J Mol Evol (2006) 62:718–729
DOI: 10.1007/s00239-005-0223-z

JOURNAL OF MOLECULAR EVOLUTION

The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes

James S. Godde, Amanda Bickerton

Department of Biology, Monmouth College, Monmouth, IL 61462, USA

Received: 21 September 2005 / Accepted: 12 January 2006 [*Reviewing Editor:* Dr. Stuart Newfeld]

Abstract. We have found direct DNA repeats 21–47 bp in length interspersed with nonrepetitive sequences of similar length, or clustered regularly interspaced short palindromic repeats (CRISPRs) in a wide range of diverse prokaryotes, including many Archaeal and Eubacterial species. A number of *cas*, CRISPR-associated genes have also been characterized in many of the same organisms. Phylogenetic analysis of these *cas* genes suggests that the CRISPR loci have been propagated via HGT, horizontal gene transfer. We suggest a mechanism by which this HGT has occurred, namely, that the CRISPR loci can be carried between cells on megaplasmids ≥40 kb in length.

Key words: Repetitive DNA — Mobile genetic elements — Lateral gene transfer

Introduction

Highly repetitive, noncoding DNA is known to make up the majority of many eukaryotic genomes. Prokaryotic genomes, however, are typically seen to contain lower levels of repetitive DNA; most phyla examined have less than 5% of their genome in the form of global DNA repeats (Ussery et al. 2004). These authors hypothesized that prokaryotes have had more replication cycles over evolutionary time to "streamline" their genomes and that the repetitive DNA found in prokaryotic genomes often has some selective advantage or other reason for its existence.

Correspondence to: James S. Godde; *email:* JGodde@monm.edu

DNA repeats have been classified as either direct, inverted, or mirror, depending on their region of symmetry. Short direct repeats are common in genomes containing repetitive DNA; these can be found arrayed in tandem, dispersed throughout a genome, or, more recently, interspersed with nonrepetitive sequences. Clustered regularly interspaced short palindromic repeats (CRISPRs) are a recently described class of repetitive DNA element that is contained exclusively in prokaryotes. CRISPRs contain 21–47 bp of a directly repeated sequence separated by similarly sized chunks of nonrepetitive DNA.

The presence of short regularly spaced repeats in prokaryotes was first described by Mojica and colleagues, whereas the name CRISPRs was later introduced by Jansen and colleagues (Mojica et al. 2000; Jansen et al. 2002a). These authors were able to locate CRISPRs in approximately half of the prokaryotic genomes available at the time, including a wide diversity of organisms belonging to both the Eubacterial and Archaeal domains (Jansen et al. 2002a,b). Jansen and colleagues also described four CRISPR-associated, or *cas*, genes that were typically found in association with the DNA repeats. The function of CRISPRs, as well as their associated genes, remains unknown. The genes do share homology with proteins involved in DNA recombination and repair. *Cas* 1, 3, and 4 are homologous to a DNA repair protein, a helicase, and a RecB exonuclease, respectively. The function of *cas*2 has not been predicted with any certainty.

Clusters of CRISPR-associated genes have previously been characterized in great detail due to their resemblance to a novel DNA repair system in certain

prokaryotes (Makarova et al. 2002). This investigation was performed before it was known that repetitive sequences were located nearby. The authors used phylogenetic evidence to show that HGT was most likely responsible for the movement of this cluster among distantly related genomes, and hypothesized that the gene cassettes disseminated as a single entity. Recently, the intervening sequences of CRISPR loci from a wide range of different prokaryotes have been found to resemble sequence from transmissible genetic elements such as bacteriophage and conjugative transposons (Mojica et al. 2005). This conclusion has been confirmed by studies which focused on the origin of intervening DNA taken from many strains of a single organism, either *Streptococcus thermophilus* or *Yersinia pestis* (Bolotin et al. 2005; Pourcel et al. 2005). All of this has lead to the speculation that CRISPRs and their associated genes represent a form of mobile genetic element which moves via HGT. In this study, we expand upon the prokaryotic genomes found to contain CRISPRs as well as *cas* genes, confirm findings that HGT has acted upon the *cas* genes in question and present evidence that this transfer has likely involved the use of conjugation.

Results

We have searched 370 prokaryotic genomes for the presence of interspersed direct repeats, three quarters of these genomes had been completed and published at the time of our analysis, the remainder of which were in various stages of completion (*supplemental material*, Table A). We initially set the limits of the repeated pattern broader than the previously reported range that CRISPRs fell into, until we were convinced that no species fell outside of this range. The average size of a CRISPR repeat was found to be 32 bp (*supplemental material*, Table B). Organisms at the lower end of this range include *Yersinia pestis* KIM and *Nostoc punctiforme*, while *Bacteroides fragilis* NCTC 9343 has the longest CRISPR repeat found to date (Table 1). Once the range became apparent, the search parameters were narrowed to include this range only in order to speed up our analysis. We have been able to greatly expand the number of species which are known to contain CRISPRs, from the 39 previously reported to the 148 described here (Jansen et al. 2002a; Table 1; *supplemental material*, Table A).

CRISPR containing species of prokaryotes are extremely diverse, belonging to the domain Archaea as well as the majority of the phyla that have been sequenced from the Eubacterial domain. In all, about 40% of the genomes that have been searched were found to contain CRISPR loci. It is not uncommon for there to be more than one CRISPR locus per genome, genera with large numbers of loci include *Anabaena* (sp. 7120), *Chloroflexus*, and *Methanocaldococcus*, which all contain 11 (*supplemental material*, Table A). Overall, about half of the CRISPR-containing organisms studied had more than one CRISPR locus in their genome (*supplemental material*, Table A). The number of interspersed repeats in these loci also varies, with many loci containing as few as 3 repeats (the lower limit set by our program) and the average number of repeats per locus being around 27. However, one species, *Thermoanaerobacter tengcongensis*, contains a locus with 217 repeats, nearly twice the upper limit previously reported (Jansen et al. 2002a). The genomic coordinates for each of the CRISPR loci studied are given in Table 1 and again in a machine-readable format in the supplemental material (Table B).

The CRISPRs described here can be seen to be quite variable in sequence. In fact, with the exception of closely related strains or species, no two CRISPR sequences were found to be alike. Most CRISPRs do, however, have an imperfect palindromic sequence which conforms to the general consensus sequence shown in Figure 1. It is not uncommon for a CRISPR sequence to begin with a G, followed by 3 Ts and either a G or a C. This is roughly palindromic to the most common ending, namely GAAAC. A run of cytosines a quarter of the way into the sequence also corresponds with a run of guanines three quarters of the way in. There is, however, little conservation of sequence in the center of the CRISPR repeat.

Many, but not all, of the CRISPR loci described here contain a number of genes in close association with the DNA repeats. A recent study has examined many CRISPR-associated genes and grouped them into 45 different protein families, providing evidence of the size, complexity, and heterogeneity of many CRISPR loci (Haft et al. 2005). Although the number of CRISPR-associated genes appears to be approximately a dozen in many of the genomes we have examined, we have chosen to concentrate on the four original *cas* genes for this investigation. Homologs of the *cas* genes were identified in many of the species found to have CRISPRs by our DNA pattern searching techniques. In order to be classified as a *cas* homolog in this study, a gene had to meet two criteria: 1) homology with known *cas* genes, and 2) proximity to a CRISPR locus. This allows a greater degree of confidence that, despite the divergence of many of the encoded proteins, they indeed belong to the *cas* gene categories indicated, although we realize that *bone fide cas* genes may have been left out by imposing these criteria. Overall, about two thirds of the 148 CRISPR-containing species revealed identifiable homologs of at least one of the four core *cas* genes, these 103 spe-

Table 1. *Cas* genes and associated CRISPRs in prokaryotic genomes

Organism name	Cas 1	Cas 2	Cas 3	Cas 4	CRISPR (Coordinates)
Acinetobacter calcoaceticus	ACIAD2484		ACIAD2477		GTTCGTCATCGCATAGATGATTTAGAAA (2448115 - 53475)
Actinobacillus pleuropneum	Ava_4176			Ava_4175	CTTCACTGCCGTATAGGCAGCTTAGAAA
Aeropyrum pernix	**ape1240**		**ape1232**	**ape1239**	*GCATATCCCTAAAGGGAATAGAAAG (786657 - 9345)*
Anabaena PCC 7120	alr0381	asr0382			GTTTCCATCCCCTTGCGGGGAAGTGATGGTAAGTAAC (445573 - 7716)
”	alr1468				GTTTCCATCCCCGTGAGGGGTAAAGGAATCAAAAC (1732927 - 3321)
”	alr1568	asr1570	alr1564	alr1567	GTTTCAATCCCTGATAGGGATTTGTGTTAATAGAAAC (1836813 - 7723)
Anabaena variabilis	Yes	Yes		Yes	*GTTTCTATTAACACAAATCCCTATCAGGGATTGAAAG*
Aquifex aeolicis	**aq_369**		**aq_371**	**aq_370**	*GTTTCAACTCCACACGGTACATTAGGAAC (244561 - 721)*
Archaeoglobus fulgidus	**AF2435**	**AF2434**	AF2406	**AF2436**	*GTTGAAATCAGACCAAAATGGGATTGAAAG (148 - 4211)*
”	**AF1878**	**AF1876**	**AF1874**	**AF1877**	*GTAAGAAAGGGAGGCTCCTGAAAATGGAGATTGAAAG (1690931 - 3946)*
Azotobacter vinelandii	Yes	Yes	Yes	Yes	GTTTCAATCCACACGCCCGCATGGGGCGTGAC
Bacillus clausii	ABC3592				ATTTCAATCCACGCACTCACAAAGAGTGCGAC (3720977 - 1732)
Bacillus halodurans	**bh0341**	**bh0342**	**bh0336**	**bh0340**	*GTCGCACTCTACATGAGTGCGTGGATTGAAAT (355361 - 78904)*
Bacteroides fragilis NCTC 9343	BF3951	BF3950	BF3954		GTTGTGATTTGCTTTCAAATTAGTATCTTTGAACCATTGGAAACAGC (4661170 - 3191)
Bacteroides fragilis YCH46	BF2544	BF2543	BF2548	BF2545	ATTTCAATTCCATAAGGTACAATTAATAC (2924393 - 922)
Campylobacter jejuni jejuni	**Cj1522c**	**Cj1521c**	**Cj1523c**		*TTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT (1455126 - 419)*
Campylobacter jejuni RM1221	*CJE1695*	*CJE1694*			TTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT (1594104 - 336)
Chlorobium tepidum	**CT1130**	**CT1129**	**CT1135**	**CT1131**	*GTTTCAATCCACGCGCCCGCGCGGGGCGCGAC (1052108 - 4923)*
”	CT1977	CT1978	CT1971	CT1974	GTCTTCCCCACGCCCGTGGGGGTGTTTC (1877311 - 8380)
Chloroflexus aurantiacus	Yes (2)	Yes	Yes	Yes	CTTTCAACAATTTCGGCTCACGGTAGAGCACTGAAAC, CAGCAGAGCATTGCCCCGCAATGAAGGGGTTTGAAAC
Chromobacterium violaceum	**CV1229**	**CV1230**	**CV1224**	**CV1228**	GTGCTCCCCACGCACGTGGGGATGAACCG (1452717 - 62362)
”	CV1756		CV1755	CV1751	TTTCTAAGCTGCCTATCCGGCAGTGAAC (1900978 - 2557)
Clostridrum tetani	**CTC01148**		**CTC01146**	**CTC01147**	GTATTAGTAGCACCATATTGGAATGTAAAT (1217308 - 9456)
”	**CTC01463**		**CTC01465**	**CTC01464**	ATTTAAATACAACTCTTGTTATTGTTCAAC (1570768 - 92352)
Clostridium thermocellum	Yes (3)	Yes	Yes (2)	Yes	GTTTTTATCGTACCTATGAGGAATTGAAAC, GTTTTTATCGTACCTATGAGGAATTGAAAT GTTTCAATCCTTGTTTTTACTGGAAGTACCTCTTCAAC
Corynebacterium diphteriae	**DIP0037**	**DIP0038**	**DIP0036**		GAAGTCTATCAGGGTTTTTGAGAACTGAACCCCAGT (39014 - 39354)
”	DIP2214	DIP2215	DIP2213	DIP2212	GTCTTCTCCGCACACGCGGAGGTATTTC (2306022 - 7630)
Corynebacterium jeikeium	jk0643		jk0649	jk0644	ND
Desulfitobacterium hafniensie	Yes (2)	Yes (2)	Yes (2)	Yes (2)	GTCACTCCTCGTATGAGGAGTGTGGATTGAAAT, GTTTCAATTCCTCATAGGTAAGCTAACAAC
Desulfotalea psychrophila			DP0186		GCTTCAATGTAGTCACCCCTTTCGAGGTGATTGATAC (198425 - 9019)
Desulfovibrio desulfuricans	Yes	Yes	Yes	Yes	GTGTTCCCCGCACCCGCGGGGATGAACCG
Escherichia coli K12	**b2755**	b2754	**b2761**	b2756	*CGGTTTATCCCCGCTAACGCGGGGAACTC (2875846 - 902355)*
Escherichia coli 0157: H7 EDL933					*CGGTTTATCCCCGCTGATGCGGGGAACAC (3665521 - 3665657)*
Escherichia coli 0157: H7 Sakai	**Z4064**	Z4062	Z4070	Z4065	*CGGTTTATCCCCGCTGGCGCGGGGAACAC (3598284 - 426)*
Erwinia carotovora	ECA3679		ECA3680	ECA3684	TTTCTAAGCTGCCTGTACGGCAGTGAAC (4124333 - 5853)
Exiguobacterium 255-15	Yes		Yes	Yes	ATTTCAATCCACGCACTCACGAGGAGTGCGAC
Fusobacterium nucleatum 25586	**FN1177**	**FN1176**	**FN1179**	**FN1178**	ATTTAAATTCTAATATAGAAATACATAAAT (1836375 - 7327)
Fusobacterium nucleatum 49256	**FNV1819**	**FNV1821**	**FNV1820**		ND

(Continued)

Table 1. Continued

Organism name	Cas 1	Cas 2	Cas 3	Cas 4	CRISPR (Coordinates)
Geobacter metallireducens	Yes	Yes	Yes	Yes	GTAGCGCCCGCCTACATAGGCGGGCGAGGATTGAAAC
Geobacter sulfurreducens		**GSU0058**	GSU0057		*GTATTCCGGGGCCATGATGCCCCGGCCTCATTGAAGC (74807 - 7398)*
''	**GSU1392**	GSU1393	**GSU1384**	GSU1389	GTGTTCCCCGCACATGCGGGGATGAACCG (1521850 - 30352)
Legionella pneumophila (Lens)	lpl2837		lpl2838	lpl2842	GTTCACTGCCGCACAGGCAGCTTAGAAA (3243127 - 4887)
Leptospira interrogans (lai)	LA0684	LA0683	LA0690	LA0689	ND
''	LA3181	LA3182	LA3190		CTGAATATAACTTTGATGCCGTTAGGCGTTGAGCAC (3163420 - 5235)
Leptospira interrogans (Cope)	**LIC10942**	**LIC10940**	**LIC10938**	**LIC10943**	ND
''		LIC12917	LIC12910(1)	LIC12915	ND
Listeria innocua	**lin2743**		**lin2744**		GTTTTGTTAGCATTCAAAATAACATAGCTCTAAAAC (2769058 - 606)
Listeria monocytogenes	**lmo2714**	**lmo2712**	**lmo2713**		GTTTTGGTAGCATTCAAAATAACATAGCTCTAAAAC
Magnetococcus sp. MC-1	Yes	Yes	Yes	Yes	GTTTCAATCCACGCCCCCGTGGGAGGGGGCGAC
Mannheimia succiniciproducens	MS0981	MS0980		MS0985	GTTTCAATTCACGCTCTCGTGTGAGAGCGGAG (954864 - 6998)
''	MS1635	MS1634(6)			GTTTCAATCCCTTTAAGACAGGGCAAGGTCTTTCGA (1630019 - 836)
Methanococcoides burtonii	Yes	Yes		Yes	GTTTCAATCCCTCTAAGGTCTGATTTTAAC, GAGTTCCCCATGCATGTGGGGATAAACCG
Methanothermobacter thermoauot.	**mth1084**	**mth1083**	**mth1086**	**mth1085**	ATTTCAATCCCATTTTGGTCTGATTTTAAC (983373 - 91506)
Methanocaldococcus jannaschii	**MJ0378**	**MJ0386**	**MJ0376**	**MJ0377**	ATTTCCATTCCCCGAGGGATCTGATTTTAC (351707 - 2387)
''		MJ1572.1	MJ1574		GTTAAAATCAGACCTCTTGGAGGATGGAAA (1570374 - 1684)
Methanopyrus kandleri	MK1312	MK1310			CTCGCAATTAACCCGTAATAATATGGGTAATGAAAC (1306914 - 7428)
Methanosarcina acetivorans	MA3670	MA3669	MA3665	MA3671	GTTTCAATCCCTCTAAGGTCTGATTTTAAC (4523522 - 5536)
Methanosarcina barkeri	Yes	Yes	Yes	Yes	GTCGCCCCCTCCCACGGGGGCGTGGATTGAAAC
Methanosarcina mazei	**MM0559**	**MM0558**	**MM0561**	**MM0557**	GTTTCAATCCTTGTTTTAATGGATCTTGCTCGCGAAT (679342 - 82521)
Methylobacillus flagellatus			Yes		GTTTCAATTCACGCACCCGCGCAGGGTGCGAC
Methylococcus capsulatus	MCA0651	MCA0650	MCA0657	MCA0652	GTTTCAATCCACTCCCGGCTATTTAGCCGGGAGATAC (680024 - 4300)
''	MCA0930		MCA0936	MCA0931	GGTCTATCCCCGCGTGTGCGGGGGAGCC (971781 - 2106)
Moorella thermoacetica	Yes (2)	Yes (2)	Yes (2)	Yes (2)	GTTCAAATTCCTCTATGGTCGATGGTCAC, GTCGCGCCCGGCTTCCATGCCGGGCGAGGGTTGAAAC
Mycobacterium tuberculosis 1551	MT2884	MT2883			*GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC (3114128 - 7776)*
Mycobacterium tuber. H37Rv	**Rv2817c**	**Rv2816c**			*GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC (3119410 - 20460)*
Mycoplasma gallisepticum	**MGA_0523**	**MGA_0525**	**MGA_0519**		GTTTTAGCACTGTACAATACTTGTGTAAGCAATAAC (911020 - 5740)
Mycoplasma mobile	**MMOB0320**		**MMOB0330**		GTTTAAGAATACATAAGAATGATACTACACCAAAAC (39815 - 43728)
Nanoarchaem equitans	**NEQ017**	**NEQ016**	**NEQ022**	**NEQ021**	CTTTCAATATTTCTAATATATTAGAAAC (13719 - 4512)
Neisseria meningitidis A	**Nma0630**	**Nma0629**	**Nma0631**		GTTGTAGCTCCCTTTCTCATTTCGCAGTGCTACAAT (608611 - 9357)
Nitrosomonas europaea	NE0844	NE0845	NE0834		GTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTC (139664 - 40171)
''	NE0111	NE0112			GTAGCGCCCGGTCACCAGACCGGGCGAGGATTGAAAC (921023 - 542)
Nocardia farcinica	nfa44220		nfa44280	nfa44230	GGGCTCATCCCCGCACGCGCGTGGAGCAC (4584078 - 403)
Nostoc PCC 7120	alr0381	asr0382			GTTACTTACCATCACTTCCCCGCAAGGGGATGGAAAC (445573 - 7693)
''	alr1468				GTTTTGATTCCTTTACCCCTCACGGGGATGGAAAC (1733001 - 316)
''	alr1568	asr1570	alr1564	alr1567	GTTTCTATTAACACAAATCCCTATCAGGGATTGAAAC (1836813 - 7701)
Nostoc punctiforme	Yes	Yes	Yes	Yes	GTTTCAATCCCTAATAGGGATTTTGATAAATTGCAAT
Pasteurella multocida	PM0311		PM0312	PM0308	GTTCACCATCGTGTAGATGGCTTAGAAAA (362647 - 4587)
''	**Pm1126**	**PM1125**	**PM1127**		GTTGTAGTTCCCTCTCTCATTTCGCAGTGCTACAAT (1322127 - 477)
Photobacterium profundum	PBPRB1995		PBPRB1994	PBPRB1991	TTTCTAAGCTGCCTGTGCGGCAGTGAAC (97592 - 9173)

(Continued)

Table 1. Continued

Organism name	Cas 1	Cas 2	Cas 3	Cas 4	CRISPR (Coordinates)
Photorhabdus luminescens			Plu0745	Plu0750	GTGCACTGCCGTACAGGCAGCTTAGAAA (2126977 - 8369)
”			Plu1791		GTTCACTGCCGTACAGGCAGCTTAGAAA (2142057 - 3271)
Picrophilus torridus	PTO0049	PTO0048			CTTCCATACTATCTAGTAATTCTTAAAC (51349 - 2261)
Porphyromonas gingivalis	*PG2014*	*PG2013*	*PG2016*	*PG2015*	GTTGGATCTACCCTCTATTCGAAGGGTACACACAAC (375957 - 6109)
Pyrobaculum aerophium	PAE0081	PAE0080	PAE0068	PAE0079	CCAGAAATCAAAAGATAGTTGAAAC (45503 - 6677)
”	**PAE0200**	**PAE0199**	**PAE0198**	**PAE0208**	GAATCTCAAAAAGAGGATTGAAAG (95531 - 100995)
Pyrococcus furiosus			PF0640	PF0649	GTTCCAATAAGACTACAAAAGAATTGAAAG (27091 - 30546)
”	**PF1118**	**PF1117**	**PF1120**	**PF1119**	CTTTCAATTCTTTTGTAGTCTTATTGGAAC (1064076 - 5461)
Pyrococcus horikoshii	**PH0173**	**PH0173.ln**	**ph0176**	**ph0175**	GTTTCCACACTACTAAGTTCTACGGAAAC (150418 - 1767)
”	**PH1245**	PH1244.1n	**PH1246**	PHS033(.ln)	CTTTCAATTCTATTTTAGTCTTATTGGAAC (1117810 - 8962)
Rhodospillium rubrum	Yes (5)	Yes (4)	Yes (3)	Yes (2)	*GGTCTATCCCCACGAGTGCGGGGGAACC, GTCTCCCGAGGCAGTAATGCCTCGGCCTCATTGAAAC, CTGTTCCCCGCATGCGCGGGGATGAACC, CTGTTCCCCGCACACGCGGGGATGAACCG, GTTTCGATCCACGCCCCCGTGAAGGGGGCGAC*
Rubrobacter xylanophilus	Yes	Yes	Yes	Yes	*ND*
Salmonella enterica Typhi CT18	STY3065	STY3064	STY3071	STY3066	*CGGTTTATCCCCGCTGGCGCGGGGAACAC (2926184 - 507)*
Salmonella enterica Typhi Ty	t2839		t2845		GTTTATCCCCGCTGGCGCGGGGAACAC (2912043 - 185)
Salmonella enterica Typhimurium	STM2938	STM2937	STM2944	STM2939	*CGGTTTATCCCCGCTGGCGCGGGGAACAC (3076613 - 8139)*
Shewanella sp. (Sargasso Sea)	EAH93242	EAH93243	EAH93237	EAH93241	GTCGCGCCTCCCACAGGCGCGTGGATTGAAAC (46346 - 54413)
Staphylococcus epidermitis RP62A	SE2463	SE2462			*GTTCTCGTCCCCTTTTCTTCGGGGTGGGTATCGATC (2517690 - 857)*
Streptococcus agalactiae NEM316	gbs0912	gbs0913	gbs0911	gbs0910	GTTTTAGAGCTGTGCTGTTTCGAATGGTTCCAAAAC (951908 - 2669)
Streptococcus agalactiae 2603	**SAG0895**	**SAG0896**	**SAG0894**	SAG0893	GTTTTAGAGCTGTGCTGTTTCGAATGGTTCCAAAAC (908164 - 9702)
Streptococcus mutans	**SMU.1404c**	**SMU.1403c**	**SMU.1405c**		ND
”	SMU.1754c*	SMU.1753c	SMU.1764c	SMU.1758c	*GTTTTAGAGCTGTGTTGTTTCGAATGGTTCCAAAAC (1327915 - 8210)*
Streptococcus pyogens 10394				M6_Spy0793	ND
Streptococcus pyogens M1 GAS	**spy1047**	**spy1048**	**Spy1046**	Spy1044	*GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC (860825 - 1175)*
”	**spy1562**	**spy1561**	**spy1567**	**spy1563**	*TTTCAATCCACTCACCCATGAAGGGTGAGAC (1283508 - 659)*
Streptococcus pyogens M49	Yes	Yes (2)	Yes (2)		TTTCAATCCACTCACCCATGAAGGGTGAGAC, GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC
Streptococcus pyogens MGAS 315	SpyM3_678	SpyM3_0679	SpyM3_0677		ND
Streptococcus pyogens SSI-1	SPs1175	SPs1174	SPs1176	SPs1177	ND
Streptococcus suis	**ssui1059**	**ssui1061**	**ssui1060**	Yes	**GTTTTACTGTTACTTAAATCTTGAGAGTACAAAAA**
Streptococcus thermophilus CNRZ	**str0658**	**str0659**	**str0657**		GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC (625101 - 7839)
Streptococcus thermophilus LMG	**stu0658**	**stu0659**	**stu0657**		GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC (629901 - 32033)
Streptomyces avermitilis	SAV7537		SAV7543	SAV7538	GTGCTCTCCGCGCGAGCGGAGGTGAACCG (8993350 - 9003039)
Sulfolobus solfataricus	**SSo1405**	**SSo1404**	**SSo1402**	**SSo1392**	CTTTCAATTCCTTTTAGGATTAATC (1254482 - 60450)
”	SSo1450			SSO1424	*GATAATCTCTTATAGAATTGAAAG (1305634 - 11637)*
Sulfolubus tokodaii	**ST0026**	**ST0025**	**ST0032**		GATGAATCCCAAAAGGAATTGAAAG (32702 - 9695)
”	ST2634	STS264	ST2639	ST2635	CTTTATTCATAATGCTAATTCCGT (2694303 - 448)
Symbiobacterium thermophilum	STH663		STH664	STH669	GTCGTCCCCACGCGCGTGGGGGTGAACCG (725421 - 8869)
Thermoanaerobacter tengcongensis	**TTE2658**	**TTE2657**	**TTE2660**	**TTE2659**	GTTTCAATCCCTCTTAGGTAGGCTAAAAAC (2537224 - 55014)
”	Tfu_1587		Tfu_1593	Tfu_1588	CGGTCCATCCCCACGTGCGTGGGGCTCTC
”	**TVn0106**	**TVn01056**			*GTTTCAATAATTCCTTAGAGGTATGGAAAC (44 - 394)*

(Continued)

Table 1. Continued

Organism name	Cas 1	Cas 2	Cas 3	Cas 4	CRISPR (Coordinates)
Thermotoga maritime	**Tm1797**	**Tm1796**	**Tm1799**	**Tm1798**	GTTTCCATACCTCTAAGGAATTATTGAAAC (1780074 - 627)
Thermus thermophilus HB27	**TTP0196**	**TTP0195**	**TTP0132**	**TTP0197**	*CTTCCATACTAACTAGTACATCTTAAAC (109188 - 10291)*
Treponema denticola	*TDE0328*	*TDE0329*	TDE0327		GTTTGAGAGTTGTGTAATTTAAGATGGATCTCAAAC (373272 - 6975)
Vibrio vulnificus YJ016	VVA1544	VVA1545(6)			GTTTCAGACATGCCCGGTTTAGACGGGATTAAGAC (1694586 - 5242)
Wolinella succinogenes	**WS1444**	**WS1443**	**WS1445**		GTTATAGCCGCCTACTCAGCCATTCCTCGCTATAAT (1386269 - 1386893)
"	WS1615	WS1616		WS1617	GCAACACTTTATAGCAAATCCGCTTAGCCTGTGAAAC (1531926 - 3345)
Xanthomonas axonopodis	**XAC3842**	**XAC3843**	**XAC3837**	**XAC3841**	GTCGCGCCCTCACGGGCGCGTGGATTGAAAC (4522474 - 3421)
Xanthomonas oryzae	XOO0867	XOO0866	XOO0872	XOO0868	GTTTCAATCCACGCGCCCGTGAGGACGCGAC (888612 - 92507)
Yersinia pestis C	YPO2468		YPO2467	YPO2462	GTTTCTAAGCTGCCTGTGCGGCAGTGAAC (2769473 - 795)
Yersinia pseudotuberculosis	YPTB2510		YPTB2509		TTTCTAAGCTGCCTGTGCGGCAGTGAAC (2964488 - 5351)
Zymomonas mobilis	ZMO0680		ZMO0681	ZMO0685	GTTCACTGCCGCACAGGCAGCTTAGAAA (1242995 - 3316)

Locus tags for *cas* genes in a particular cluster are listed as well as the CRISPR sequence that is found associated with that cluster. Numbers in parentheses after locus tags denote a second homolog in this region, designated by the given tag with the final number replaced by the parenthetical number. If permanent locus tags have not yet been assigned, the presence of a *cas* homolog is simply noted. The coordinates of the CRISPRs in question are given in parentheses after the CRISPR sequence (if known). CRISPR sequences and *cas* genes are in bold if they were described previously and are italicized if they are correctly identified in the NCBI database (Jansen et al. 2002a; Bolotin et al. 2005).

cies are presented in Table 1, along with the locus tags of the homologs in question.

We performed phylogenetic analysis of the identified *cas* genes in order to determine whether horizontal gene transfer, or HGT, of these elements had occurred during prokaryotic evolution. Although we initially analyzed each of the *cas* genes separately, we found four separate phylogenetic trees containing approximately 100 members each to be rather unwieldy (*supplemental material*, Fig. A). Multiple alignment data as well as smaller individual trees for most of the *cas* genes in question have been presented previously by others (Jansen et al. 2002a; Makarova et al. 2002; Bolotin et al. 2005). The latter authors have suggested that each of the *cas* gene classes, with the exception of *cas*4, divide into two distinct clades and have thus divided off *cas*1b, *cas*6, and *cas*5 from what we have called *cas* 1, 2 and 3, respectively. We do not see such a clear distinction in our phylogenetic trees and have chosen to keep the classes more inclusive for this analysis (*supplemental material*, Fig. A).

Since the conclusions drawn from each individual tree were qualitatively similar, we chose to create a ''total evidence'' tree to represent our data. Since it is likely that the *cas* genes are transferred among prokaryotes as an entire cassette (Makarova et al. 2002), we decided to join their amino acid sequences prior to phylogenetic analysis. This approach did require that we limit our analysis to loci that contained representatives of all four *cas* genes; this amounted to about one half of the organisms presented in Table 1, specifically 59 loci from 52 different species. The amino acid sequences of the *cas* genes were concatenated to form long sequences with an average size of 1413 residues, and this representation of the data was submitted to multiple alignment followed by the generation of an unrooted phylogenetic tree (*supplemental material*, Figs. B-F; Fig. 2). Certain features of this tree are notable. For instance, major phylogenetic groups which should form their own clade if no HGT had taken place are seen dispersed throughout the tree. The Archaea, Proteobacteria, and Firmicutes are all designated on the tree, none of which cluster together into a single clade.

In searching for CRISPR loci and associated *cas* genes, we noticed that a number of them were not located on the bacterial chromosome but were found on megaplasmids, very large plasmids ≥ 40 kb in size (Table 2). Of the ten megaplasmids described, four of these are the only CRISPR loci found within a particular species, while the rest have CRISPR sequences and *cas* genes on the corresponding chromosome as well. The number of loci associated with megaplasmids varies between one and eight, while the size of the plasmids themselves varies between 39 and 257

Fig. 1. Sequence logo representing the consensus sequence for all CRISPRs characterized to date. The relative size of each nucleotide represents the probability that it is found at each position. The logo has been truncated by 5 bp at each end to remove non-consensus sequence.

kb. The interspersed sequence in one of these plasmids, *Sulfolobus* pNOB8, has been described previously but was not identified as a CRISPR at that time (She et al. 1998). Two other plasmids on our list have their *cas* genes identified in GenBank but the presence of CRISPR loci on these plasmids was not discussed in the corresponding publications (Heidelberg et al. 2004; Henne et al. 2004).

While some of the CRISPR sequences are unique to the megaplasmid on which they are found, others are shared with bacterial chromosomes or other megaplasmids (Tables 1 & 2). The pTT27 megaplasmid from *Thermus thermophilus HB8*, for instance, shares a repeating sequence with the chromosome of the subspecies HB27, while the corresponding plasmid in HB27 shares a different repeating unit with both the HB8 chromosome and megaplasmid. We have extended our analysis by comparing the two pTT27 megaplasmids using a Pustell DNA Matrix (Fig. 3). It is clear that the two plasmids are highly conserved at the level of DNA sequence, with most of their non-conserved regions being made up of CRISPR loci which differ in both sequence and location. Gaps occur in the diagonal in regions which correspond to CRISPR loci coordinates, while segments which deviate from this line can often be traced to the one CRISPR sequence which the two plasmids share, albeit at different locations. It also appears that much of the size difference between the two plasmids (24 kb) can be accounted for by the presence of two more *cas* gene containing loci in the HB8 subspecies. The inter-relatedness of some of the megaplasmid-borne *cas* genes we have described is also apparent from our phylogenetic analysis, namely the *cas* genes found on the *Thermus HB8* megaplasmid show a close phylogenetic relationship with those found on the pSYSA plasmid from *Synechocystis* (Fig. 2).

Additional clues concerning the origin and method of propagation of CRISPRs and their associated genes comes from recent studies of environmental samples taken from the Sargasso Sea. Venter and colleagues have embarked upon a project to obtain large amounts of DNA sequence from uncultured samples taken from the environment (Venter et al. 2004). While it is still relatively uncommon to have sequence data available for a number of different strains of the same species (with obvious exceptions occurring for pathologically important specimens, see Pourcel et al. 2005), environmental sampling has provided a number of closely related genomes belonging to non-pathogenic species.

We searched for CRISPRs using this data and noticed that there was a CRISPR locus in the SAR2 scaffold that corresponds to the complete genome of a Shewanella sp. obtained from the Sargasso Sea. This environmental sample was compared to the *Shewanella oneidensis* genome, a closely related species that completely lacks CRISPR loci (Fig. 4). It can be seen that, while the genes flanking the CRISPR locus share a high degree of homology between SAR2 and *S. oneidensis*, the CRISPR locus itself appears as wide gap of non-conserved sequence. The environmental sample has apparently picked up the CRISPR locus via HGT and incorporated the locus into its genome. Although the four *cas* genes from the *Shewanella* environmental sample are somewhat divergent from their homologs on the *Desulfovibrio* megaplasmid listed in Table 2 at the level of amino acid sequence (Fig. 2, see arrow), this supposed insertion displays a high degree of synteny with the CRISPR locus from the megaplasmid, the order of not only the 4 core *cas* genes but also three additional CRISPR-associated ORFs, *cas5d*, *csd1*, and *csd2*, is completely conserved between the two (Haft et al. 2005; Fig. 5). This arrangement of genes has been deemed the Dvulg subtype by the above authors and it is found not only in *Desulfovibrio vulgaris*, which gives this subtype its name, but also in the 15 other species which share a large clade with the *Desulfovibrio* megaplasmid in our phylogenetic analysis (Fig. 2, unpublished data). It appears that sharing this subtype of CRISPR-associated genes is one of the few things these organisms have in common; they consist of a mix of Proteobacteria, Firmicutes, Spirochetes, and Chlorobi, to name a few phyla.

Through the comparison of these two strains of *Shewanella*, it is possible to estimate the size of at least one CRISPR locus insertion. The SAR2 scaffold shows no similarity to the published genome of *S. oneidensis* between bases 14,480 and 54,988, for a total length of 40,508 bp. The sequences which flank this region, however, show a high degree of similarity to the *Shewanella* genome, with no deletions of genomic DNA detectable. To rule out the possibility that the differences between these genomes

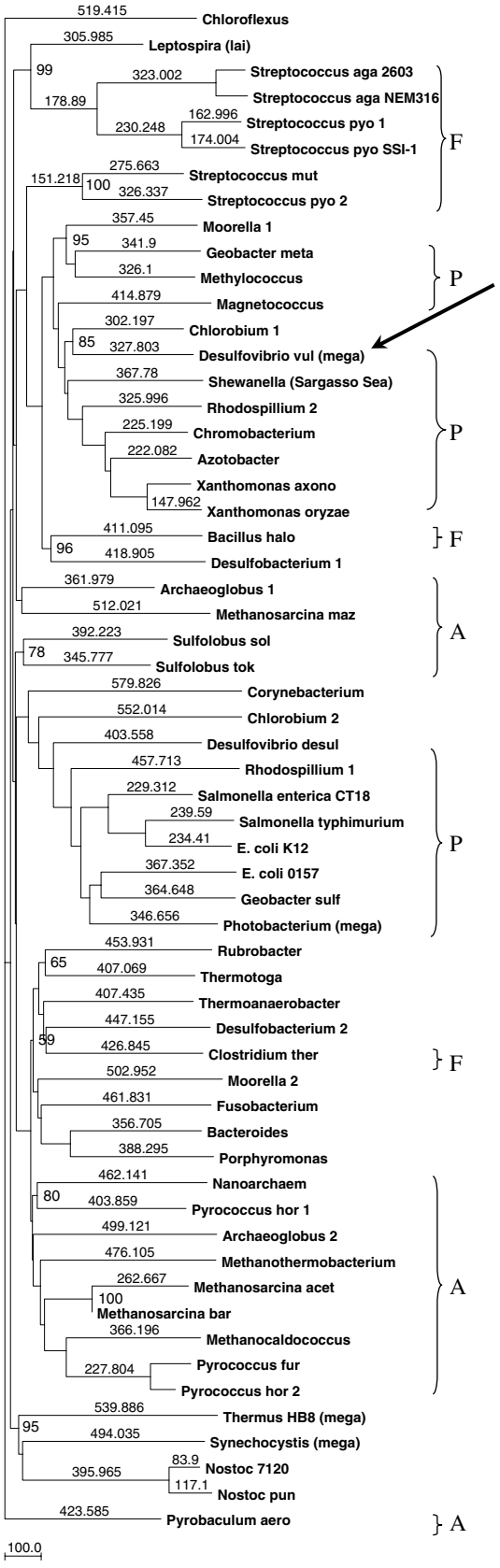

Fig. 2. Total evidence phylogenetic tree for all four *cas* genes. The organism in which *cas* genes are found is listed, these are numbered if more than one cluster is found in a given organism. Members of the division Archaea are indicated by brackets and the letter "A". Members of the phyla Proteobacteria and Firmicutes are indicated with the letters "P" and "F", respectively. Distances between organisms and the next node are given above their branches, bootstrap values greater than 55 (of 1000 replicates) are given next to their corresponding nodes. An arrow indicates the Shewanella sp. from the Sargasso Sea environmental sample and the *Dusulfovibrio vulgaris* megaplasmid.

could be accounted for by a deletion of approximately 40 kb from the organism represented by SAR2 to create the sequence of genes seen in *S. oneidensis*, we have analyzed the arrangement of genes in a third genome, *Shewanella baltica*, that is likely to have diverged from *S. oneidensis* before this predicted insertion took place (Murray et al. 2001). The draft genome of *S. baltica* is almost completely syntenic with the corresponding region in *S. oneidensis*, the locus tags of the *S. baltica* homologs are given in Fig. 5. A single ORF of 112 amino acids, Sbal_1424, lying within the region of our predicted insertion does not contain a *S. oneidensis* homolog in the corresponding region. The number of DNA bases which fall between SO1154 and SO1155 in *S. oneidensis* is 757, while the number of bases between Sbal_1423 and Sbal_1425 in *S. baltica* is 1150, for a difference of 393 bases which contains the ORF described above. Since the size difference between these divergent genomes (*baltica* vs. *oneidensis*) is 100 times less than we find between SAR2 and *S. oneidensis*, we conclude that a deletion event was not responsible for difference between the latter two strains. Our conclusion is that the opposite has occurred- that about 40 kb has been inserted into the *Shewanella* genome to form the SAR2 strain.

Since 40 kb is approximately the size of the smallest CRISPR-containing megaplasmid we have characterized, and all of the megaplasmids presented contain additional genes which lie outside of the CRISPR loci, this raises the possibility that various plasmid-associated genes may be transferred along with a CRISPR locus but would not be identifiable as being CRISPR-associated *per se*. We have identified five additional ORFs upsteam of *cas*3 in the SAR2 scaffold that do not appear in the *Desulfovibrio* megaplasmid and were probably specific to the vector which was responsible for the transfer of this CRISPR locus. These ORFs include a putative integrase, transposase, and plasmid replication protein, as well as TraD (a putative DNA transport protein) and TrwC (a putative conjugative relaxase). The presence of these ORFs in the SAR2 genome also point to a conjugative origin of the adjoining CRISPR locus. It is likely that additional transferred genes which are not required by the CRISPR locus would eventually be deleted by the host genome so that analogous conjugation-associated genes are not evident in many of the loci examined.

Table 2. *Cas* genes and associated CRISPRs from megaplasmids

Megaplasmid [approximate size in kb]	Cas 1	Cas 2	Cas 3	Cas 4	CRISPR
Aquifex aeolicis plasmid ece1 [39]					CTTTCTATCCCATATATGGGAACTAAAAC (437 - 664)
Azoarcus sp. EbN1 plasmid 1 [207]					GTGTTCCCCGCGCATCGCGGGGGTTGAAG (55 - 4714)
Desulfovibrio vulgaris (Hildenborough) [202]	**DVUA0134**	**DVUA0135**	**DVUA0129**	**DVUA0133**	GTCGCCCCCCACGCGGGGGCGTGGATTGAAAC (175898 - 177573)
Haloarcula marismortui plasmid pNG400 [50]	**pNG4053**	**pNG4054**	**pNG4049**	**pNG4052**	GTTACAGACGGACCCTCGTGGGGTTGAAGC (1032 - 1184)
,,					GTTTCAGACGGACCCTTGGGCGGTTGAAGT (35013 - 36556)
,,					GTTACAGACGGACCCTCGTGGGGTTGAAGC (46654 - 49973)
Legionella pneumophila (Lens) plasmid pLPL [60]	plpl0052		plpl0051	plpl0047	TTTCTAAGCTGCCTGTACGGCAGTGAAC (43277 - 46356)
Photobacterium profundum plasmid pPBPR1 [80]	PBPRC0034	PBPRC0033	PBPRC0040	PBPRC0039	CGGTTCAGCCCCGTGAGTACGGGGAACAC (39353 - 43166)
Sulfolobus sp. NOB8H2 pNOB8 [41]					**GATAATCTACTATAGAATTGAAAG** (23581 - 23921)
Synechocystis sp. 6803 plasmid pSYSA [103]	Slr7016	Ssr7017	Slr7010	Slr7015	CTTTCCTTCTACTAATCCCGGCGATCGGGACTGAAAC (17488 - 21365)
,,	Slr7071	Ssr7072			GTTCAACACCCTCTTTTCCCCGTCAGGGACTGAAAC (74027 - 78308)
,,	Slr7092	Ssr7093			GTCTCCACTCGTAGGAGAAATTAATTGATTGGAAAC (97433 - 100351)
Thermus thermophilus HB8 plasmid pTT27 [257]					GTCGCAATCCCCTTACGGGGCTCAATCCCTTGCAAC (18048 - 18224)
,,	TTHB145				GTTGCAAGGGATTGAGCCCCGTAAGGGGATTGCGAC (133766 - 133950[a])
,,	TTHB193		TTHB187		CGGTCCATCCCCACGTGCGTGGGGACTAC (189692 - 190939)
,,					GTAGTCCCCACGCGTGTGGGGATGGACCG (200811 - 201880)
,,	TTHB224	TTHB223	TTHB230	TTHB225	GTTTCAATCCTCTACGAGGCTGACGGGGTTTGCAAC (227472 - 228078)
Thermus thermophilus HB27 plasmid pTT27 [233]					GTTGCAAGGGATTGAGCCCCGTAAGGGGATTGCGAC (92158 - 92340[b])
,,			TTP0132	TTP0136	GTTTCAATCCTCACCGGCCCTTTCGGGCCGGTGCAAC (131658 - 132055)
,,					GTTGCACCGGCCCGAAAGGGCCGGTGAGGATTGAAAC (138802 - 138978)
,,					GTTGCAAGGGATTGAACCCCGTAAGGGGATTGCGAC (153106 - 153289)
,,					GTTTCAATCCTCTACGAGGCTAACGAGGTTTGCAAC (191888 - 192277)
,,					GTTGCAAACCTCGTTAGCCTCGTAGAGGATTGAAAC (203222 - 204045)
,,					GTCGCAATCCCCTTACGGGGCTCAATCCCTTGCAAC (212649 - 213290)

[a]Also 135156 - 136016, 144129 - 144691, and 146042 - 146823.

[b]Also 100454 - 101538.

Listing of locus tags for *cas* genes and their associated CRISPR sequence. The organism that contains the specified megaplasmid is listed. Bold print indicates that *cas* genes are correctly identified in the NCBI database and that the CRISPR sequence has been described previously (She et al. 1998).

Discussion

Previous studies have suggested that the CRISPR-associated genes play a role in DNA recombination and repair (Jansen et al. 2002a). Additional CRISPR-associated genes have been found that act as hydrolases as well as members of the RAMP superfamily (Repair Associated Mysterious Protein) (Makarova et al. 2002). Phylogenetic analyses by these authors indicated the possibility that *cas* genes have been propagated via HGT. These investigators surmised that what came to be known as CRISPR loci represented a novel DNA repair pathway which is found primarily in Archaea as well as hyperthermophilic Eubacteria. Our current work has greatly expanded the number of bacteria known to have CRISPR loci, including a diverse group of mesophiles, thermophiles, aerobes, anaerobes, enterics, heterotrophs, photoautotrophs, etc. It is clear from perusing the 148 species that contain CRISPR loci that they hail from all walks of prokaryotic life and that they do not belong exclusively to any specific groups.

Multiple events of HGT throughout evolution are suggested by our phylogenetic analysis of concatenated sequences from *cas*1-4. While some have argued that homologs of *cas*2 should not be included in the conserved "core" of genes usually found in what are now known as CRISPR loci (Makarova et al. 2002), this work supports their inclusion since we have now expanded the number of these homologs to well within the range of the other three core *cas* genes. The separation of species in the phylogenetic tree we have generated that one would expect to find in close association, as well as the close association in the tree of divergent species that one would expect to be separated, both indicate that these genes have at times been passed by means other than vertical transmission. The precise number of HGT events which have taken place are impossible to estimate due to the complexity of the phylogeny, but that there have been numerous events is without a doubt. Recent studies of the non-repetitive intervening sequences have suggested the origins of this spacer DNA originate from elsewhere in genomes, namely bacteriophage DNA or conjugative transposons (Mojica et al. 2005; Bolotin et al. 2005; Pourcel et al. 2005).

The presence of megaplasmids that contain CRISPR loci suggest a mode of transmission by which CRISPRs have spread so widely throughout prokaryotes. The 40 kb size of the predicted insertion into the *Shewanella oneidensis* genome is approximately the size of the smallest CRISPR-containing megaplasmid characterized to date (plasmid ece1 from *Aquifex aeolicis*, which is 39 kb), both of which are on the same size scale as bacteriophage and conjugative transposons. One difference the CRISPR-containing megaplasmids share with most conjugative transposons described to date is the absence of genes conveying resistance to antibiotics (Scott & Churchward, 1995). This is not to say that other, non-CRISPR-associated genes are not found on CRISPR containing megaplasmids. The pTT27 plasmid, for instance, contains enzymes required for the final stages of carotenoid as well as cobalamin biosynthesis (Henne et al. 2004).

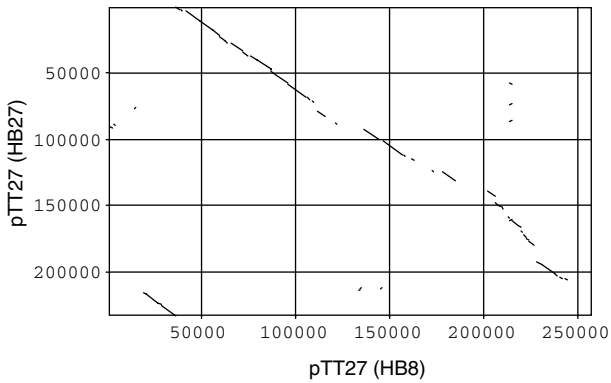

Fig. 3. Comparison of megaplasmids from *Thermus thermophilus*. A Pustell DNA Matrix highlighting conserved sequence between pTT27 plasmids from strains HB8 and HB27. An unbroken diagonal line represents regions of high sequence conservation.

One question that remains is: if megaplasmids are ultimately responsible for the propagation of CRISPRs throughout prokaryotic genomes, why have only ten CRISPR-containing megaplasmids been found, compared to the 148 CRISPR-containing genomes? The answer might be that most megaplasmids are not stably maintained in their host cells and that they often pay transient visits to the species in question, what we have defined as megaplasmids account for one fifth to one quarter of the total number of plasmids contained on the NCBI website for Archaea and Eubacteria, respectively. Notable exceptions to this proposed transience include the pTT27 plasmid discussed above which contains needed biosynthetic enzymes. Other CRISPR-containing megaplasmids may have developed alternate means by which they became stably maintained within a host. To fully answer this question, further studies are needed. Future characterization of CRISPR containing megaplasmids as well as the genes contained therein should shed insight into the function and amazing propagation of the class of DNA repetitive elements known as CRISPRs.

Experimental Procedures

Finding CRISPRs

Whole genomes were downloaded and searched using the program PatScan which was obtained from

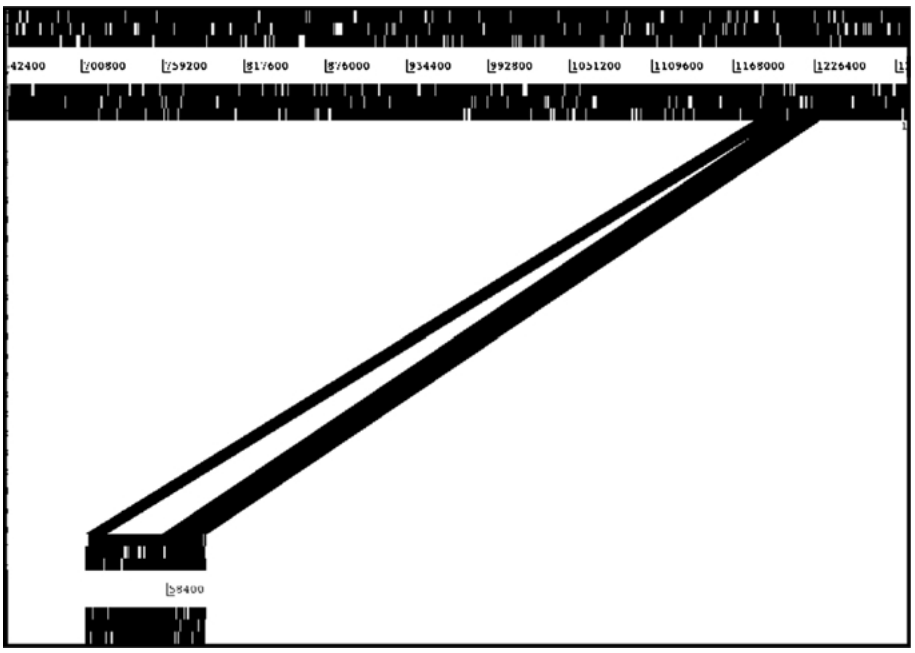

Fig. 4. Alignment of a portion of the *Shewanella oneidensis* MR-1 genome (top, accession # NC_004347.1) with the SAR2 environmental sample taken from the Sargasso sea (bottom, # AACY01119384.1) using the Artemis Comparison Tool (ACT). The CRISPR locus is seen as a gap between overlapping regions. The coordinates of each genome are given.

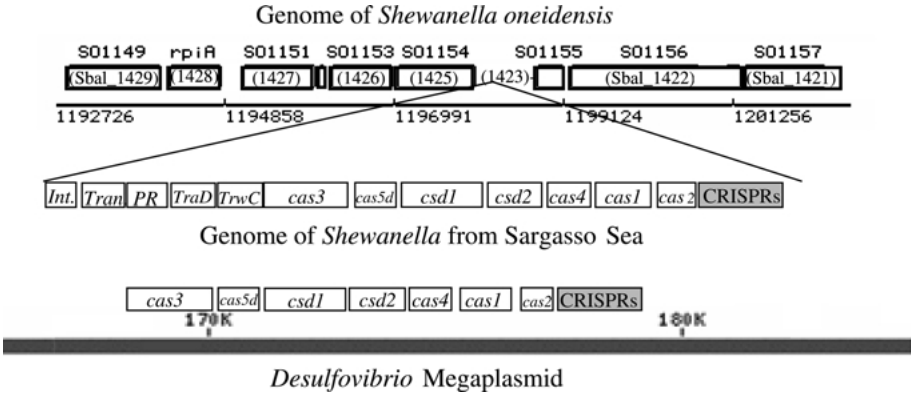

Fig. 5. Model for insertion of CRISPRs and *Cas* genes into *Shewanella*. This is one model which would explain our findings concerning the presence of a CRISPR locus in the SAR2 environmental sample. The synteny between the *Desulfovibrio* megaplasmid and the *Shewanella* sample from the Sargasso Sea is displayed. The *cas* genes are identified by name, as are five additional ORFs which have not been identified as CRISPR-associated. Int., Tran., and PR stand for a putative integrase, transposase, and plasmid replication protein, respectively. The locus tags for *S. baltica* homologs are given in parentheses in the boxes for their corresponding *S. oneidensis* genes.

Argonne National Laboratory, Argonne, IL USA (http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/patscan.html). Initially, the pattern used to find interspersed repeats was: p1 = 15 ... 70 15 ... 70 p1 15 ... 70 p1 (algorithm 1), but this was changed to: p1 = 21 ... 39 15 ... 45 p1 15 ... 45 p1 (algorithm 2) since it was found to speed up our analysis without detrimentally affecting the results obtained. The specific algorithm used to obtain our results is noted in the supplemental materials (Table A). Degenerate tandem repeats were discarded by hand following the analysis. A sequence logo representing the consensus sequence of all CRISPRs characterized was created by submitting a multiple alignment of the sequences to NetLogo at http://weblogo.berkeley.edu. A multiple alignment of 360 sequences was created by aligning each sequence, along with its reverse complement using the ClustalW function of MacVector 7.2.2. The open gap penalty was set at 10, while the extend gap penalty was set at 5, transitions were weighted with a divergent delay of 40%.

Finding Cas *Genes*

Existing *cas* genes (Jansen et al. 2002a) were used as the query for pBLAST searches using the NCBI database (http://www.ncbi.nlm.nih.gov/BLAST/). *Cas* genes were identified based both on their homology to the query, as well as their proximity to a

CRISPR sequence. Often, the newly identified *cas* gene with the lowest amount of homology to the query (E-value $\leq .7$) was then used as a query in a new search to identify divergent members of the *cas* family.

Phylogenetic Analysis of Cas *Genes*

Organisms which contained all four known *cas* genes in a particular CRISPR locus were compared phylogenetically. All phylogenetic analysis was performed using the MacVector 7.2.2 suite of software. The *cas* genes were first concatenated in the following order: *cas* 1, 2, 4, then 3. *Cas*3 was the last protein sequence included since it is the most variable in size. These concatenated sequences were aligned using ClustalW, using an MD 350 matrix for pairwise alignment and an MD series for multiple alignment. Open gap penalties of 10 and extend gap penalties of 0.1 and 0.05 were used in the respective alignments. The neighbor joining method with random tie breaking was used to construct the phylogenetic tree. Gaps were distributed proportionally. The concatenated alignment was visually inspected for significant overlap between the different *cas* genes and none was found.

Alignment of Genomes

Genomic sequences were aligned using the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/ACT/). Input comparison files were generated at the Double ACT website (http://193.129.245.227/pise/double_act.html). A cutoff score of 200 was used. Megaplasmid sequences were compared using a Pustell DNA Matrix generated in MacVector 7.2.2. A window size of 232 was used with a minimum % score of 60, a hash value of 6, and a jump value equal to 1.

Acknowledgments. We would like to acknowledge Jeff Elhai at Virginia Commonwealth University for his many helpful comments regarding this manuscript. We also thank Tom Murphy, who initiated this work at the Bioinformatics and Bioengineering Summer Institute at VCU. We also acknowledge John Iverson, Kabi Neupane, and Sara Penhale for their assistance with the phylogenetic analysis.

References

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

Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A guide of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Computational Biology 1:474–483

Heidelberg JF, Seshadri R, Haveman SA, Hemme CL, Paulsen IT, Kolonay JF, et al. (2004) The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. Nat Biotechnol 22:554–559

Henne A, Bruggemann H, Raasch C, Wiezer A, Hartsch T, Liesegang H, et al. (2004) The genome sequence of the extreme thermophile Thermus thermophilus. Nat Biotechnol 22:547–553

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

Jansen R, van Embden JDA, Gaastra W, Schouls LM (2002) Identification of a novel family of sequence repeats among prokaryotes. Omics 6:23–33

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

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

Mojica FJ, Diez-Villasenor C, Soria E, Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol 36:244–246

Murray AE, Lies D, Nealson K, Zhou J, Tiedje JM (2001) DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc Natl Acad Sci (USA) 98:9853–9858

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

Scott JR, Churchward GG (1995) Conjugative transposition. Annu Rev Microbiol 49:367–397

She Q, Phan H, Garrett RA, Albers S-V, Stedman KM, Zillig W (1998) Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon. Extremophiles 2:417–425

Ussery DW, Binnewies TT, Gouveia-Oliveira R, Jarmer H, Hallin PF (2004) Genome update: DNA repeats in bacterial genomes. Microbiology 150:3519–3521

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74