Nelson KE (2006) Chromosome evolution in the *Thermotogales*: large-scale inversions and strain diversification of CRISPR sequences. *J Bacteriol* 188:2364–2374
- Godde JS, Bickerton A (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* 62:718–729
- Grissa I, Vergnaud G, Pourcel C (2007a) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 37:W52–W57
- Grissa I, Vergnaud G, Pourcel C (2007b) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinform* 8:172
- 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* 1:474–483
- Hale C, Kleppe K, Terns RM, Terns MP (2008) Prokaryotic silencing (psi) RNAs in *Pyrococcus furiosus*. *RNA* 14:1–8
- Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* 190:1401–1412
- Jansen R, Embden VJDA, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43:1565–1575
- Kunin V, Sorek R, Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* 8:R61
- Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res* 41:8034–8044
- Lillestol RK, Redder P, Garrett RA, Brugger K (2006) A putative viral defense mechanism in archaeal cells. *Archaea* 2:59–72
- Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV (2002) A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* 30:482–496
- Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 1:7
- 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 (2012) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 9:467–477
- Marraffini LA, Sontheimer EJ (2009) Invasive DNA, chopped and in the CRISPR. *Structure* 7:786–788
- Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the archaea *Haloferax mediterranei* and *Haloferax volcanii*. *Mol Microbiol* 17:85–93
- 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
- Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323–329
- Peng W, Feng M, Feng X, Liang YX, She Q (2015) An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res* 43:406–417
- Quax TE, Voet M, Sismeiro O, Dillies MA, Jagla B, Coppée JY, Sezonov G, Forterre P, van der Oost J, Lavigne R, Prangishvili D (2013) Massive activation of archaeal defense genes during viral infection. *J Virol* 87:8419–8428
- Sasaki MS, Takata M, Sonoda E, Tachibana A, Takeda S (2004) Recombination repair pathway in the maintenance of chromosomal integrity against DNA interstrand crosslinks. *Cytogenet Genome Res* 104:28–34

- Shah SA, Erdmann S, Mojica FJ, Garrett RA (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* 10:891–899
- 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
- Spilman M, Cocozaki A, Hale C, Shao Y, Ramia N, Terns R, Terns M, Li H, Stagg S (2013) Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell* 52:146–152
- 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
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 24:4876–4882
- Tomoyasu Y, Miller SC, Tomita S, Schoppmeier M, Grossmann D, Bucher G (2008) Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in *Tribolium*. *Genome Biol* 9:R10
- Vestergaard G, Garrett RA, Shah SA (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol* 11:156–167
- Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV (2012) Viral diversity threshold for adaptive immunity in prokaryotes. *MBio* 3:e00456-12
- Zebec Z, Manica A, Zhang J, White MF, Schleper C (2014) CRISPR-mediated targeted mRNA degradation in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Res* 42:5280–5288
- Zivanovic YP, Lopez P, Philippe H, Forterre P (2002) *Pyrococcus* genome comparison evidences chromosome shuffling driven evolution. *Nucleic Acids Res* 30:1902–1910