ARTICLE

Journal of Cellular Biochemistry 113:2464–2473 (2012)

Journal of Cellular Biochemistry

Comparative Genomic Structures of *Mycobacterium* CRISPR-Cas

Liming He, Xiangyu Fan, and Jianping Xie*

Institute of Modern Biopharmaceuticals, State Key Laboratory Breeding Base of Eco-Environment and Bio-Resource of the Three Gorges Area, School of Life Sciences, Southwest University, Beibei, Chongqing 400715, China

ABSTRACT

Clustered regularly interspaced short palindromic repeats (CRISPR) are inheritable genetic elements of many archaea and bacteria, conferring acquired immunity against invading nucleic acids. CRISPR might be indicative of the bacterial niche adaptation and evolutionary. Mycobacterium is an important genus occupying diverse niches with profound medical and environmental significance. To present a comparative genomic landscape of the *Mycobacterium* CRISPR, the feature of mycobacterium CRISPR structures with sequenced complete genomes were bioinformatically analyzed. The results show that CRISPR structures can be found among 14 mycobacteria, and all loci are chromosomally located. Long CRISPRs present in three species, namely *M. tuberculosis*, *M. bovis*, and *M. avium*. Integrated CRISPR-Cas system can only be found in *M. tuberculosis* and *M. bovis*, with highly conserved repeat sequences, very short leaders, and promoterless. *M. tuberculosis* and *M. bovis* repeat sequences cannot form stable RNA secondary structure, consistent with a Cas6-binding sequence. *M. avium* repeat sequences can form classical stem-loop structure. A three-step model of *M. tuberculosis* CRISPR-Cas system action was put forward based on the composition and function of *cas* genes cluster. *M. tuberculosis* and *M. bovis* CRISPRs might interfere with the invading nucleic acids, but have somehow lost the capacity to incorporate new spacers and co-evolve with corresponding mycobacteriophages. J. Cell. Biochem. 113: 2464–2473, 2012.

KEY WORDS: GENOME; REPEAT; SPACE; ACQUIRED IMMUNITY

Clustered regularly interspaced short palindromic repeats (CRISPRs) are highly diverse inheritable components widespread across many bacteria (~ 40%) and most archaea (~ 90%) [Sorek et al., 2008; Makarova et al., 2011]. Initially reported in *Escherichia coli* in 1987 [Ishino et al., 1987], the name CRISPR was not widely adopted [Jansen et al., 2002] until 2002. Most species contain two or more CRISPR loci. CRISPR loci contain short direct repeats, spacers, and leader. The size of the highly conserved repeat varies between 21 and 47 bp, with an average of 32 bp [Godde and Bickerton, 2006]. Spacers are short sequences with similar size but interposing in two consecutive repeated elements. The leader, consisting of several 100 bp, is a non-coding sequence rich in A/T located at the 5′ end of the first repeat. CRISPR-associated (cas) genes, often adjacent to CRISPR, encode a large protein families. Specific functional domains identified in Cas proteins include nucleases, helicases, polymerases, DNA-, and RNA-binding proteins [Haft et al., 2005]. CRISPR along with Cas proteins is generally known as the CRISPR-Cas system. This can act as acquired immunity system against exogenous nucleic acids (viruses and plasmids) [Barrangou et al., 2007; Garneau et al., 2010], functionally comparable to the eukaryotic RNA interference (RNAi) [Carthew and Sontheimer, 2009].

Mycobacterium is Gram-positive genus bacteria belonging to Actinobacteria. The genus includes a wide variety of organisms of medical, agricultural, and environmental importance, notably the pathogens known to cause serious diseases in humans and animals, such as tuberculosis (*Mycobacterium tuberculosis*) and leprosy (*M. leprae*). With the advent of extensively drug-resistant strains of *M. tuberculosis*, tuberculosis continue to plague global public health. Some mycobacteria appear to be parasites, exemplified by

Grant sponsor: National Megaproject for Key Infectious Disease; Grant number: 2012ZX10003-003; Grant sponsor: Fundamental Research Funds for the Central Universities; Grant numbers: XDJK2009A003, XDJK2011D006; Grant sponsor: National Natural Science Foundation; Grant number: 81071316; Grant sponsor: Excellent PhD Thesis Fellowship of Southwest University; Grant numbers: kb2010017, ky2011003; Grant sponsor: New Century Excellent Talents in Universities (NCET-11-).

*Correspondence to: Jianping Xie, Institute of Modern Biopharmaceuticals, State Key Laboratory Breeding Base of Eco-Environment and Bio-Resource of the Three Gorges Area, School of Life Sciences, Southwest University, Beibei, Chongqing 400715, China. E-mail: georgex@swu.edu.cn, jianpingxie@vip.sina.com

Manuscript Received: 16 February 2012; Manuscript Accepted: 29 February 2012

Accepted manuscript online in Wiley Online Library (wileyonlinelibrary.com): 6 March 2012

DOI 10.1002/jcb.24121 • © 2012 Wiley Periodicals, Inc.

the causative agent of tuberculosis and leprosy. Hence, the diversity of mycobacteria provide an ideal source to dissect the CRISPRs.

M. tuberculosis CRISPR was regarded as the rapidest evolving unit in the genome [Hermans et al., 1991]. The number of repeats and the sequence of specific spacers vary with *M. tuberculosis* strains [Groenen et al., 1993], whereby the basis for spacer-oligotyping or spoligotyping of clinical isolates [Kamerbeek et al., 1997]. This genotyping is also applicable to other bacteria (such as *Salmonella enterica* subsp. *enterica* [Liu et al., 2011] and *Corynebacterium diphtheriae* [Mokrousov et al., 2007]). Is there any defensive role of mycobacterium CRISPR-Cas system? Whether the same proto-spacer can be found among mycobacteriophage genomes? To this end, the occurrence of CRISPR loci in the complete genomes of sequenced Mycobacterium were explored, and a comparative analysis the leader sequences, repeats, spacers, and *cas* genes was performed. The evolving stages of acquired immunity of *M. tuberculosis* CRISPR was proposed based on the *cas* genes content and architecture.

MATERIALS AND METHODS

Complete genome sequences were downloaded from GenBank at the National Center for Biotechnology Information [Benson et al., 2011] (http://www.ncbi.nlm.nih.gov/genomes/). CRISPR information (contain the loci in genomes, repeat, and spacer sequences) were retrieved from the CRISPRdb database [Grissa et al., 2007] (http://crispr.u-psud.fr/crispr). The non-coding sequences at the immediate upstream of the first CRISPR repeat were selected as the putative leader sequences and compared using Vector NTI. Then the leader sequences of long CRISPR whose repeat numbers were greater than 5 were predicted for promoter by online tools BDGP Neural Network Promoter Prediction [Reese, 2001] (http://www.fruitfly.org/seq_tools/promoter.html). Identification of *cas* genes was achieved by NCBI BLAST [Altschul et al., 1990] (http://blast.ncbi.nlm.nih.gov/Blast.cgi). BLAST was also used to search the identical sequences with CRISPR spacers in the GenBank database limited organism to Bacteria (taxid:2) or Viruses (taxid:10239). RNA secondary structure prediction was performed by RNAfold [Schuster et al., 1994; Hofacker, 2003] (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

CRISPR loci was allegedly fallen outside the coding domain [Sorek et al., 2008]. However, each CRISPR information within all-inclusive CRISPRdb database must be manually proofread. Loci located within coding area or repeat size larger than 48 bp were discarded. Additionally, for the convenience of description, the loci excluded were called not real CRISPR locus, and the loci with small repeat numbers (2–5) were named questionable CRISPR locus. Determination of the 5′ end of the long CRISPR was based on the GAAA(C/G) signature at the 3′ terminus of repeats.

RESULTS

OCCURRENCE OF CRISPR LOCI IN MYCOBACTERIUM GENOMES

Twenty-one CRISPR loci (include questionable locus) have been found from 22 mycobacterium strains belonging to 14 species. All loci locate at the chromosome. CRISPRs information were summarized in Table I. The loci with greater repeat numbers (>5) can only be found in three mycobacteria (*M. tuberculosis*, *M. bovis*, and *M. avium*), others are questionable CRISPR loci. Our focus is the long CRISPR.

All five *M. tuberculosis* strains harbor two longer CRISPR loci, CRISPR1 adjacent to *cas* gene cluster, and more distant CRISPR2. The number of *M. tuberculosis* H37Rv repeat is 24 and 18, respectively. The absence of duplicate spacer between the two CRISPRs implicates that they are distinct CRISPRs. The relative position of CRISPR and *cas* gene and their numbers are unique in *M. tuberculosis*. In other bacteria, the common scenario is one CRISPR locus followed by one *cas* genes, or one CRISPR locus flanked by one *cas* gene cluster. In *M. tuberculosis*, two tandem CRISPRs are followed by nine consecutive *cas* genes whose names were cas2, cas1, csm6, csm5, csm4, csm3, csm2, cas10(csm1), and cas6 from 5′ to 3′(Figs. 1 and 3). Two transposase genes belonging to the IS6110 family interpose in the two CRISPR loci. The structures of the CRISPR1 of other four strains are the same as *M. tuberculosis* H37Rv with difference in the number of repeats and spacers. The CRISPR2 of *M. tuberculosis* H37Ra and *M. tuberculosis* F11 are exactly same.

The CRISPR structure of *M. bovis* is similar to *M. tuberculosis*, with same repeat, leader, and *cas* genes. More than half of the spacer sequences are same between them, suggestive of shared phages between *M. tuberculosis* and *M. bovis*. The number of repeat of two CRISPR in *M. bovis* is 25 and 17, respectively, *M. bovis* BCG has 30 and 19, together with two questionable loci. *M. bovis* BCG contains all the spacers of *M. bovis*, arranged dispersedly. *M. bovis* BCG harbors some unique CRISPRs and several identical to that of *M. tuberculosis*. This seems counterintuitive, since the habitat of *M. bovis* BCG might have fewer invader than *M.bovis*. There is a pseudogene between cas1 and csm6 of *M. bovis* BCG *cas* gene cluster, wherein *M. bovis* genome is a non-coding sequence. The pseudogenes and non-coding sequences share a high homology of 84.9%. This pseudogene sequence is a duplicate of the 5′ end of *M. tuberculosis* csm6 (Rv2818c), but the identity with a middle sequence of adjacent csm6 (BCG_2837c) is only 48%.

One long CRISPR and one questionable locus exist in two *M. avium* genomes. The long CRISPR has unique repeat and spacer sequences and no discernable flanking *cas* gene cluster. The number of repeat is 13, we designate it as CRISPR3.

LEADER SEQUENCES

Typical leader is a AT-rich sequence immediate upstream of the first CRISPR repeat and lacks an open reading frame. New repeat-spacer unit tends to interpose between leader and the earlier unit [Barrangou et al., 2007], the leader was the preferable recognition sequence for the insertion of new spacers. Furthermore, the leader can also function as the promoter and transcription factor-binding site of the transcribed CRISPR in some strains, such as *E. coli* K12 [Westra et al., 2010] and *Pyrococcus abyssi* [Phok et al., 2011]. The leader sequences of *M. tuberculosis* and *M. bovis* CRISPR1 and CRISPR2 are identical. However, no conserved consensus sequences can be found from all CRISPR leaders. This was consistent with observation that leaders vary with species [Sorek et al., 2008]. The

TABLE I. The CRISPR Loci in *Mycobacterium* Genomes

Name	Number of CRISPR loci	Number of cas genes	CRISPR locus name	Begin to end position in chromosome	Number of repeats	Repeat size (bp)	Spacer size (bp) (min-max)	cas genes near CRISPR	Representative repeat sequence (5′ → 3′)
Mycobacterium tuberculosis H37Rv	3	13	Mtb-H37Rv 1[a]	692025–692101	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			Mtb-H37Rv 2	3119185–3120468	18	36	35–40	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			Mtb-H37Rv 3	3121862–3123576	24	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Mycobacterium tuberculosis H37Ra	3	13	Mtb-H37Ra 1[a]	693335–693411	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			Mtb-H37Ra 2	3131153–3132436	18	36	35–40	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			Mtb-H37Ra 3	3133830–3135567	24	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Mycobacterium tuberculosis F11	3	13	Mtb-F 1[a]	693335–693411	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			Mtb-F 2	3131140–3132423	18	36	35–40	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			Mtb-F 3	3133859–3135205	19	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Mycobacterium tuberculosis CDC1551	3	13	Mtb-CDC 1[a]	693473–693549	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			Mtb-CDC 2	3113903–3115257	19	36	35–40	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			Mtb-CDC 3	3116651–3117781	16	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Mycobacterium tuberculosis KZN 1435	4	13	Mtb-KZN 1[a]	692277–692353	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			Mtb-KZN 2	1290260–1291827	22	36	25–41	9	GTCGTCAGACCCAAAACCCCGAGAGGGGACGGAAAC
			Mtb-KZN 3	1293263–1294472	17	36	35–40	9	GTCGTCAGACCCAAAACCCCGAGAGGGGACGGAAAC
			Mtb-KZN 4[a]	3148815–3148904	2	29	31	No	TGGGCTTTGCGAATCTCGGCAACAACAAC
Mycobacterium bovis AF2122/97	3	13	Mbo 1[a]	693270–693346	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			Mbo 2	3075735–3076948	17	36	36–43	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			Mbo 3	3078342–3080138	25	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Mycobacterium bovis BCG str. Pasteur 1173P2	5	13	BCG-P 1[a]	722910–722986	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			BCG-P 2[b]	2966330–2966511	3	26	42–62	No	GCGCGCTCGTACTGTTGAGGTCGTCG
			BCG-P 3	3072633–3073989	19	36	35–43	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			BCG-P 4	3075383–3077553	30	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			BCG-P 5[b]	4083791–4083879	2	31	27	No	GCTCGGCGACGATGCGGGCCGGATGACGGCC
Mycobacterium bovis BCG str. Tokyo 172	3	13	BCG-T 1[a]	693271–693347	2	23	31	No	TGAGGTGCGGCGTGAGCGCGGGT
			BCG-T 2	3065406–3066762	19	36	35–43	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
			BCG-T 3	3068156–3070326	30	36	25–41	9	GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Mycobacterium sp. MCS	2	5	MCS 1[b]	3167235–3167317	2	26	31	No	CCACCGACGGCGGCGTCGCAGAGGCG
			MCS 2[a]	4603408–4603494	2	27	33	No	GCATCATCCGCCGTCACCCGACCGGCG
Mycobacterium sp. KMS	4	5	KMS 1[a]	3185123–3185205	2	26	31	No	CCACCGACGGCGGCGTCGCAGAGGCG
			KMS 2[b]	4088587–4088690	2	24	56	No	GCGCTGAGGGCCTGGGCAAGTCTC
			KMS 3[b]	4533157–4533261	2	25	55	No	GAGACTTGGCCCAGGCCCTCAGCGC
			KMS 4[a]	4642613–4642699	2	27	33	No	GCATCATCCGCCGTCACCCGACCGGCG
Mycobacterium sp. JLS	3	5	JLS 1[a]	2523548–2523632	2	25	35	No	CCGTCTGACGACGATGCGGGGCGAG
			JLS 2[a]	3147134–3147216	2	26	31	No	CCACCGACGGCGGGGTCGCAGAGGCG
			JLS 3[a]	4951810–4951896	2	27	33	No	GCATCATCCGCCGTCACCCGACCGGCG
Mycobacterium avium 104	7	6	Mav 1	279270–280041	13	27	34–37	No	TGCTCCCCGCGCAAGCGGGGATGAACC
			Mav 2[a]	2322571–2322653	2	26	31	No	AGGAGCCGGGCGATTCAGCTTGGGCC
			Mav 3[a]	2505613–2505719	2	41	25	No	CGTTCGAGTGCCGCTTCGGCGTACGGCGCGCCGCGTTCGGC
			Mav 4[a]	2745832–2745930	2	24	51	No	CGGTTGCGGCCCGACCCCCAGCGG
			Mav 5[a]	3515395–3515479	2	27	31	No	CCTACCCCGCTTCGCGGCGCAGCGCGG
			Mav 6[a]	3625433–3625509	2	24	29	No	CAGTCGGCGTCATGGGCGGTCATC
			Mav 7[a]	4410369–4410459	2	33	25	No	AGCCCGGTGACGATGCCCGCCGCAGCGCGGCGG
Mycobacterium avium subsp. paratuberculosis K-10	1	5	Mav-P[b]	3124763–3124843	2	26	29	No	GGCGCAGCGGGTCACCGCCATTAGCG
Mycobacterium gilvum PYR-GCK	1	5	Mgi[b]	4160507–4160585	2	24	31	1	GTAACGAACTGACGAACTGACTCA
Mycobacterium marinum M	5	4	Mma 1[a]	564508–564900	5	30	48–69	No	TGCCGAACAGCAGGCCCCCCGCCCCGCCGG
			Mma 2[a]	959046–959118	2	25	23	No	GCCAAGAAGGCTCCCGCCAAGAAGG
			Mma 3[a]	3306511–3306847	5	28	29–68	No	CGCCGCCATTGCCGCCGTTGCCGATCAG
			Mma 4[a]	5580059–5580477	8	25	20–50	No	TGTTTGGCAACGGCGGAGCCGGCGG
			Mma 5[a]	5580581–5580689	2	25	59	No	TGTTTGGCAACGGCGGAGCCGGCGG
Mycobacterium ulcerans Agy99	1	4	Mul[a]	5045124–5045195	2	24	24	No	GCCAAGAAGGCTCCCGCCAAGAAG

(*Continued*)

TABLE I. (*Continued*)

Name	Number of CRISPR loci	Number of cas genes	CRISPR locus name	Begin to end position in chromosome	Number of repeats	Repeat size (bp)	Spacer size (bp) (min-max)	cas genes near CRISPR	Representative repeat sequence (5′ → 3′)
Mycobacterium vanbaalenii PYR-1	5	5	Mva 1[a]	912184–912267	2	25	34	No	CGGGGGCCCGCACTCCCGTCTTGAC
			Mva 2[a]	959137–959234	2	23	52	No	GGTCTGCTGTGGGGCAGCGGCGG
			Mva 3[b]	2440518–2440615	2	26	46	No	AACGGGGGAAACGCCGGCCTTTTCGG
			Mva 4[a]	2938891–2938985	2	23	49	No	GCCGGACGGCGGCGAGGATTCCG
			Mva 5[a]	5468404–5468509	2	25	56	No	CGCGTGACGACGAAGGCGGCGTGGG
Mycobacterium abscessus	0	6	–	–	–	–	–	–	–
Mycobacterium leprae TN/Br4923	0	1/1	–	–	–	–	–	–	–
Mycobacterium smegmatis str. MC2 155	0	4	–	–	–	–	–	–	–
Mycobacterium sp. Spyr1	0	4	–	–	–	–	–	–	–

[a]Not real CRISPR locus.
[b]Questionable CRISPR structures.

size of the leader of CRISPR1 is 48 bp. The size of the leader of CRISPR2 is 97 bp, with a palindromic sequence "CCCCGAG" separated by 12 bp. Similarly, the leader of CRISPR3, with a size of 90 bp, contains a palindromic sequence "CGCCGCG" separated by 20 bp and two shorter palindromic sequences. No promoter possibilities can be found for CRISPR1 and CRISPR2. Slight possibility of promoter can be found in CRISPR3 (scored 0.19, range: 0–1). The sequences 1,000 bp upstream of the CRISPR were predicted to be promoter. A promoter scored 0.51 located 500 bp upstream of first repeat in CRISPR1 was found. Several promoters were predicted in CRISPR2, the sequence with the highest score of 0.68 located 300 bp upstream.

REPEAT SEQUENCES

Repeats can be grouped into 33 clusters based upon sequence similarity [Kunin et al., 2007], 12 of which included 10 or more members. The repeats from *M. tuberculosis* and *M. bovis* fall into cluster 8. Many repeats contain a conserved 3′ GAAA(C/G)terminus, potential-binding site for one or more of the conserved Cas proteins. GAAA(C/G) motif can be found in mature CRISPR RNA (crRNA), suggestive of possible recognition or binding site of Csm protein complexes. Two repeats can be found in mycobacteria long CRISPR, one is the repeat of CRISPR1 and CRISPR2 (repeat1, Fig. 2A), the other is the repeat of CRISPR3 (repeat2, Fig. 2B). 3′ Terminus of repeat1 is GAAAC, consistent with previous reports; while GAACC in repeat2, similar to the conserved motif. Predicted RNA secondary structures of CRISPR repeats were displayed in Figure 2. Repeat1 was predicted to have an unstable secondary structure by RNAfold, with a large middle loop and two small loops at both ends. After CRISPR transcribed into primary CRISPR RNA (pre-crRNA), about 10 nucleotides at the 5′ end of the repeat RNA was enfolded by two ferredoxin-like domains of Cas6, and the 3′ end located in the enzyme active center [Wang et al., 2011]. Based on the Cas6-binding structure with repeat, highly stable structure cannot be formed by repeat1. Mature crRNAs invariably retain a partial (8-nucleotide) repeat sequence upstream of the spacer sequence [Brouns et al., 2008; Hale et al., 2009], cleavage site might be formed between G28 and A29. Figure 2B represents a classical stem-loop stable structure for repeat2 characterized by the one stem of G:C base pairs, a large and a small loop at both ends. This stem-loop structure facilitates the recognition of repeat for Cas proteins-binding RNA, and the CRISPR traps foreign DNA or RNA in the form of repeat-spacer unit.

IDENTICAL SEQUENCES OF SPACERS

Spacer sequences derived from the prior infection phage or plasmid genomes can confer immunity to second infection of the same phages [Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005]. Therefore, spacer of different CRISPR system is unique, few exceptions, such as two tandem spacer resulting from duplication. Incorporation of a spacer sequence into the CRISPR locus necessitates the duplication of a repeat, which will produce a new spacer-repeat unit. Eighty-three spacer patterns can be found from the 259 spacers of the 21 CRISPR of mycobacteria, with varying length from 25 to 62 bp. The spacer sequences of *M. tuberculosis* and *M. bovis* are same. However, the slight variations of several spacers among five *M. tuberculosis* strains

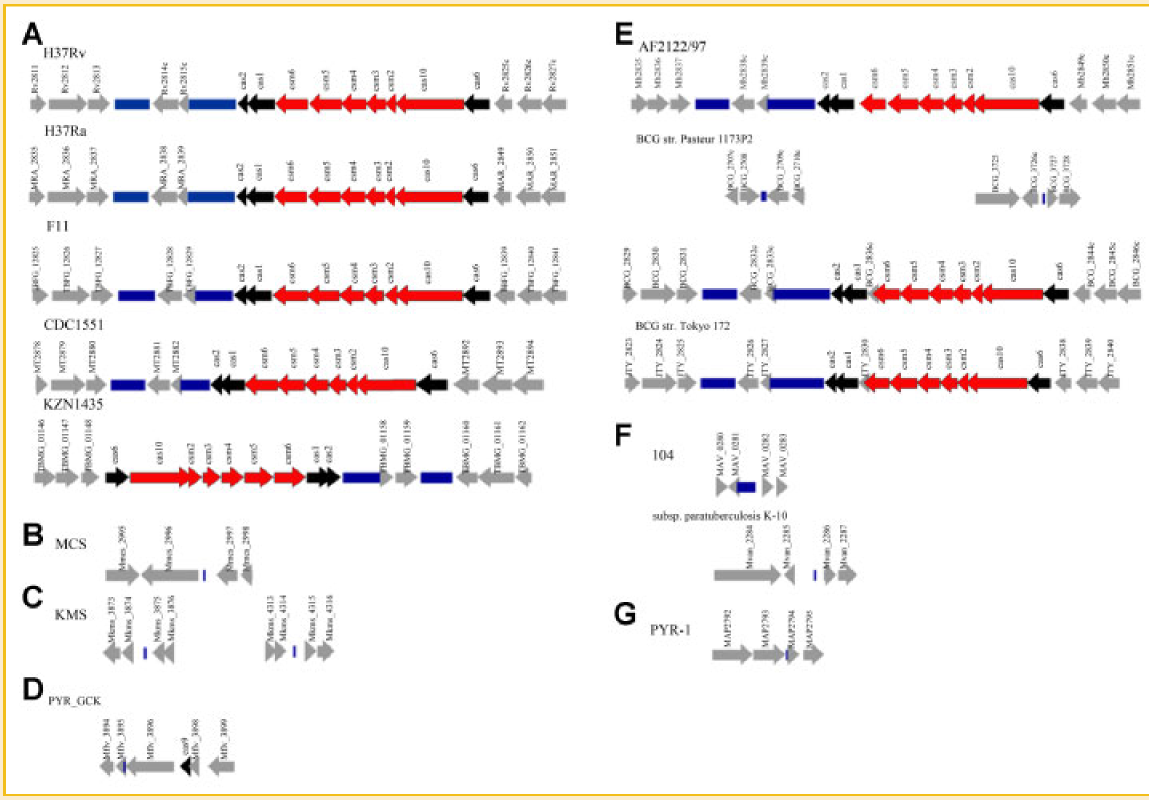

Fig. 1. Graphic representation of Mycobacterium CRISPR loci. The blue rectangles represent the CRISPR loci; the black and red arrows represent the cas genes, the red are the csm genes; the gray arrows represent the genes surrounding the CRISPR loci. The characters above each gene indicate the name or Gene symbol, respectively. A: CRISPR loci in *M. tuberculosis*; H37Rv, *M. tuberculosis* H37Rv; H37Ra, *M. tuberculosis* H37Ra; F11, *M. tuberculosis* F11; CDC1551, *M. tuberculosis* CDC1551; KZN1435, *M. tuberculosis* KZN1435. B: CRISPR loci in *Mycobacterium* sp. MCS. C: CRISPR loci in *Mycobacterium* sp. KMS. D: CRISPR loci in *M. gilvum* PYR-GCK. E: CRISPR loci in *M. bovis*; AF2122/97, *M. bovis* AF2122/97; BCG str. Pasteur 1173P2, *M. bovis* BCG str. Pasteur 1173P2; BCG str. Tokyo 172, *M. bovis* BCG str. Tokyo 172. F: CRISPR loci in *M. avium*; 104, *M. avium* 104; subsp. *paratuberculosis* K-10, *M. avium* subsp. *paratuberculosis* K-10. G: CRISPR loci in *M. vanbaalenii* PYR-1. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/jcb]

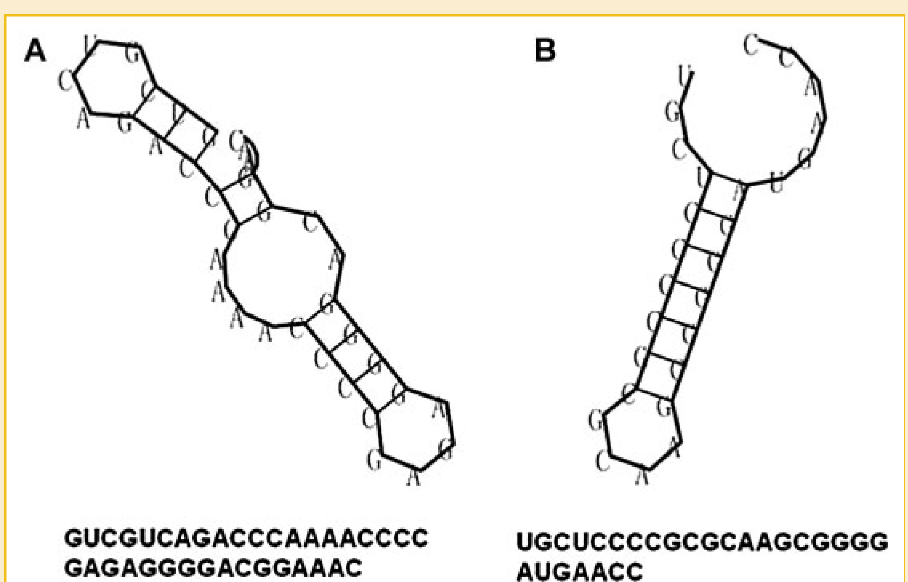

Fig. 2. Predicted RNA secondary structure of the repeat sequences by RNAfold. The repeat RNA sequences are shown at the bottom of the figure. A: Predicted secondary structure of repeat of *M. tuberculosis* and *M. bovis*, unstable. B: Predicted secondary structure of repeat of *M. avium* 104, stable.

implicate diverse phage exposure during each strain life history. Contrary to the rather short evolutionary history of CRISPR [Karginov and Hannon, 2010], the spacers at the leader end of CRISPR loci were conserved among five strains, with hypervariable distal end.

Search for the matches of 83 spacers showing 100% identity over the whole length spacer sequences using NCBI-Blast. We cannot found the counterpart of proto-spacer from known mycobacteriophage. No homology to the genome or insertion sequences of *M. microti*, *M. bovis*, and *M. bovis* BCG can be found too. One explanation might be the intrinsic limitation of current mycobacteriophages discovery methodologies. Nearly all mycobacteriophages were firstly isolated using *M. smegmatis* as host, followed by test their host spectrum using *M. tuberculosis* and other mycobacterium. Therefore, it is logical the host of most mycobacteriophage screened was *M. smegmatis* in which no CRISPR loci can be found [Pope et al., 2011]. The possibilities of as-yet untapped mycobacteriophages or CRISPR resistant phages might exist too [Hatfull, 2008]. The latter immune evasion-like phenomenon had been observed in *S. thermophilus* [Deveau et al., 2008; Karginov and Hannon, 2010].

CAS GENE FAMILIES

Based upon phylogenetic and comparative analyses of the genomic context to these 45 families, three basic types of *cas* gene family were proposed [Haft et al., 2005]: core cas genes, subtype-specific genes, and modular genes. This original classification is simple and widely used. To account for the distant relationships among Cas proteins and the evolutionary relationships among the CRISPR-Cas systems, a "polythetic" nomenclature [Makarova et al., 2011] was adopted in this study.

The cas genes cluster consisted of nine cas genes that were cas2–cas1–csm6–csm5–csm4–csm3–csm2–cas10(csm1)–cas6 (Figs. 1 and 3) from 5′ to 3′ in eight mycobacterial strains which contain complete CRISPR-Cas structure. In addition, csm5, csm4, csm3, and cas6 also belonged to repeat-associated mysterious protein (RAMP) superfamily. Some of the RAMPs had been shown to act as sequence- or structure-specific RNase functional in the processing of pre-crRNA transcripts [Haurwitz et al., 2010]. Csm and Cmr proteins are over-represented in archaea, and the above-mentioned components of cas proteins match to that of *Thermoplasma volcanium*. Csm6 was the least conserved Cas protein among the eight strains with 70.2%, and the percentage identity was 88.6% in five *M. tuberculosis* strains, the others were highly conserved with identity above 99%. A pseudogene between cas1 and csm6 was only found in *M. bovis* BCG. Cas genes exist in all mycobacteria studied, some in clusters, such as *M. tuberculosis* and *M. bovis*, others interspersed. Free cas genes were invariably a function domain of genes. Some free cas genes in the genomes of four mycobacteria lack CRISPR locus (Table I). Cas4 was the widest distributed cas gene, a RecB-like nuclease possibly involved in DNA metabolism or gene expression with three cysteine C-termini cluster [Jansen et al., 2002; Haft et al., 2005; Makarova et al., 2006].

Cas proteins are key player of the immunity against invading genetic elements. The defense act in three stages: adaptation (spacers were integrated into the CRISPR loci), expression (CRISPR was transcribed into pre-crRNA and processed into short mature crRNA), and interference (foreign genome was targeted by crRNA and

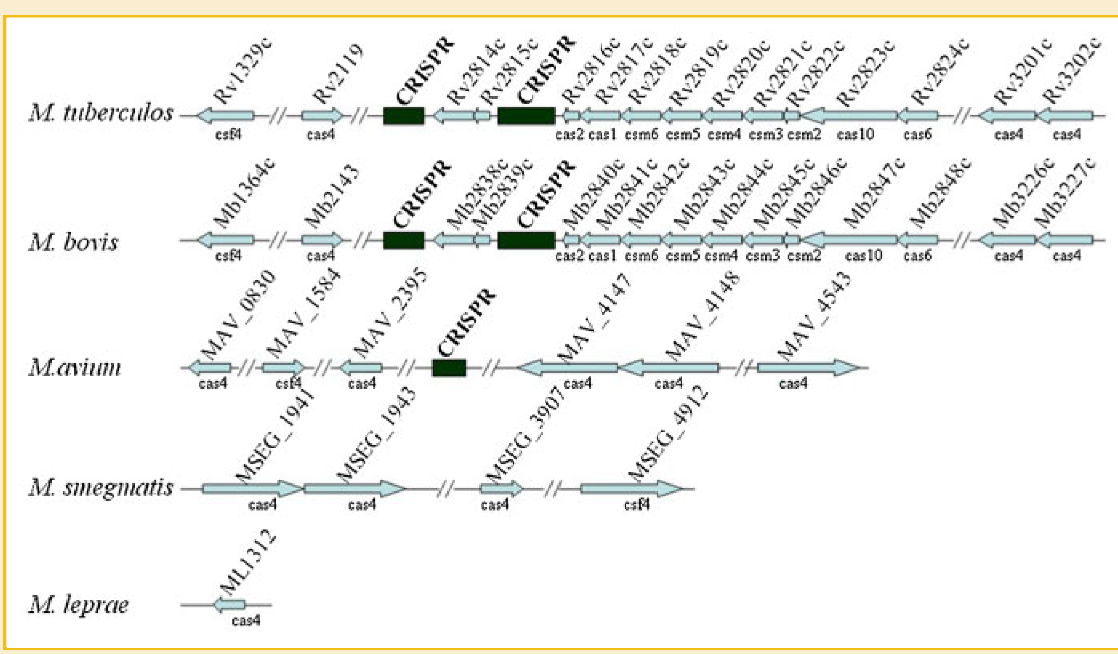

Fig. 3. Graphic representation of CRISPR loci and all cas genes in five mycobacterial genomes. The bluish arrows represent the cas genes, the rectangles represent the CRISPR loci. The words below each cas gene indicate the name of that one, and the corresponding gene symbol is depicted on the top. The niche of five mycobacteria in graph: *M. tuberculosis*, parasitism; *M. bovis*, parasitism; *M. avium*, water and soil; *M. smegmatis*, soil; *M. leprae*, intracellular parasitism. [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/jcb]

cleaved by Cas proteins). Cas1 and Cas2, the most conserved Cas proteins, are presumably crucial for spacer acquisition. Cas proteins, active in other two stages, often act in the form of Cas proteins complexes.

RELATIONSHIP OF MYCOBACTERIUM CRISPR-CAS STRUCTURE WITH NICHES

The widespread distribution of CRISPR-Cas system implicates that it is not niche-specific. This holds true for Mycobacterium. The CRISPR loci and the distribution of cas genes in genomes of several representative mycobacteria were summarized in Figure 3. The presence of cas gene is not associated with host lifestyle and surrounding. CRISPR structures of *M. tuberculosis* and *M. bovis* are highly similar, while quite different among other mycobacteria. This reflects the close evolution relationship between *M. tuberculosis* and *M. bovis*, and consistent with the 16S rRNA-based phylogenetic tree [Brosch et al., 2001].

DISCUSSION

The distribution and the constituents of *Mycobacterium* CRISPR-Cas with completed genome sequences were analyzed. Mycobacteria long CRISPR can be divided into two types. One is the CRISPR1 and CRISPR2 of *M. tuberculosis* and *M. bovis*, directly adjoining *cas* gene cluster. The other is the CRISPR3 of *M. avium*, without apparent *cas* gene cluster. The leaders of CRISPR1 and CRISPR2 were much shorter than the common leader (up to 550 bp), and without promoter function. A predicted promoter can be found in the leader of CRISPR3, but with very low score. The leader of CRISPR2 had similar palindromic sequence with the leader of CRISPR3. The repeat of long CRISPR can also be divided into two types. The representative repeat sequences were shared between *M. tuberculosis* and *M. bovis*, consistent with previous findings [Kunin et al., 2007] that similar repeats can be found in the same- or close-related species. No a stable secondary structure can be found in the RNA of repeat 1, which is agreement with the prediction in *P. furiosus* repeat RNA [Wang et al., 2011]. No mycobacteriophage perfectly matches the spacer can be found.

Typical *cas* gene cluster arrangement, namely cas2–cas1–csm6–csm5–csm4–csm3–csm2–cas10–cas6, was present in *M. tuberculosis* and *M. bovis*. The middle part of Csm6, identical in eight strains, might be the function domain of Csm6. Nucleotide sequence of pseudogene of *M. bovis* BCG was the duplication of the 5′ end of *M. tuberculosis* H37Rv Rv2818c (csm6), and similar to the BCG_2837c middle domain. These observation led to the hypothesis that this pseudogene was likely formed before the N-terminal of BCG_2837c change.

The CRISPR-Cas system is comparable to eukaryotic RNAi [Makarova et al., 2006, 2011]. The disparity is the spacer. Spacer is a DNA fragment derived from exogenous nucleic acid and integrated into the host CRISPR locus [Makarova et al., 2006]. The spacer can recognize the same invader like a "memory stick" via self-matching the complementary sequence in invasive genome [Bolotin et al., 2005; Pourcel et al., 2005]. The matched invader will then be cleaved by Cas protein complexes. The *M. tuberculosis* and *M. bovis* csm genes can be grouped into subtype III-A according to the Makarova "polythetic" classification [Makarova et al., 2011]. *M. tuberculosis* CRISPR-Cas systems mediated immunity against invading mycobacteriophage or plasmid can be divided into three stages (Fig. 4). The first step is adaptation, the proto-spacer in virus genome or plasmid is inserted into the leader side of CRISPR locus. It is widely accepted that Cas1 and Cas2 are crucial Cas proteins in this process [Karginov and Hannon, 2010]. The ubiquitous Cas protein Cas1, a Mn^{2+}- or Mg^{2+}-dependent double-stranded DNA (dsDNA) endonuclease with promiscuous sequence specificity, is presumable a component of the machinery disposing foreign genetic materials [Wiedenheft et al., 2009; Cady and O'Toole, 2011]. The role of Cas2, a sequence-specific endoribonuclease that cleaves uracil-rich single-stranded RNAs (ssRNAs), remains unclear [Beloglazova et al., 2008]. How these two proteins interact or cooperate, and is there any other elements involved, remain an open question. Further investigation of the regulatory factors involving in this process is justified. The second step is expression, the CRISPR is transcribed into pre-crRNA and processed into mature crRNAs by Cas6 and other Cas proteins. Based on the type III systems, the pre-crRNA transcript is cleaved into crRNA units by a single endoribonuclease Cas6 [Makarova et al., 2011]. Cas6 binds to 2–9 nucleotides near the 5′ end of the pre-crRNA repeat sequence and its cleavage site lie in approximately 20 nucleotides on the opposite side of the binding site [Carte et al., 2010; Wang et al., 2011]. The short crRNA processed by Cas6 was delivered to Cas protein complex where the nucleotides at the 3′ end will be further degenerated [Carte et al., 2008; Hale et al., 2008], subsequently the mature crRNA is generated. The final step is mere conjecture. The proto-spacer sequence is targeted by crRNA and destructed by Csm complex. Discrimination of the chromosomal CRISPR locus and the invading DNA fragment during this step is required. Interference which might not occur can take place simply because the base paired to the 5′ repeat fragment of the mature crRNAn [Marraffini and Sontheimer, 2010; Makarova et al., 2011].

The CRISPR-Cas system is regulated by global regulatory factors, such as LeuO, H-NS, and LRP in *E. coli* [Westra et al., 2010], *Salmonella* [Medina-Aparicio et al., 2011], and many other prokaryotes. Homologs can be found in *M. tuberculosis*. *M. tuberculosis* Lsr2 (Rv3597c), a repressor implicated in virulence [Gordon et al., 2010], is a unique H-NS-like protein. *M. tuberculosis* Lrp (Rv3291c) and Rv2779c are regulatory proteins crucial to nutrient limitation and persistence [Betts et al., 2002; Thaw et al., 2006]. Two Lrps homologs, Rv2529 and Rv2324, present in *M. tuberculosis* genome too. But no LeuO protein homolog can be found.

The mechanisms of short-term CRISPR evolution mentioned above means that the ancestral spacers at the distal end of the locus are shared among strains, but the "newer" spacers next to the leader are polymorphic [He and Deem, 2010]. Among the five *M. tuberculosis* strains, the spacers at the leader-proximal end of the CRISPR are common within each strain, but the distal end of the leader is unique. Besides, the CRISPR leader is so short and has no promoter, most probably cannot function as the recognition site for the insertion of new spacers with truncation. Perhaps to assist the

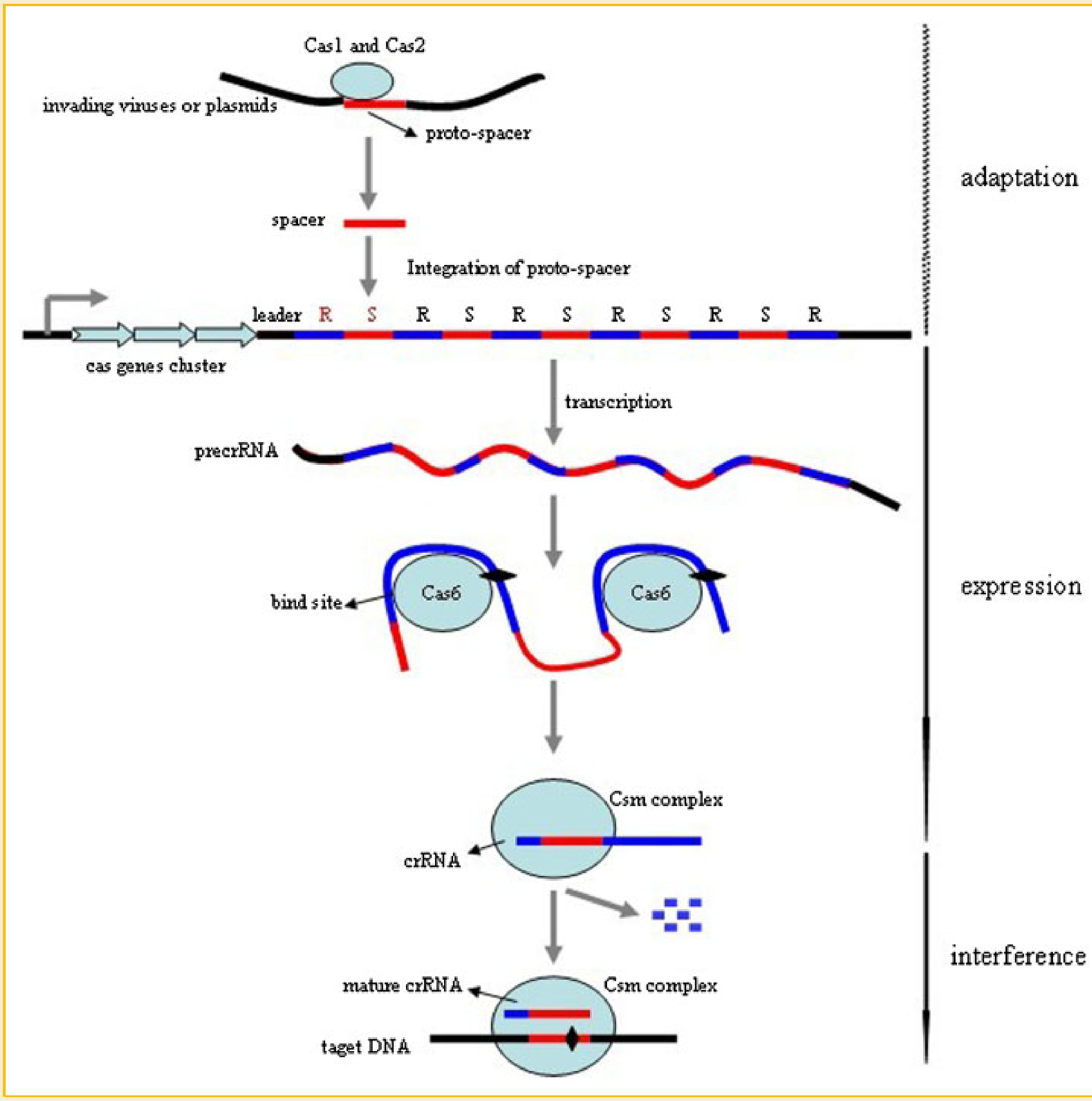

Fig. 4. Overview of the *M. tuberculosis* CRISPR–Cas system action. After a novel spacer derived from viruses or plasmids is actively incorporated into the leader end of the CRISPR locus by Cas1 and Cas2, the CRISPRCas system can recognize and resist the same invader. The CRISPR repeat-spacer array is transcribed into a pre-crRNA that is shorn a set of small RNAs by Cas6 and processed into mature crRNAs degenerated the 3′ end. If the virus or plasmid invade once again, the crRNA will guide the destruction of corresponding invading nucleic acid by Cas complex. Repeats are represented as red fragment, marked "R," spacers as blue fragment, marked "S," and the additional unit is marked red word. Filled diamonds represent the cleavage site. The three stages of CRISPR–Cas action are shown on the right. The adaptation stage is represented by dotted arrows because it is absence in *M. tuberculosis* (see below). [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/jcb]

host limit the expansion of the CRISPR locus, the internal and trailer spacer deletions have been reported. Furthermore, because of the high rate of evolution for phage genomes, the resistance against viruses provided by "older" spacers may be historical, the loss of spacers is necessary to maintain a dynamic level in size. So the polymorphism between the distal end of the cluster contains spacers is likely to due to the difference with the deletion. In summary, this led to the hypothesis that the *M. tuberculosis* CRISPRs are inactive remnants, they cannot appear to incorporate new spacers. Simultaneously, they have complete cas genes cluster and repeat-spacer structure, thus we consider this CRISPRs can accomplish the expression and interference stages and interfere with the corresponding invading nucleic acid.

After a series of comparative analyses among the evolutionary clusters of repeats, cas1 and 16S rRNA genes sequences from 100 different bacteria, Chakraborty et al. [2010] revealed that repeat and cas1 genes are coevolving and have analogous ancestral origin. Therefore, the cas gene cluster of *M. avium*, which only have CRISPR locus now, would have lost at some time of evolutionary history. Moreover, *M. tuberculosis* repeat and cas1 share identical clades with 10 other bacteria whose genus even orders are different with *M. tuberculosis*, such as *Bifidobacterium adolescentis*, *Methylococ-*

cus capsulatus, and *Leptothrix cholodnii*, respectively [Chakraborty et al., 2010]. This supports the possibility of horizontal transfer event of CRISPR locus.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403–410.

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(5819):1709–1712.

Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF. 2008. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. J Biol Chem 283(29):20361–20371.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2011. GenBank. Nucleic Acids Res 39(Database issue):D32–D37.

Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K. 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. Mol Microbiol 43(3):717–731.

Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151(Pt 8):2551–2561.

Brosch R, Pym AS, Gordon SV, Cole ST. 2001. The evolution of mycobacterial pathogenicity: Clues from comparative genomics. Trends Microbiol 9(9): 452–458.

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(5891):960–964.

Cady KC, O'Toole GA. 2011. Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. J Bacteriol 193(14): 3433–3445.

Carte J, Wang R, Li H, Terns RM, Terns MP. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 22(24):3489–3496.

Carte J, Pfister NT, Compton MM, Terns RM, Terns MP. 2010. Binding and cleavage of CRISPR RNA by Cas6. RNA 16(11):2181–2188.

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. Cell 136(4):642–655.

Chakraborty S, Snijders AP, Chakravorty R, Ahmed M, Tarek AM, Hossain MA. 2010. Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria. Mol Phylogenet Evol 56(3):878–887.

Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol 190(4): 1390–1400.

Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468(7320):67–71.

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(6):718–729.

Gordon BR, Li Y, Wang L, Sintsova A, van Bakel H, Tian S, Navarre WW, Xia B, Liu J. 2010. Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 107(11):5154–5159.

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.

Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. 1993. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; Application for strain differentiation by a novel typing method. Mol Microbiol 10(5):1057–1065.

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(6):e60.

Hale C, Kleppe K, Terns RM, Terns MP. 2008. Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. RNA 14(12):2572–2579.

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 139(5):945–956.

Hatfull GF. 2008. Bacteriophage genomics. Curr Opin Microbiol 11(5):447–453.

Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. 2010. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 329(5997):1355–1358.

He J, Deem MW. 2010. Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). Phys Rev Lett 105(12):128102.

Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect Immun 59(8):2695–2705.

Hofacker IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Res 31(13):3429–3431.

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. J Bacteriol 169(12):5429–5433.

Jansen R, Embden JD, Gaastra W, Schouls LM. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43(6): 1565–1575.

Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35(4):907–914.

Karginov FV, Hannon GJ. 2010. The CRISPR system: Small RNA-guided defense in bacteria and archaea. Mol Cell 37(1):7–19.

Kunin V, Sorek R, Hugenholtz P. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8(4):R61.

Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG. 2011. Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. Appl Environ Microbiol 77(6):1946–1956.

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. 2011. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9(6):467–477.

Marraffini LA, Sontheimer EJ. 2010. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 463(7280):568–571.

Medina-Aparicio L, Rebollar-Flores JE, Gallego-Hernandez AL, Vazquez A, Olvera L, Gutierrez-Rios RM, Calva E, Hernandez-Lucas I. 2011. The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* Serovar *Typhi*. J Bacteriol 193(10):2396–2407.

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(2):174–182.

Mokrousov I, Limeschenko E, Vyazovaya A, Narvskaya O. 2007. *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. Biotechnol J 2(7):901–906.

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 Genomics 12:312.

Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, Alexander LM, Alfano MB, Alford ST, Amy NE, Anderson MD, Anderson AG, Ang AA, Ares M, Jr., Barber AJ, Barker LP, Barrett JM, Barshop WD, Bauerle CM, Bayles IM, Belfield KL, Best AA, Borjon A, Jr., Bowman CA, Boyer CA, Bradley KW, Bradley VA, Broadway LN, Budwal K, Busby KN, Campbell IW, Campbell AM, Carey A, Caruso SM, Chew RD, Cockburn CL, Cohen LB, Corajod JM, Cresawn SG, Davis KR, Deng L, Denver DR, Dixon BR, Ekram S, Elgin SC, Engelsen AE, English BE, Erb ML, Estrada C, Filliger LZ, Findley AM, Forbes L, Forsyth MH, Fox TM, Fritz MJ, Garcia R, George ZD, Georges AE, Gissendanner CR, Goff S, Goldstein R, Gordon KC, Green RD, Guerra SL, Guiney-Olsen KR, Guiza BG, Haghighat L, Hagopian GV, Harmon CJ, Harmson JS, Hartzog GA, Harvey SE, He S, He KJ, Healy KE, Higinbotham ER, Hildebrandt EN, Ho JH, Hogan GM, Hohenstein VG, Holz NA, Huang VJ, Hufford EL, Hynes PM, Jackson AS, Jansen EC, Jarvik J, Jasinto PG, Jordan TC, Kasza T, Katelyn MA, Kelsey JS, Kerrigan LA, Khaw D, Kim J, Knutter JZ, Ko CC, Larkin GV, Laroche JR, Latif A, Leuba KD, Leuba SI, Lewis LO, Loesser-Casey KE, Long CA, Lopez AJ, Lowery N, Lu TQ, Mac V, Masters IR, McCloud JJ, McDonough MJ, Medenbach AJ, Menon A, Miller R, Morgan BK, Ng PC, Nguyen E, Nguyen KT, Nguyen ET, Nicholson KM, Parnell LA, Peirce CE, Perz AM, Peterson LJ, Pferdehirt RE, Philip SV, Pogliano K, Pogliano J, Polley T, Puopolo EJ, Rabinowitz HS, Resiss MJ, Rhyan CN, Robinson YM, Rodriguez LL, Rose AC, Rubin JD, Ruby JA, Saha MS, Sandoz JW, Savitskaya J, Schipper DJ, Schnitzler CE, Schott AR, Segal JB, Shaffer CD, Sheldon KE, Shepard EM, Shepardson JW, Shroff MK, Simmons JM, Simms EF, Simpson BM, Sinclair KM, Sjoholm RL, Slette IJ, Spaulding BC, Straub CL, Stukey J, Sughrue T, Tang TY, Tatyana LM, Taylor SB, Taylor BJ, Temple LM, Thompson JV, Tokarz MP, Trapani SE, Troum AP, Tsay J, Tubbs AT, Walton JM, Wang DH, Wang H, Warner JR, Weisser EG, Wendler SC, Weston-Hafer KA, Whelan HM, Williamson KE, Willis AN, Wirtshafter HS, Wong TW, Wu P, Yang Y, Yee BC, Zaidins DA, Zhang B, Zuniga MY, Hendrix RW, Hatfull GF. 2011. Expanding the diversity of mycobacteriophages: Insights into genome architecture and evolution. PLoS ONE 6(1):e16329.

Pourcel C, Salvignol G, Vergnaud G. 2005. Elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151(Pt 3): 653–663.

Reese MG. 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput Chem 26(1): 51–56.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: A case study in RNA secondary structures. Proc Biol Sci 255(1344):279–284.

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(3):181–186.

Thaw P, Sedelnikova SE, Muranova T, Wiese S, Ayora S, Alonso JC, Brinkman AB, Akerboom J, van der Oost J, Rafferty JB. 2006. Structural insight into gene transcriptional regulation and effector binding by the Lrp/AsnC family. Nucleic Acids Res 34(5):1439–1449.

Wang R, Preamplume G, Terns MP, Terns RM, Li H. 2011. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: Recognition and cleavage. Structure 19(2):257–264.

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(6):1380–1393.

Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA. 2009. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure 17(6):904–912.