Chapter 1

General introduction

Gene repertoires and genome organizations differ between closely related microbial organisms depending on the ecological characteristics of each habitat (Cohan and Koeppel 2008). The cyanobacterial *Prochlorococcus* spp. account for a significant fraction of primary production in the ocean (Goericke and Welschmeyer 1993) and show physiological features relevant to the different ecological niches within a stratified oceanic water column (Moore et al. 1998; West et al. 2001). The whole-genomic comparisons of the *Prochlorococcus* spp. strains show gross signatures according to this niche differentiation (Rocap et al. 2003). Alpha-proteobacterium *Pelagibacter ubique* which belongs to the SAR11 clade in the phylogenetic tree based on the 16S rRNA gene is the most abundant microorganism in the ocean (Morris et al. 2002). The genomes of the SAR11 isolates are highly conserved in the core genes that are common to all strains (Medini et al. 2005) and show synteny (the conservation of DNA sequence and gene order) (Bentley and Parkhill 2004). However, variations exist among genes for phosphorus metabolism, glycolysis, and C_1 metabolism, suggesting that adaptive specialization in nutrient resource utilization is important for niche partitioning (Grote et al. 2012). This adaptation at the genomic level was also observed in archaea. The members of the genus *Pyrococcus* are anaerobic and hyperthermophilic archaea (Fiala and Stetter

1986). The archaeal *Pyrococcus* spp. strains also encode genes for survival under high hydrostatic pressure which has been subject to positive selection (Gunbin et al. 2009).

Aeropyrum species are heterotrophic, aerobic, neutrophilic, and hyperthermophilic archaea (Sako et al. 1996). They grow at temperatures up to 100°C and this is the highest growth temperature among the strictly aerobic organisms (Sako et al. 1996). The two currently known species, *Aeropyrum pernix* and *Aeropyrum camini*, were isolated from geographically distinct locations (Sako et al. 1996; Nakagawa et al. 2004). The type strain of the type species, *A. pernix* K1, was isolated from a coastal solfataric vent on Kodakara-Jima Island in southwestern Japan (Sako et al. 1996), and 11 additional strains were isolated from a coastal shallow hydrothermal vent and a coastal hot spring in southwestern Japan (Nomura et al. 2002). The complete genome sequence of *A. pernix* K1 was determined (Kawarabayasi et al. 1999). The type strain *A. camini* SY1 was isolated from a deep-sea hydrothermal vent chimney at the Suiyo Seamount in the Izu-Bonin Arc, Japan, at a depth of 1,385 m (Nakagawa et al. 2004). In chapter 2, I report the complete genome sequence of *A. camini* and compare it to the genome sequence of *A. pernix* in order to examine the genetic differences depending on habitats. This comparative genomic analysis showed that the genomic variation between *A. camini* and *A. pernix* is exerted partly by viruses, although they possess small and highly syntenic genomes.

Aeropyrum spp. belong to the archaeal phylum crenarchaeota. The

crenarchaeota comprises of thermophilic and hyperthermophilic organisms. They inhabit solfataric hot spring or marine hydrothermal vent (Garrity and Holt 2001). The majority of them grow optimally at temperature > 80°C and utilize sulfur compounds widely present in thermal environments (Garrity and Holt 2001). In general, their genome sizes are relatively small and range from 1.3 to 3.0 Mbp (Podar et al. 2008). A phylogenetic birth-and-death maximum likelihood model suggests that this is attributed to extensive gene loss especially during the diversification of taxonomic family-level groups (Csűrös and Miklós 2009; Wolf et al. 2012). Therefore, I hypothesized that some crenarchaea as well as *Aeropyrum* spp. are specialized in their own habitat with small and conservative genome; nevertheless their genomic diversification is driven by viruses through coevolution between hosts and viruses. In chapter 3, I performed a comprehensive comparative analysis of closely related crenarchaeal genomes.

Chapter 2

Comparative genomic analysis of the hyperthermophilic archaea *Aeropyrum camini* and *Aeropyrum pernix*

Introduction

Aeropyrum spp. are aerobic, heterotrophic, and hyperthermophilic marine archaea. *A. pernix* K1 was isolated from a coastal solfataric vent on Kodakara-Jima Island in southwestern Japan (Sako et al. 1996), and 11 additional strains were isolated from a coastal shallow hydrothermal vent and a coastal hot spring in southwestern Japan (Nomura et al. 2002). *A. camini* SY1 was isolated from a deep-sea hydrothermal vent chimney at the Suiyo Seamount in the Izu-Bonin Arc, Japan, at a depth of 1,385 m and is recognized as the first aerobic hyperthermophilic archaeon from a deep-sea hydrothermal environment (Nakagawa et al. 2004). *A. camini* is the sole strain from a deep-sea environment among *Aeropyrum* strains. Despite the geographically distinct habitats of *A. camini* and *A. pernix*, they are phylogenetically closely related based on their 16S rRNA gene sequences (99%, Fig. 2-1) and are similar in morphology and growth characteristics, except for some distinguishable physiological properties such as optimum temperature and pH range for growth (Nakagawa et al. 2004). In this chapter, I determined the complete genome sequence of *A. camini* and compared it with the *A. pernix* genome to determine the genetic differences between close relatives in distinct habitats.

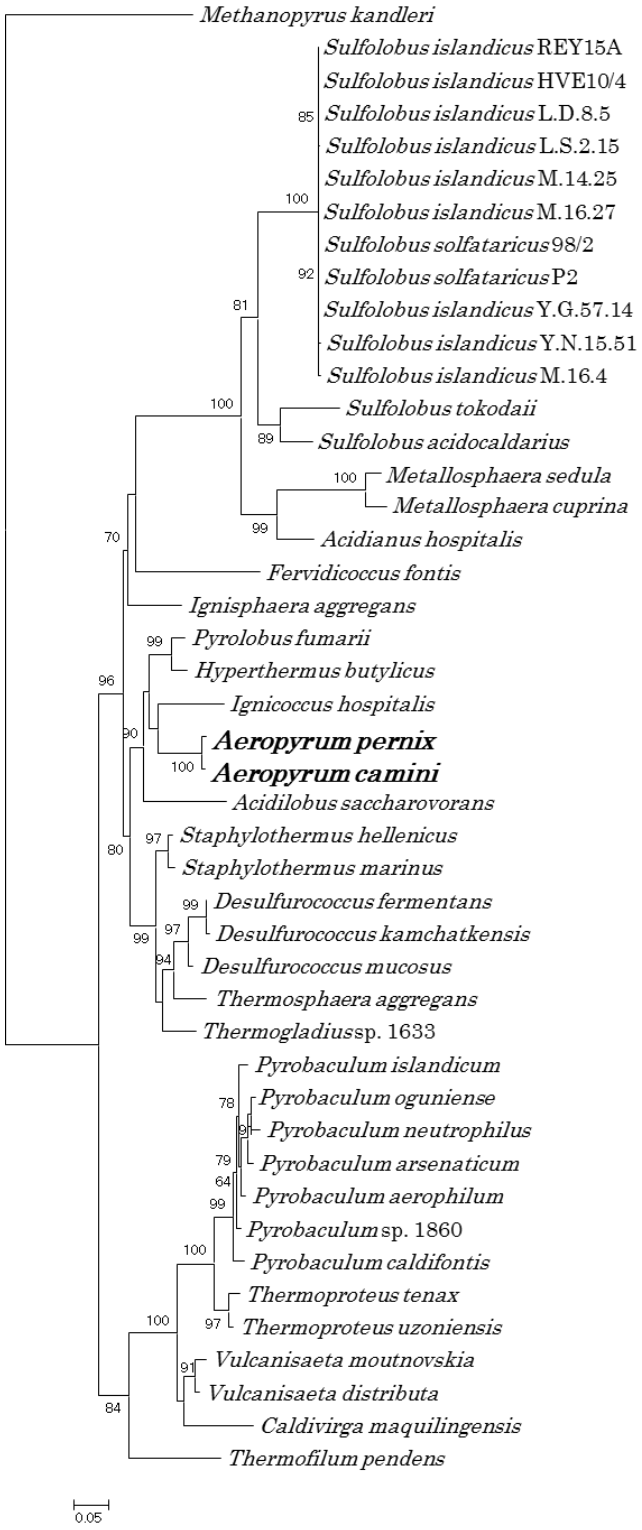

Figure 2-1. Maximum likelihood tree of the16S rRNA genes of the members in the phylum crenarchaeota. *Methanopyrus kandleri* was used as an outgroup in the analysis. Bootstrap values higher than 50 from 100 samplings are shown at branch points.

Materials and Methods

Strain and DNA extraction

A. camini SY1 was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ, Braunschweig, Germany) as DSMZ 16960. *A. camini* cells were grown in a cotton-plugged 2,000 ml Erlenmeyer flask containing 500 ml JXTm medium (Nomura et al. 2002), using an air-batched rotary shaker (RGS-32.TT; Sanki Seiki, Osaka, Japan) at 120 rpm. The pH of the medium was 8.0 and the incubation temperature was 85°C. Cells in the mid-exponential growth phase were harvested by centrifugation at 10,000 × *g* for 1 min at 4°C. Cell pellets were stored at -80°C. DNA was extracted by using a Wizard genomic DNA purification system (Promega, Madison, WI, USA) according to the manufacturer's instructions. DNA was further purified by using phenol-chloroform-isoamyl alcohol (25:24:1) treatment and precipitated with 2-propanol. The DNA was dissolved in 100 µl distilled deionized water.

Genome sequencing and functional annotation

The genome of *A. camini* SY1 was sequenced by using a Roche 454 GS FLX Titanium pyrosequencing platform (Roche Diagnostics, Burgess Hill, West Sussex, United Kingdom) with an 8-kb paired-end library. GS FLX sequencing (one-quarter plate) resulted in the generation of about 116-Mb sequences with an average read length of 342 bases, providing approximately 73-fold coverage of the genome. Reads were assembled onto a

scaffold including 10 large contigs (> 500 bases), using GS *De Novo* Assembler version 2.3. The gaps between the contigs were filled by sequencing of PCR products using a 3130 capillary sequencer (Applied Biosystems, Foster City, CA). The genome sequence was automatically annotated with Microbial Genome Annotation Pipeline version 2.02 (Sugawara et al. 2009). For each predicted open reading frame (ORF), validity was confirmed manually by searching for a putative ribosome binding site (RBS) upstream of the start codon. I modified the position of the start codon of ORFs with no RBS according to the orthologous counterpart encoded on the *A. pernix* genome and confirmed its RBS upstream of the newly predicted start codon. Protein-coding sequences were assigned to clusters of orthologous groups of proteins (COGs) by using RPS-BLAST (Marchler-Bauer et al. 2002), with an E value threshold of 10^{-6} at an effective database size of 10^7. The origins of chromosomal DNA replication were predicted with the Ori-Finder tool (Gao and Zhang 2008).

Comparative genomics

I calculated a genomic similarity score (GSS) to compute similarity between genomes. This measurement is based on the sum of bit scores of shared orthologs, detected as reciprocal best hits (RBHs), and normalized against the sum of bit scores of the compared genes against themselves (self-bit scores). The score has a range from 0 to 1, with a maximum reached when two compared proteomes are identical (Alcaraz et al. 2010). Overall similarity between genomes was generated with the genome-to-genome

distance calculator (GGDC) (Auch et al. 2010). This system calculates the genomic distance and estimates DNA-DNA hybridization (DDH) values from a set of formulas (1, HSP [high-scoring segment pair] length/total length; 2, identities/HSP length; 3, identities/total length). Synteny plots were generated as alignments of the complete genome nucleotide sequences by using MUMMER 3.0 (Kurtz et al. 2004) and Mauve 2.3.1 (Darling et al. 2010). Insertion sequence (IS) elements were identified by using the ISfinder database (Siguier et al. 2006). Multiple copies < 600 bp long flanked by inverted repeats were identified as miniature inverted-repeat transposable elements (MITEs) by using the Einverted program from EMBOSS (Rice et al. 2000).

CRISPR analysis

CRISPR (clustered regularly interspaced short palindromic repeat) elements and spacers were identified by using CRISPRFinder (Grissa et al. 2007) with manual validation. The spacer sequences were clustered by using CD-HIT-EST (Li and Godzik 2006), with a local sequence identity threshold of 90%, an alignment coverage threshold for a shorter sequence of 60%, and a word size set at 7. Two available viral metagenomes from Yellowstone hot springs (Schoenfeld et al. 2008) and from the Juan de Fuca ridge (Anderson et al. 2011) were retrieved from the GenBank trace archive and from the CAMERA database (Seshadri et al. 2007), respectively. A similarity search of spacer sequences was performed against the NCBI nonredundant (nr) database and the viral metagenomes by using BLASTN (Altschul et al. 1990),

with an E value threshold of 10^{-5} and a word size set at 7.

Comparison of protein-coding sequences

A. pernix and *Hyperthermus butylicus* genome sequences were downloaded from the RefSeq database (Pruitt et al. 2007). Putative orthologous genes were identified as RBHs by using BLASTP (Altschul et al. 1997), with a coverage threshold of 50% for both gene sequences and an E value threshold of 10^{-6} at an effective database size of 10^7. Paralogous genes were identified by searching nonorthologous genes against their own proteomes using BLASTP (Altschul et al. 1997), with the parameters noted above and a local identity threshold of 75%. ORFans were identified as sequences without a significant match to those in the NCBI nr database by using BLASTP (Altschul et al. 1997), with an E value threshold of 10^{-6} at an effective database size of 10^7.

Genes acquired by horizontal gene transfer (HGT) events were predicted as previously described (Rhodes et al. 2011). Genes were compared to the nr database by using BLASTP (Altschul et al. 1997), with an E value of 10^{-5} and default parameters. Each gene whose top nonidentical hit was not a gene of a member of the order *Desulfurococcales*, that had a normalized bit score (BLAST bit score to the homolog divided by the BLAST bit score to self) > 25% higher than the best hit to a *Desulfurococcales* gene, and that had a bit score of > 67 was flagged as a putative interorder HGT gene. The donor species were assigned according to the top nonidentical comparisons.

The unclassified genes in the analysis noted above were further

inspected by searching the distributions of homologs in crenarchaeal genomes. In *A. camini*, genes that are homologous to *A. pernix* genes and to its own genes were predicted to be orthologs and paralogs, respectively. Genes whose homologs were distributed in up to five genomes and over five genomes were predicted to be HGT genes and depleted genes in *A. pernix*, respectively. The identical criteria were applied to *A. pernix*.

Results and Discussion

General features

The genome of *A. camini* consisted of a single circular chromosome with no extrachromosomal elements. The general features of the circular chromosome were compared with those of *A. pernix* (Table 2-1). The chromosomes were similar in size (*A. camini*, 1,595,994 bp; *A. pernix*, 1,669,696 bp) and in percent G+C content (*A. camini*, 56.7%; *A. pernix*, 56.3%). Each genome had a single copy of the 16S-23S rRNA operon, a single distantly located 5S rRNA gene, and a total of 47 tRNA genes coding for all 20 amino acids (Table 2-2). Similar numbers of ORFs were identified (*A. camini*, 1,645; *A. pernix*, 1,700). Of all the ORFs, 70.6% and 70.9% were classified into COG categories in *A. camini* and *A. pernix*, respectively.

Table 2-1. Genome statistics of *Aeropyrum* species.

	Value for species	
Attribute	*A. camini*	*A. pernix*
Genome size (bp)	1,595,994	1,669,696
G+C content (%)	56.7	56.3
Total genes	1695	1750
RNA genes (% of total genes)	50 (2.95%)	50 (2.86%)
No. of ORFs (% of total genes)	1645 (97.1%)	1700 (97.1%)
Genes assigned to COGs (% of total ORFs)	1162 (70.6%)	1205 (70.9%)

Table 2-2. tRNA gene assignment in *A. camini*.

UUU (Phe)		UCU (Ser)		UAU (Tyr)		UGU (Cys)	
UUC (Phe)	○	UCC (Ser)	○	UAC (Tyr)	◎	UGC (Cys)	◎
UUA (Leu)	○	UCA (Ser)	○	UAA (Stop)		UGA (Stop)	
UUG (Leu)	○	UCG (Ser)	◎	UAG (Stop)		UGG (Trp)	○
CUU (Leu)		CCU (Pro)		CAU (His)		CGU (Arg)	
CUC (Leu)	○	CCC (Pro)	◎	CAC (His)	○	CGC (Arg)	○
CUA (Leu)	○	CCA (Pro)	○	CAA (Gln)	○	CGA (Arg)	○
CUG (Leu)	○	CCG (Pro)	◎	CAG (Gln)	○	CGG (Arg)	○
AUU (Ile)		ACU (Thr)		AAU (Asn)		AGU (Ser)	
AUC (Ile)	○	ACC (Thr)	○	AAC (Asn)	○	AGC (Ser)	○
AUA (Ile)		ACA (Thr)	○	AAA (Lys)	◎	AGA (Arg)	◎
AUG (Met)	◎[a]	ACG (Thr)	◎	AAG (Lys)	◎	AGG (Arg)	○
GUU (Val)		GCU (Ala)		GAU (Asp)		GGU (Gly)	
GUC (Val)	○	GCC (Ala)	○	GAC (Asp)	◎	GGC (Gly)	○
GUA (Val)	○	GCA (Ala)	○	GAA (Glu)	○	GGA (Gly)	○
GUG (Val)	○	GCG (Ala)	○	GAG (Glu)	○	GGG (Gly)	○

tRNA genes identified are indicated by circles and those containing introns by double-circles.

[a]Two of the three Met-tRNA genes possess an intron.

Although most archaeal genes are predicted to use an AUG start codon, a large percentage of the predicted start codons were GUG (*A. camini*, 27%; *A. pernix*, 30%) or UUG (*A. camini*, 41%; *A. pernix*, 38%). Similar values in the start codon usage were obtained from the archaeon *H. butylicus* (Brügger et al. 2007).

Orthologous genes between *A. camini* and *A. pernix* were identified by using the RBH approach. Each of the genomes carried 1,455 (86 to 88%) orthologous genes (Table 2-3). Genes involved in the Embden-Meyerhof pathway and the tricarboxylic acid cycle were conserved in both genomes.

Table 2-3. Characteristics of protein coding genes encoded on the *A. camini* and *A. pernix* genomes.

	Value for species	
Characteristic	*A. camini*	*A. pernix*
No. of ORFs	1,645	1,700
Orthologous genes	1,455	1,455
Paralogous genes	5	16
ORFans	86	31
Proviral genes	0	70
HGT genes	22	45

The closest relative of *Aeropyrum* spp. is a peptide-fermenting, sulfur-reducing, and hyperthermophilic archaeon, *H. butylicus* (Zillig et al. 1990); *A. camini* and *H. butylicus* shared 772 (46 to 47%) orthologous genes, and *A. pernix* and *H. butylicus* shared 769 (45 to 46%) orthologous genes

(Fig. 2-2). The functional distribution of nonorthologous genes between *Aeropyrum* spp. and *H. butylicus* was inspected (Fig. 2-3). The COG category with the greatest number of nonorthologous genes was energy production and conversion (C), except for two categories, general function prediction only (R) and function unknown (S). This was consistent with the fact that *Aeropyrum* spp. are aerobic, whereas *H. butylicus* is an anaerobic sulfur reducer. *Aeropyrum* spp. contained genes encoding COXs, and *H. butylicus* contained genes encoding a sulfur reductase instead of COXs. Genome variation between *A. camini* and *A. pernix* is described below in detail.

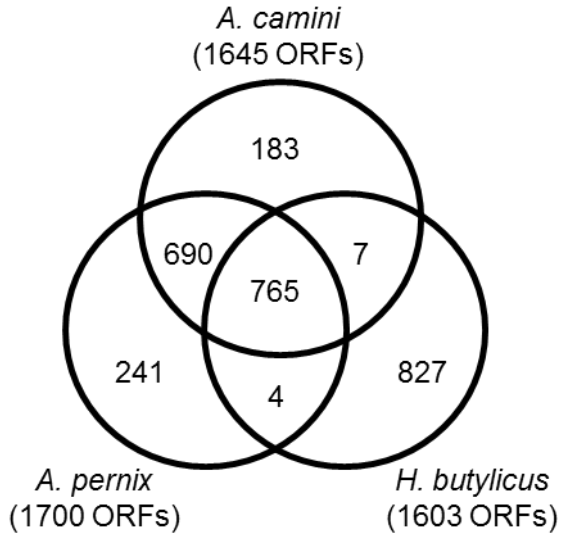

Figure 2-2. Orthologous genes of *A. camini, A. pernix*, and *H. butylicus.* The overlapping circle plots show the numbers of the orthologous genes shared between the genomes.

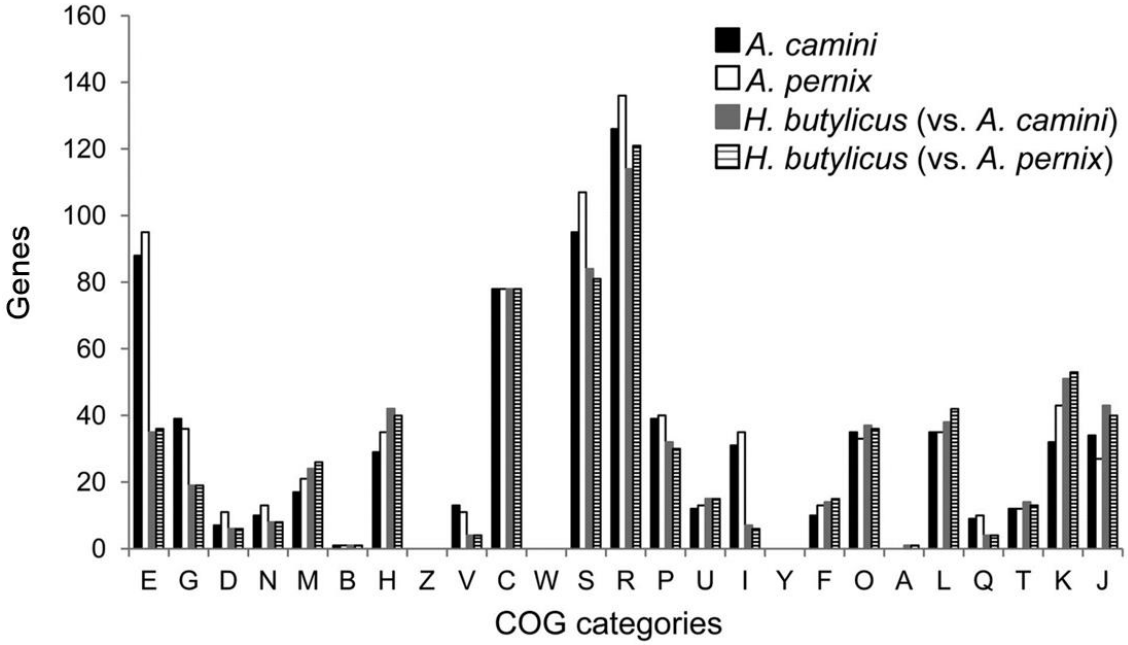

Figure 2-3. Numbers of non-orthologous genes between *Aeropyrum* spp. and *H. butylicus* assigned to COG functional categories. The one letter code for COG categories is the following: E, amino acid transport and metabolism; G, carbohydrate transport and metabolism; D, cell division and chromosome partitioning; N, cell motility and secretion; M, cell envelope biogenesis, outer membrane; B, chromatin structure and dynamics; H, coenzyme metabolism; Z, cytoskeleton; V, defense mechanisms; C, energy production and conversion; W, extracellular structures; S, function unknown; R, general function prediction only; P, inorganic ion transport and metabolism; U, intracellular trafficking and secretion; I, lipid metabolism; Y, nuclear structure; F, nucleotide transport and metabolism; O, posttranslational modification, protein turnover, chaperones; A, RNA processing and modification; L, DNA replication, recombination, and repair; Q, secondary metabolites biosynthesis, transport, and catabolism; T, signal transduction mechanisms; K, transcription; J, translation, ribosomal structure and biogenesis.

The *A. pernix* genome harbors at least two *oriC* sites on noncoding regions containing crenarchaeal origin recognition boxes (ORBs), the binding sites for Orc1/Cdc6 proteins, and *ori*-specific uncharacterized motifs (UCMs) (Robinson and Bell 2007). In the *A. camini* genome, I predicted two *oriC* sites on noncoding regions located between ACAM_0493 and ACAM_0494. Both *oriC* sites coincided with two GC disparity minima described by a Z-curve analysis (Fig. 2-4A). Four copies of the ORB and an UCM were present between ACAM_0493 and ACAM_0494, and an UCM was present between ACAM1253 and ACAM 1254 (Fig. 2-4B and C).

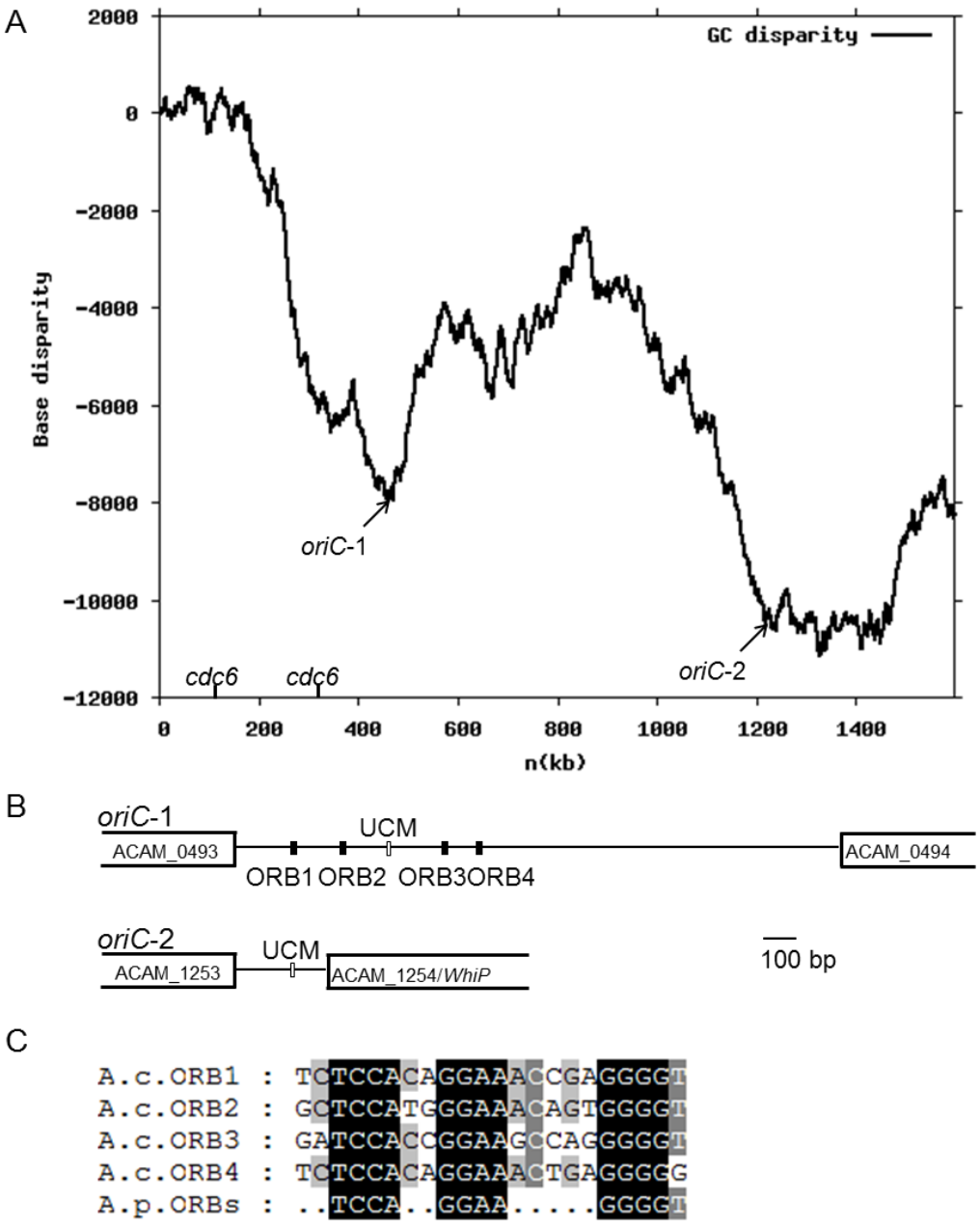

Figure 2-4. Prediction of *A. camini* replication origins. (A) The GC disparity curve for the *A. camini* genome. In the genome map, predicted *oriC* and *cdc6* genes are shown. (B) The structure of the predicted *oriC* region is shown. ORB elements, UCMs, and ORFs flanking the *oriC* site are shown as black boxes, white boxes, and open rectangles, respectively. (C) Alignments of ORB sequences are presented. The four ORB sequences in *A. camini* (A.c.ORB1-4) are compared to the consensus ORB sequences in *A. pernix* (A.p.ORBs), where dots indicate nonconserved bases.

Genome phylogenetics

DDH values estimated by three GGDC formulas were 63.6, 18.9, and 52.0, respectively. Given that the DDH values for the species delineation cutoff are above 70 (Wayne et al. 1987), these data were comparable to a previous report that *A. camini* is a different species from *A. pernix* (Nakagawa et al. 2004). The GSS based on orthologous genes was 0.87, and nucleotide identity was 73.2 to 76.6%, with a range of 86.2 to 90.2% for the two chromosomes, indicating the close relationship of *A. camini* and *A. pernix*. Genome synteny decreases with phylogenetic distance, although this relationship varies depending on the group examined (Tamames 2001; Rocha 2003). Next, I analyzed the degree of genome synteny between *A. camini* and *A. pernix*.

Genome synteny between *A. camini* and *A. pernix*

There were no large-scale rearrangements in the nucleotide alignment of *A. camini* and *A. pernix* chromosomes, confirming the close relationship of them (Fig. 2-5). Comparisons of closely related archaeal and bacterial genomes generally show disruptions of synteny with a characteristic X-shape pattern in the dot plots (Novichkov et al. 2009). The factors that affect genome rearrangements are not well understood but presumably may be associated with the state of recombination systems and the abundance of mobile elements in the respective genomes (Koonin and Wolf 2008). It has been suggested that the low frequency of recombination in *Corynebacterium* spp. is likely due to the absence of RecBCD, a

well-characterized recombinational enzyme complex in bacteria (Nakamura et al. 2003). The RecBCD system was missing in archaea (Blackwood et al. 2013), including *Aeropyrum* spp. Thermoacidophilic archaeal *Sulfolobus* spp. show poor genome synteny owing to genome rearrangements induced by a large number of mobile elements such as IS elements (34 to 201 IS elements) and MITEs (61 to 143 MITEs) (Brügger et al. 2004). The *A. camini* genome carried two IS elements (ACAM_0659 and ACAM_0660) belonging to the IS*607* family and four MITEs, and *A. pernix* carried no IS element and 26 MITEs, indicating that homologous recombinations are less likely to occur at mobile elements. Furthermore, hyperthermophilic organisms are highly specialized in their narrow range of habitat and are isolated from one another by geographic barriers (Whitaker et al. 2003). *Aeropyrum* spp. therefore can be defined as specialists in the concept of specialists as opposed to generalists, where specialists often have small genomes harboring genes essential for cell maintenance and most generalists have large genomes harboring additional genes for signal transduction or metabolism, allowing survival in variable environments (Koonin and Wolf 2008; Newton et al. 2010). In the highly "specialized" small genomes of *Aeropyrum* spp., the disruption of gene regulation derived from synteny breaks may be limited due to elimination of individuals associated with reduced fitness.

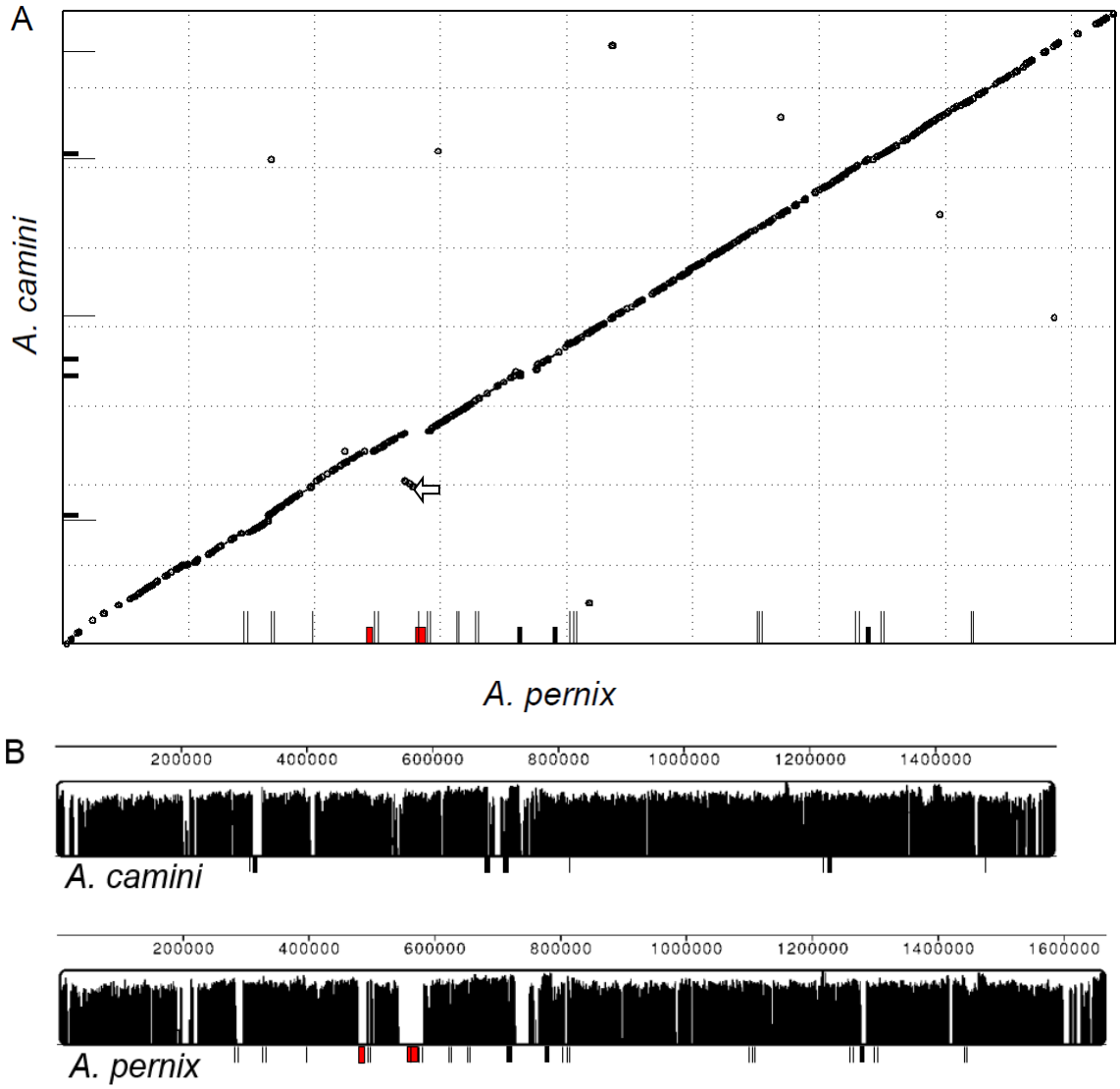

Figure 2-5. Comparison of the chromosomes of *A. camini* and *A. pernix*. (A) MUMMER nucleotide alignment, where dots indicate similar sequences shared by the two species. (B) Mauve nucleotide alignment, where the height of plots is proportional to the level of sequence identity in that region. Proviral regions, CRISPR elements, and MITEs are shown on the map in red boxes, filled boxes, and thin lines, respectively, on the two nucleotide alignments. A translocated inversion upstream of the proviral region was indicated by an empty arrow.

Virus-related elements

Although both genomes showed synteny, I observed some synteny disruptions. The most prominent disruptions were contained in virus-related elements. First, *A. pernix* contains two proviral regions that were induced under suboptimal conditions (Mochizuki et al. 2011). Two viruses containing circular double-stranded DNA (dsDNA) genomes were isolated and named *Aeropyrum pernix* spindle-shaped virus 1 (APSV1) and *Aeropyrum pernix* ovoid virus 1 (APOV1), respectively (Mochizuki et al. 2011). The proviral sequences were absent from the *A. camini* genome at the conserved tRNA sequences homologous with *attP* sites (Fig. 2-6), the recombination sites for viruses, although I could not rule out the possibility that *A. camini* was cured of the proviruses in the isolation step, repeated at least three times (Nakagawa et al. 2004). A translocated inversion of a 2-kbp sequence was identified upstream of the integrated APSV1 genome (Fig. 2-5 A, an empty arrrow). The inversion might be caused by a 12-bp inverted repeat observed in that region.

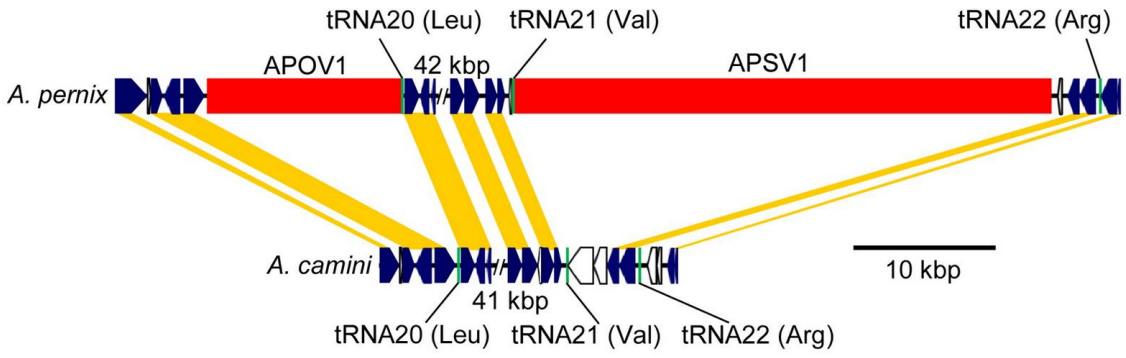

Figure 2-6. Two proviral regions (APOV1 and APSV1) were present in the *A. pernix* genome and absent in the *A. camini* genome. Proviral regions, tRNAs, and ORFs are shown as red boxes, green vertical lines, and arrows, respectively. Orthologous genes are shown in navy blue highlighted by orange.

Second, synteny disruptions were observed in the CRISPR elements (Fig. 2-5). The CRISPR system is a recently recognized defense mechanism in archaea and bacteria against foreign DNA such as viruses and plasmids (Sorek et al. 2008). CRISPR allows cells to specifically recognize and destroy target sequences using spacers derived from invaders and, in many respects, parallels the function of the eukaryotic RNA interference system (Makarova et al. 2006). CRISPR spacers effectively act as libraries of previous invasion by extrachromosomal elements. In practice, host-virus interactions are investigated by the analysis of CRISPR spacers in the natural cyanobacterial community (Kuno et al. 2012). *A. camini* contained four CRISPR loci (Aca_1 to Aca_4), composed of 14, 15, 27, and 3 repeat-spacer units, respectively (Table 2-4). Aca_1 and Aca_3 were interrupted by genes that did not show similarity to any other available protein sequences. According to the CRISPRdb database, the *A. pernix* genome carried three CRISPR loci (Ape_1 to Ape_3), composed of 26, 41, and 18 repeat-spacer units, respectively (Table 2-4). Each noncoding sequence upstream of the first CRISPR of all CRISPR elements was AT rich (percent G+C content ranging from 31.5 to 52.0% lower than that of each genome sequence) and included a RBS, a TATA box, and a B recognition element. Therefore, the sequences were considered to be leader sequences that are transcription initiation sites for the CRISPR (Fig. 2-7, empty boxes) (Lillestøl et al. 2006). CRISPR-associated (*cas*) genes were adjacent to Aca_1, Aca_3, and Ape_2 (Fig. 2-7) but not to the others. CRISPR/Cas types are classified on the basis of their repeat sequences, leader sequences, and *cas* genes (Makarova et al. 2011a).

Table 2-4. Characteristics of the CRISPR elements of *A. camini* and *A. pernix*.

Species	CRISPR locus	CRISPR type[a]	Position	No. of repeat-spacer units	Typical repeat sequences (5'-3')	No. of spacers with significant hits[b]	
						APSV1	APOV1
A. camini	Aca_1	-	313030..313907, 314206..314270	14	GAATCTTCGCGATAGAATTGCGAG	-	-
	Aca_2	-	679471..680511	15	GAATCTTCGAGATAGAATTGCAAG	-	-
	Aca_3	I-A	737714..738255, 738626..739902	27	GCATATCCCTAAAGGGAATAGAAAG	2	1
	Aca_4	-	1224281..1224496	3	GAATCTTCGAGATAGAATTGCAAG	-	-
A. pernix	Ape_1	-	717248..718997	26	GAATCTTCGAGATAGAATTGCAAG	1	-
	Ape_2	I-A	786657..789355	41	GCATATCCCTAAAGGGAATAGAAAG	-	-
	Ape_3	-	complement (1277299..1278486)	18	CTTGCAATTCTATCTCGAAGATTC	-	-

[a] Dashes indicate that the CRISPR type cannot be identified.

[b] Dashes indicate that no comparison was found for the spacers.

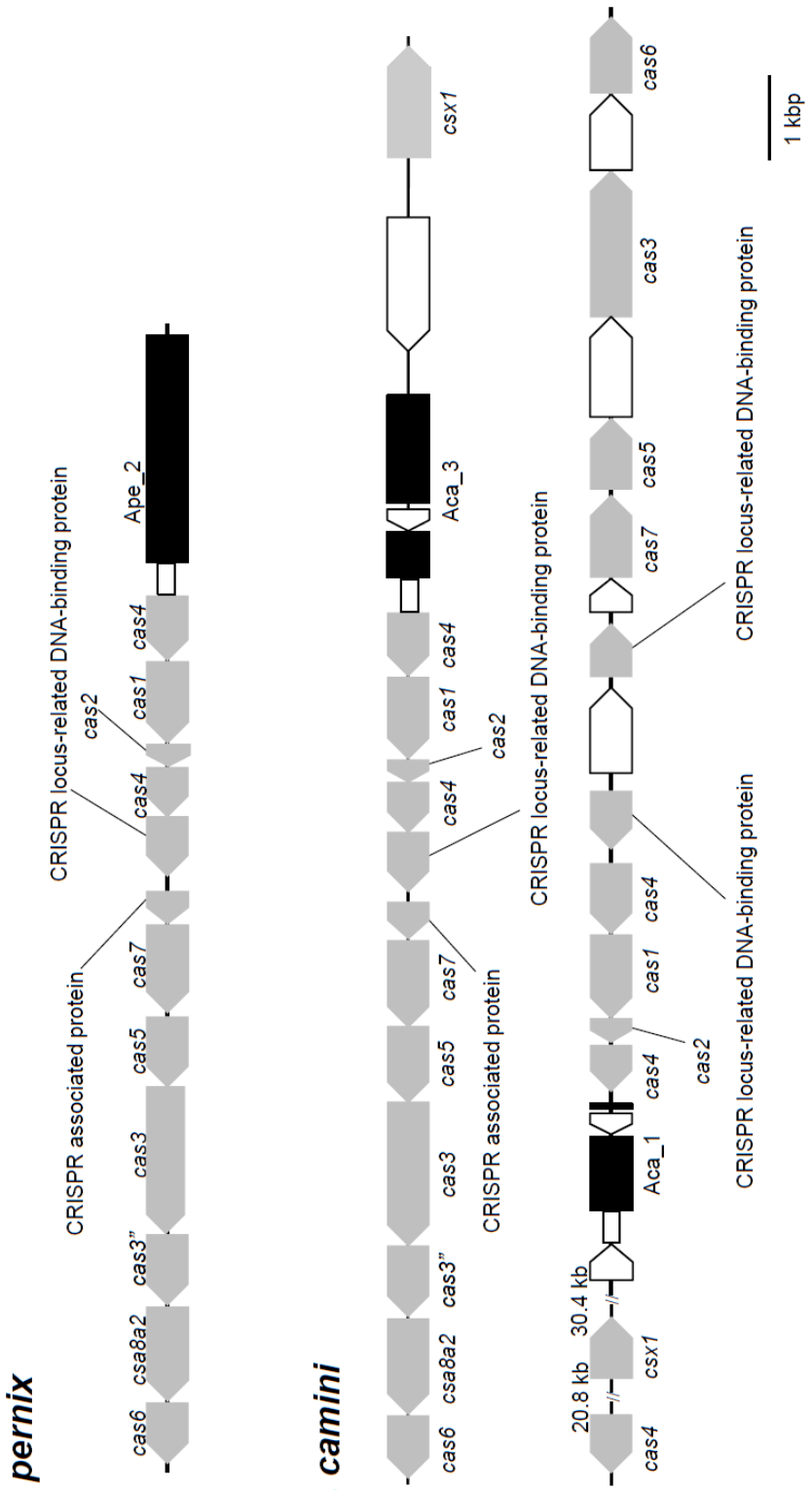

Figure 2-7. Schematic representations of the Ape_2, Aca_3, and Aca_1 CRISPR/Cas systems. ORFs, leader sequences, and CRISPRs are shown as arrows, empty boxes, and filled boxes, respectively. *Cas* genes are indicated in gray.

In a previous report, the Ape_2 CRISPR/Cas system was determined to belong to DNA-targeting subtype I-A (Makarova et al. 2011a). A subtype I-A CRISPR/Cas system homologous to the Ape_2 CRISPR/Cas system was identified in *A. camini* (Aca_3 CRISPR/Cas system) (Fig. 2-7). The CRISPR type of the other CRISPR (Aca_1, Aca_2, Aca_4, Ape_1, and Ape_3) was not identified due to the uniqueness of the typical repeats and the leader sequences or the absence of signature genes for a subtype of the CRISPR/Cas system notwithstanding the presence in the Aca_1 CRISPR/Cas system of *cas3* peculiar to the type I CRISPR/Cas system (Fig. 2-7).

Fifty-nine and 85 CRISPR spacers were retrieved from *A. camini* and *A. pernix*, respectively, and no significant matches were found among them. When all 144 spacers were compared to the NCBI nr nucleotide database, 3 spacers (2 spacers in Aca_3 and a spacer in Ape_1) and a spacer in Aca_3 showed a significant match to the genomes of APSV1 and APOV1, respectively (Table 2-4). This strongly suggested that *Aeropyrum* spp. CRISPR/Cas may have been functional, at least in the past. *A. pernix* encoded a spacer (Ape_1_4) significantly matched with the genome of APSV1 integrated into its genome. In general, single-nucleotide mutation of targeted sequences can render CRISPR/Cas ineffectual (Barrangou et al. 2007; Deveau et al. 2008). *A. pernix* may avoid destroying its own genome due to 5 nucleotide discrepancies between the Ape_1_4 spacer and the proviral sequence (Table 2-5). Of three provirus-matching spacers in Aca_3, two spacers showed synonymous and nonsynonymous substitutions compared with their corresponding putative protospacers in proviruses

(Table 2-5), indicating that *A. camini* interacted with viruses that were closely related to APSV1 and APOV1. All CRISPR spacers did not show a significant match with any other nucleotide sequences in the nr database or viral metagenomes from Yellowstone hot springs (Schoenfeld et al. 2008) and the Juan de Fuca ridge (Anderson et al. 2011) other than the APSV1 and APOV1 genomes. It is noteworthy that none of the CRISPR spacers matched the nonorthologous genes of *Aeropyrum* spp., which are extrachromosomal elements in most cases (described below). In this research, I examined only 144 spacers collected from two *Aeropyrum* spp. More CRISPR spacers might enable us to identify the spacers matched with nonorthologous genes.

Table 2-5. Spacers compared to APSV1 and APOV1 for putative proto-spacers.

Spacer/virus gene[a]	Nucleotide sequence[c]	Predicted amino acid sequence[c]
Ape_1_4	GGTCCTGGTCTTGCTCCCCCGCGACTACTGGCAGCTCTTCCAGGG	VLVLLPRDYWQLFQ
ORF52 (APSV1)	. . . **GT** . . . . **GC** . **C**	. . . . .
Aca_3_12	AGCCCCCTGGCTCCATGGAAGCGTATAGCAAGAATAGTACCGG	P PGSMEAYSKNST
ORF53 (APSV1)	. C *G* . . *T* C	AS . . . . .
Aca_3_19[b]	CGCTGGGCATACCGCCCAGCAGCACACACGGGCTCATGCAG	LGIPPSSTHGLMQ
ORF4 (APOV1)	. . . . . **A** . . . . . *G*	. . . . . A . . . . .
Aca_3_25	GGCGGGCGTGGACTACAGGCTCCAGCCGTACCTGCCAA	GGRGLQAPAVPAX
ORF51 (APSV1)	. . . . .	. . . . .

[a] In each row, the spacer (top) and the corresponding putative protospacer (bottom) are shown.

[b] A reverse complementary sequence is shown.

[c] Identical nucleotides and amino acids are indicated by dots. Synonymous and nonsynonymous substitutions are shown in bold faces and italic types, respectively.

Nonorthologous genes in *A. camini* and *A. pernix*

Along with the virus-related elements that primarily contributed to the synteny disruptions, nonorthologous genes were located on nonsyntenic regions scattered over the *A. camini* and *A. pernix* genomes. In the *A. camini* genome, on the other hand, 56 nonorthologous genes (29%) were localized at kbp 13 to 22, kbp 314 to 331, kbp 407 to 411, and kbp 687 to 715. In the *A. pernix* genome, except for the proviral regions, 73 nonorthologous genes (30%) were localized at kbp 190 to 211, kbp 284 to 305, kbp 726 to 764, kbp 1279 to 1286, and kbp 1599 to 1644.

Of these genes, notable metabolic pathways missed in *A. camini* were on the nonsyntenic regions at kbp 726 to 764 and kbp 1279 to 1286 in *A. pernix*, reflecting the smaller genome of *A. camini* than that of *A. pernix*. L-Rhamnose is a common component of the cell wall in bacteria (Giraud and Naismith et al. 2000) and is also found in the cytoplasmic membrane of the methanogenic archaeon *Methanospirillum hungatei* (Sprott et al. 1983). The *rmlABCD* genes, involved in a nucleotide-activated L-rhamnose (dTDP-L-rhamnose) synthesis pathway, were identified (APE_1178 to APE_1181) on the nonsyntenic region at kbp 726 to 764 in *A. pernix*. Cobamides (e.g., coenzyme B_{12}) are unique for their structural complexity, and archaea synthesize them through salvaging cobinamide from the environment (Escalante-Semerena 2007). Clustered genes involved in the cobinamide-salvaging pathway were found on the nonsyntenic region at kbp 1279 to 1286 in *A. pernix*. These facts implied that *A. camini* may not be able to synthesize L-rhamnose and cobamides.

A previous report showed the geographical distribution of gene contents (e.g., mobile elements) among *Sulfolobus islandicus* strains from hot springs separated by distance (Reno et al. 2009). This suggests that the variation of metabolic pathways in *Aeropyrum* implies their locality, although the pathways are not necessarily responsible for environmental adaptation. Meanwhile, genetic islands are found within genomes of *S. islandicus* strains from a single hot spring (Cadillo-Quiroz et al. 2012). The variation among strains might be found in future analyses of more *Aeropyrum* species genomes.

Of all the nonorthologous genes, paralogous genes were identified (5 genes for *A. camini* and 16 genes for *A. pernix*) in the range of 3 to 7% by performing searches against their own proteomes (Table 2-3). In the *A. pernix* genome, eight paralogous genes were annotated as encoding hypothetical proteins with no conserved domains; however, these nucleotide sequences contained the MITEs noted above. The other genes were classified into ORFans (86 genes for *A. camini* and 31 genes for *A. pernix*), which did not show similarity to any other available protein sequences in the nr database; HGT genes (22 genes for *A. camini* and 45 genes for *A. pernix*); and proviral genes (70 genes for *A. pernix*) (Table 2-3). HGT events are likely to occur among organisms with similar life-styles and habitats, in particular among archaeal and bacterial hyperthermophiles (Rhodes et al. 2011). The donors of the HGT genes identified in the *Aeropyrum* spp. genomes were thermophiles or derived from environmental sequences collected from the thermophilic environment in the range of 82 to 84% (Table 2-6). These data

were compatible with the concept that *Aeropyrum* spp. are specialized in the thermophilic environment. The unclassified genes in the analysis described above were inspected further (Table 2-7).

Table 2-6. Donor of the HGT genes in *A. camini* and *A. pernix*.

ORF	Donor	Thermal environment[a]
ACAM_0016	*Alicyclobacillus acidocaldarius*	+
ACAM_0017	*Clostridium scindens*	−
ACAM_0018	uncultured marine microorganism HF4000_ANIW141A21	−
ACAM_0231	*Kyrpidia tusciae*	+
ACAM_0232	*Vulcanisaeta distributa*	+
ACAM_0233	*Metallosphaera sedula*	+
ACAM_0344	*Acidilobus saccharovorans*	+
ACAM_0357	*Acidilobus saccharovorans*	+
ACAM_0363	*Aciduliprofundum boonei*	+
ACAM_0365	*Archaeoglobus profundus*	+
ACAM_0374	*Vulcanisaeta distributa*	+
ACAM_0743	*Candidatus* Caldiarchaeum subterraneum	+
ACAM_0756	*Halorubrum lacusprofundi*	−
ACAM_0757	*Acidilobus saccharovorans*	+
ACAM_0765	*Nodularia spumigena*	−
ACAM_1511	*Sphaerobacter thermophilus*	+
ACAM_1512	*Candidatus* Caldiarchaeum subterraneum	+
ACAM_1606	*Ferroglobus placidus*	+
ACAM_1607	*Ferroglobus placidus*	+
ACAM_1611	*Archaeoglobus profundus*	+
ACAM_1614	*Archaeoglobus profundus*	+
ACAM_1621	*Pyrobaculum aerophilum*	+
APE_0025.1	*Pyrococcus furiosus*	+
APE_0026	*Thermobispora bispora*	+
APE_0028	*Stackebrandtia nassauensis*	−

Table 2-6. Continued.

APE_0031.1	*Pyrobaculum aerophilum*	+
APE_0203.1	*Pyrobaculum calidifontis*	+
APE_0266.1	*Candidatus* Caldiarchaeum subterraneum	+
APE_0275.1	*Aciduliprofundum boonei*	+
APE_0276.1	*Thermococcus* sp. AM4	+
APE_0276a	*Thermococcus gammatolerans*	+
APE_0287	*Candidatus* Caldiarchaeum subterraneum	+
APE_0288	*Actinosynnema mirum*	−
APE_0302.1	*Candidatus* Korarchaeum cryptofilum	+
APE_0303	*Candidatus* Korarchaeum cryptofilum	+
APE_0304.1	*Candidatus* Korarchaeum cryptofilum	+
APE_0472	*Thermococcus* sp. AM4	+
APE_0472a	*Thermofilum pendens*	+
APE_0688	*Clavibacter michiganensis* subsp. *michiganensis*	−
APE_1061.1	*Acidilobus saccharovorans*	+
APE_1183	*Sulfolobus islandicus* M.14.25	+
APE_1188	*Desulfovibrio fructosovorans*	−
APE_1189.1	*Pyrobaculum aerophilum*	+
APE_1191	*Pyrobaculum islandicum*	+
APE_1192	*Ferroglobus placidus*	+
APE_1209b	*Archaeoglobus fulgidus*	+
APE_1245.1	*Thermofilum pendens*	+
APE_1275	*Thermanaerovibrio acidaminovorans*	+
APE_1473a	*Acidilobus saccharovorans*	+
APE_1558	*Candidatus* Caldiarchaeum subterraneum	+
APE_1588	*Caldicellulosiruptor obsidiansis*	+
APE_1921	*Vulcanisaeta distributa*	+
APE_1929.1	*Candidatus* Caldiarchaeum subterraneum	+
APE_2041.1	*Acidilobus saccharovorans*	+
APE_2042.1	*Flavobacterium johnsoniae*	−
APE_2240	*Sulfolobus acidocaldarius*	+
APE_2242.1	*Chloroflexus aurantiacus*	+
APE_2380.1	*Archaeoglobus fulgidus*	+
APE_2520.1	*Pyrococcus horikoshii*	+

Table 2-6. Continued.

APE_2521.1	uncultured archaeon	−
APE_2522.1	*Achromobacter piechaudii*	−
APE_2523.1	*Thermotoga lettingae*	+
APE_2524.1	*Archaeoglobus profundus*	+
APE_2577.1	*Acidianus* two-tailed virus	+
APE_2581	*Sulfolobus solfataricus* 98/2	+
APE_2604a.1	*Acidilobus saccharovorans*	+
APE_2617.1	*Thermococcus gammatolerans*	+

[a] Pluses and dashes indicate that the donors are from thermal environment and non-thermal environment, respectively.

Table 2-7. Characteristics of nonorthologous genes in *A. camini* and *A. pernix*.

ORF	Position	Strand	Product length (amino acids)	COG No.	Code	COG Function	Designation
ACAM_0002	1746..2324	−	193	−	−	−	paralogous gene
ACAM_0004	3137..3580	+	148	−	−	−	ORFan
ACAM_0009	9376..9939	−	188	−	−	−	ORFan
ACAM_0013	13618..14220	−	201	−	−	−	ORFan
ACAM_0015	15372..15938	+	189	1225	O	Bcp, Peroxiredoxin	paralogous gene[a]
ACAM_0016	15963..16667	+	235	−	−	−	HGT gene
ACAM_0017	16652..16984	+	111	3118	O	Thioredoxin domain−containing protein	HGT gene
ACAM_0018	17098..17670	−	191	−	−	−	HGT gene
ACAM_0019	17683..17982	−	100	−	−	−	ORFan
ACAM_0020	17993..20368	−	792	0843	C	CyoB, Heme/copper−type cytochrome/quinol oxidases, subunit 1	paralogous gene[a]
ACAM_0021	20370..21110	−	247	1622	C	CyoA, Heme/copper−type cytochrome/quinol oxidases, subunit 2	paralogous gene[a]
ACAM_0022	21132..21734	−	201	−	−	−	ORFan
ACAM_0069	71004..71132	+	43	−	−	−	ORFan
ACAM_0091	96079..96228	+	50	−	−	−	depleted in *A. pernix*[a]

Table 2-7. Continued.

ACAM_0156	157663..159921	+	753	1205	R	Distinct helicase family with a unique C−terminal domain including a metal−binding cysteine cluster	paralogous gene[a]
ACAM_0183	183256..183525	−	90	−	−	−	orthologous gene[a]
ACAM_0203	200811..201065	+	85	2034	S	predicted membrane protein	ORFan
ACAM_0207	203287..203478	+	64	−	−	−	ORFan
ACAM_0208	203475..203702	+	76	−	−	−	HGT gene[a]
ACAM_0209	203710..203922	+	71	−	−	−	paralogous gene[a]
ACAM_0210	205162..205368	−	69	−	−	−	paralogous gene
ACAM_0211	205833..206255	−	141	−	−	−	HGT gene[a]
ACAM_0212	206300..206512	−	71	−	−	−	ORFan
ACAM_0213	207140..207592	−	151	1848	R	Predicted nucleic acid−binding protein, contains PIN domain	paralogous gene[a]
ACAM_0214	207589..207786	−	66	−	−	−	paralogous gene
ACAM_0215	208128..208325	−	66	−	−	−	ORFan
ACAM_0216	208243..208527	−	95	−	−	−	ORFan
ACAM_0217	208475..209503	+	343	−	−	−	ORFan
ACAM_0219	210351..212474	−	708	−	−	−	ORFan
ACAM_0226	215709..216128	−	140	−	−	−	orthologous gene[a]
ACAM_0231	220811..222031	+	407	1960	I	CaiA, Acyl−CoA dehydrogenases	HGT gene
ACAM_0232	222039..222569	+	177	2030	I	MaoC, Acyl dehydratase	HGT gene

Table 2-7. Continued.

ACAM_0233	222544..224007	+	488	0427	C	ACH1, Acetyl−CoA hydrolase	HGT gene
ACAM_0243	232452..232787	+	112	−	−	−	ORFan
ACAM_0248	236956..237300	−	115	−	−	−	ORFan
ACAM_0280	264288..264557	−	90	−	−	−	ORFan
ACAM_0298	281040..281777	+	246	−	−	−	ORFan
ACAM_0299	281795..282232	+	146	−	−	−	ORFan
ACAM_0334	306069..306308	−	80	−	−	−	ORFan
ACAM_0343	313912..314166	−	85	−	−	−	ORFan
ACAM_0344	314428..314991	−	188	1468	L	RecB family exonuclease	HGT gene
ACAM_0345	315008..315301	−	98	1343	L	Uncharacterized protein predicted to be involved in DNA repair	paralogous gene[a]
ACAM_0346	315292..316287	−	332	1518	L	Uncharacterized protein predicted to be involved in DNA repair	paralogous gene[a]
ACAM_0347	316297..317148	−	284	4343	S	Uncharacterized protein conserved in archaea	paralogous gene[a]
ACAM_0348	317302..317997	−	232	−	−	−	depleted in *A. pernix*[a]
ACAM_0349	318211..319233	+	341	2905	T	Predicted signal−transduction protein containing cAMP−binding and CBS domains	paralogous gene[a]
ACAM_0350	319349..319984	+	212	−	−	−	depleted in *A. pernix*[a]

Table 2-7. Continued.

ACAM_0351	320118..320519	+	134	—	—	—	ORFan
ACAM_0352	320534..321550	+	339	1857	L	Uncharacterized protein predicted to be involved in DNA repair	depleted in *A. pernix*[a]
ACAM_0353	321587..322432	+	282	—	—	—	depleted in *A. pernix*[a]
ACAM_0354	322438..323658	+	407	—	—	—	ORFan
ACAM_0355	323655..325400	+	582	1203	R	Predicted helicases	paralogous gene[a]
ACAM_0356	325406..326305	+	300	—	—	—	HGT gene[a]
ACAM_0357	326320..327231	+	304	—	—	—	HGT gene
ACAM_0359	327843..328130	—	96	—	—	—	orthologous gene[a]
ACAM_0360	328127..328366	—	80	—	—	—	orthologous gene[a]
ACAM_0363	330165..330431	—	89	2026	J / D	Cytotoxic translational repressor of toxin—antitoxin stability system	HGT gene
ACAM_0364	330469..330654	—	62	—	—	—	ORFan
ACAM_0365	330937..331284	—	116	—	—	—	HGT gene
ACAM_0371	334808..335092	+	95	—	—	—	ORFan
ACAM_0373	336950..337243	—	98	—	—	—	ORFan
ACAM_0374	337240..337797	—	186	—	—	—	HGT gene
ACAM_0379	343111..343308	+	66	—	—	—	ORFan
ACAM_0441	399793..399945	—	51	—	—	—	ORFan
ACAM_0451	406523..406786	—	88	—	—	—	ORFan
ACAM_0452	407060..408100	+	347	—	—	—	HGT gene[a]

Table 2-7. Continued.

ACAM_0453	408261..408929	−	223	−	−	−	ORFan
ACAM_0454	408971..409183	−	71	−	−	−	ORFan
ACAM_0455	409272..410198	−	309	−	−	−	orthologous gene[a]
ACAM_0456	410311..410430	−	40	−	−	−	orthologous gene[a]
ACAM_0457	410558..410923	−	122	−	−	−	ORFan
ACAM_0520	481239..482126	+	296	1529	C	CoxL, Aerobic−type carbon monoxide dehydrogenase, large subunit CoxL/CutL homologs	paralogous gene[a]
ACAM_0522	483522..483728	+	69	−	−	−	ORFan
ACAM_0529	490342..490530	+	63	−	−	−	ORFan
ACAM_0572	532779..533051	−	91	−	−	−	HGT gene[a]
ACAM_0575	534961..536718	−	586	1111	L	MPH1, ERCC4−like helicases	paralogous gene[a]
ACAM_0576	536758..537639	−	294	−	−	−	HGT gene[a]
ACAM_0579	540562..541215	−	218	−	−	−	orthologous gene[a]
ACAM_0580	541179..541457	−	93	−	−	−	ORFan
ACAM_0584	544111..544785	+	225	1136	V	SalX, ABC−type antimicrobial peptide transport system, ATPase component	paralogous gene[a]
ACAM_0585	544769..547150	+	794	−	−	−	ORFan
ACAM_0587	548398..548628	+	77	−	−	−	ORFan
ACAM_0633	591211..591378	+	56	−	−	−	ORFan
ACAM_0653	609434..609757	+	108	0130	J	TruB, Pseudouridine synthase	orthologous gene[a]

Table 2-7. Continued.

ACAM_0659	613995..614510	+	172	2452	L	Predicted site−specific integrase−resolvase	depleted in *A. pernix*[a]
ACAM_0660	614503..615801	+	433	0675	L	Transposase and inactivated derivatives	depleted in *A. pernix*[a]
ACAM_0661	615754..615945	−	64	−	−	−	ORFan
ACAM_0678	632419..632919	+	167	2426	S	Predicted membrane protein	orthologous gene[a]
ACAM_0689	642979..643389	−	137	0492	O	TrxB, Thioredoxin reductase	depleted in *A. pernix*[a]
ACAM_0690	643428..643991	−	188	0492	O	TrxB, Thioredoxin reductase	depleted in *A. pernix*[a]
ACAM_0727	673617..674012	+	132	−	−	−	paralogous gene
ACAM_0740	687140..687388	−	83	−	−	−	depleted in *A. pernix*[a]
ACAM_0741	687563..688027	+	155	−	−	−	HGT gene[a]
ACAM_0742	688033..688467	+	145	−	−	−	ORFan
ACAM_0743	688487..689362	+	292	−	−	−	HGT gene
ACAM_0744	689546..690067	+	174	−	−	−	ORFan
ACAM_0745	690082..690345	−	88	−	−	−	orthologous gene[a]
ACAM_0746	690245..690454	+	70	−	−	−	ORFan
ACAM_0748	691562..692974	+	471	−	−	−	orthologous gene[a]
ACAM_0751	694249..694464	+	72	−	−	−	HGT gene[a]
ACAM_0755	698979..699689	−	237	−	−	−	orthologous gene[a]
ACAM_0756	699799..700707	−	303	−	−	−	HGT gene
ACAM_0757	700704..701867	−	388	0438	M	RfaG, Glycosyltransferase	HGT gene

Table 2-7. Continued.

ACAM_0758	701879..702736	−	286	1216	R	Predicted glycosyltransferases	ORFan
ACAM_0759	702751..704490	−	580	−	−	−	ORFan
ACAM_0760	704487..705761	−	425	−	−	−	ORFan
ACAM_0761	705758..706645	−	296	0463	M	WcaA, Glycosyltransferases	paralogous gene[a]
ACAM_0765	711798..712586	+	263	3217	R	Uncharacterized Fe−S protein	HGT gene
ACAM_0766	712718..713905	+	396	1960	I	CaiA, Acyl−CoA dehydrogenases	paralogous gene[a]
ACAM_0767	714027..714464	−	146	4113	R	Predicted nucleic acid−binding protein, contains PIN domain	depleted in *A. pernix*[a]
ACAM_0768	714440..714703	−	88	−	−	−	HGT gene[a]
ACAM_0771	720784..721011	+	76	−	−	−	ORFan
ACAM_0789	738307..738561	−	85	−	−	−	ORFan
ACAM_0790	740447..742036	−	530	−	−	−	ORFan
ACAM_0791	742748..744082	+	445	−	−	−	HGT gene[a]
ACAM_0794	746657..746902	+	82	−	−	−	ORFan
ACAM_0799	752880..753488	−	203	−	−	−	ORFan
ACAM_0800	753826..754650	−	275	−	−	−	ORFan
ACAM_0803	756179..756430	−	84	−	−	−	ORFan
ACAM_0810	761726..762301	−	192	−	−	−	ORFan
ACAM_0811	762475..762882	−	136	−	−	−	ORFan
ACAM_0840	793543..793833	−	97	−	−	−	ORFan
ACAM_0847	801373..802044	+	224	−	−	−	ORFan

Table 2-7. Continued.

ACAM_0858	818068..818265	−	66	−	−	−	ORFan
ACAM_0873	834781..835056	+	92	−	−	−	ORFan
ACAM_0930	888913..889803	−	297	2431	S	Predicted membrane protein	orthologous gene[a]
ACAM_0931	889889..890806	+	306	1808	S	Predicted membrane protein	orthologous gene[a]
ACAM_0974	930185..931981	−	599	0038	P	EriC, Chloride channel protein EriC	orthologous gene[a]
ACAM_0984	939549..941558	−	670	2217	P	ZntA, Cation transport ATPase	paralogous gene[a]
ACAM_0989	946573..947211	+	213	2020	O	STE14, Putative protein−S−isoprenylcysteine methyltransferase	paralogous gene
ACAM_1001	960187..960492	+	102	−	−	−	ORFan
ACAM_1015	974872..975036	+	55	−	−	−	ORFan
ACAM_1069	1024266..1026653	−	796	1196	D	Smc, Chromosome segregation ATPases	orthologous gene[a]
ACAM_1077	1034151..1034285	+	45	−	−	−	ORFan
ACAM_1205	1161495..1161719	−	75	−	−	−	ORFan
ACAM_1206	1163529..1163726	−	66	−	−	−	ORFan
ACAM_1232	1189995..1190174	−	60	−	−	−	orthologous gene[a]
ACAM_1243	1200203..1200421	−	73	−	−	−	orthologous gene[a]
ACAM_1252	1205950..1206189	+	80	−	−	−	ORFan
ACAM_1253	1206552..1206791	+	80	−	−	−	orthologous gene[a]
ACAM_1265	1216395..1216685	−	97	−	−	−	ORFan

Table 2-7. Continued.

ACAM_1297	1249065..1249307	−	81	−	−	−	ORFan
ACAM_1298	1249304..1249480	−	59	−	−	−	ORFan
ACAM_1330	1280791..1280994	−	68	−	−	−	ORFan
ACAM_1349	1303369..1303947	−	193	4721	S	Predicted membrane protein	depleted in *A. pernix*[a]
ACAM_1350	1304134..1305531	+	466	1123	R	ATPase components of various ABC−type transport systems, contain duplicated ATPase	orthologous gene[a]
ACAM_1366	1321004..1321186	−	61	2443	U	Preprotein translocase subunit Sss1	orthologous gene[a]
ACAM_1373	1325359..1325781	−	141	−	−	−	orthologous gene[a]
ACAM_1384	1338978..1340831	−	618	1955	N / U	FlaJ, Archaeal flagella assembly protein J	orthologous gene[a]
ACAM_1396	1352410..1352664	−	85	−	−	−	ORFan
ACAM_1398	1354681..1354941	−	87	−	−	−	ORFan
ACAM_1402	1357446..1357661	+	72	−	−	−	HGT gene[a]
ACAM_1403	1357818..1358120	+	101	−	−	−	HGT gene[a]
ACAM_1408	1361579..1361848	−	90	−	−	−	ORFan
ACAM_1437	1401904..1402104	−	67	−	−	−	ORFan
ACAM_1453	1416919..1417146	−	76	1350	R	Predicted alternative tryptophan synthase beta−subunit	paralogous gene[a]
ACAM_1459	1421436..1421744	+	103	−	−	−	ORFan
ACAM_1466	1425955..1426128	−	58	−	−	−	ORFan

Table 2-7. Continued.

ACAM_1481	1437534..1437926	−	131	1585	O / U	Membrane protein implicated in regulation of membrane protease activity	ORFan
ACAM_1508	1462011..1462601	+	197	−	−	−	HGT gene[a]
ACAM_1509	1462672..1463433	+	254	−	−	−	HGT gene[a]
ACAM_1510	1463644..1464174	+	177	−	−	−	ORFan
ACAM_1511	1464189..1465382	+	398	−	−	−	HGT gene
ACAM_1512	1465497..1466879	+	461	−	−	−	HGT gene
ACAM_1559	1514308..1514469	−	54	−	−	−	ORFan
ACAM_1561	1514750..1514914	−	55	−	−	−	ORFan
ACAM_1576	1529403..1529729	+	109	4748	S	Uncharacterized conserved protein	HGT gene[a]
ACAM_1577	1529769..1530143	+	125	−	−	−	ORFan
ACAM_1587	1538900..1539850	+	317	0667	C	Tas, Predicted oxidoreductases	paralogous gene[a]
ACAM_1591	1543729..1544607	+	293	−	−	−	depleted in *A. pernix*[a]
ACAM_1596	1548259..1548525	−	89	−	−	−	orthologous gene[a]
ACAM_1600	1553421..1553915	+	165	−	−	−	ORFan
ACAM_1605	1558770..1558961	+	64	−	−	−	ORFan
ACAM_1606	1559003..1559731	+	243	0683	E	LivK, ABC−type branched−chain amino acid transport systems, periplasmic component	HGT gene
ACAM_1607	1559650..1560015	+	122	−	−	−	HGT gene

Table 2-7. Continued.

ACAM_1608	1560017..1560232	+	72	—	—	—	ORFan
ACAM_1609	1560329..1560859	+	177	0559	E	LivH, Branched—chain amino acid ABC—type transport system, permease components	ORFan
ACAM_1610	1560790..1561251	+	154	0559	E	LivH, Branched—chain amino acid ABC—type transport system, permease components	ORFan
ACAM_1611	1561263..1562270	+	336	4177	E	LivM, ABC—type branched—chain amino acid transport system, permease component	HGT gene
ACAM_1612	1562286..1562516	+	77	0411	E	LivG, ABC—type branched—chain amino acid transport systems, ATPase component	paralogous gene[a]
ACAM_1613	1562597..1562890	+	98	0411	E	LivG, ABC—type branched—chain amino acid transport systems, ATPase component	ORFan
ACAM_1614	1562875..1563072	+	66	0411	E	LivG, ABC—type branched—chain amino acid transport systems, ATPase component	HGT gene

Table 2-7. Continued.

ACAM_1619	1567424..1567663	+	80	1853	R	Conserved protein/domain typically associated with flavoprotein oxygenases	orthologous gene[a]
ACAM_1621	1568862..1570037	+	392	2133	G	Glucose/sorbosone dehydrogenases	HGT gene
ACAM_1629	1578758..1578943	−	62	−	−	−	ORFan
ACAM_1639	1590826..1591275	−	150	1848	R	Predicted nucleic acid−binding protein, contains PIN domain	paralogous gene[a]
ACAM_1640	1591272..1591475	−	68	−	−	−	HGT gene[a]
ACAM_1643	1593242..1593472	+	77	−	−	−	HGT gene[a]
ACAM_1644	1593469..1593873	+	135	4113	R	Predicted nucleic acid−binding protein, contains PIN domain	depleted in *A. pernix*[a]
ACAM_1645	1594851..1595441	+	197	−	−	−	ORFan
APE_0001	213..938	−	241	−	−	−	ORFan
APE_0002	938..1276	−	112	1695	K	Predicted transcriptional regulators	HGT gene[a]
APE_0006.1	2270..2836	+	188	−	−	−	ORFan
APE_0024.1	16021..16419	−	132	−	−	−	HGT gene[a]
APE_0025.1	16416..16823	−	135	1378	K	Predicted transcriptional regulators	HGT gene
APE_0026	16932..18800	+	622	0574	G	PpsA, Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	HGT gene

Table 2-7. Continued.

APE_0028	18728..19384	+	218	0574	G	PpsA, Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	HGT gene
APE_0031.1	19396..20850	+	484	2814	G	AraJ, Arabinose efflux permease	HGT gene
APE_0203.1	149164..149958	−	264	0428	P	Predicted divalent heavy−metal cations transporter	HGT gene
APE_0239	173560..173886	−	108	−	−	−	orthologous gene[a]
APE_0242.1	175118..175429	+	103	−	−	−	depleted in *A. camini*[a]
APE_0264.1	190586..191218	−	210	−	−	−	ORFan
APE_0265	191705..192070	+	121	−	−	−	ORFan
APE_0267	193024..193455	+	143	2524	K	Predicted transcriptional regulator, contains C−terminal CBS domains	paralogous gene[a]
APE_0266.1	193544..195373	−	609	2414	C	Aldehyde:ferredoxin oxidoreductase	HGT gene
APE_0266a.1	196170..196424	+	84	2034	S	Predicted membrane protein	ORFan
APE_0274	199327..199677	−	116	−	−	−	HGT gene[a]
APE_0274a	199916..200131	−	71	−	−	−	HGT gene[a]
APE_0275.1	200817..201218	−	133	0122	L	AlkA, 3−methyladenine DNA glycosylase/8−oxoguanine DNA glycosylase	HGT gene
APE_0275a	201925..202152	−	75	−	−	−	HGT gene[a]
APE_0275b.1	202422..202748	−	108	−	−	−	ORFan

Table 2-7. Continued.

APE_0276.1	203152..203670	−	172	2405	R	Predicted nucleic acid−binding protein, contains PIN domain	HGT gene
APE_0276a	203657..203926	−	89	−	−	−	HGT gene
APE_0278	204280..204756	−	158	5378	R	Predicted nucleotide−binding protein	ORFan
APE_0278a	204720..204920	−	66	−	−	−	ORFan
APE_0279.1	205301..205759	−	152	4113	R	Predicted nucleic acid−binding protein, contains PIN domain	paralogous gene[a]
APE_0279a.1	205752..205952	−	66	−	−	−	paralogous gene
APE_0283.1	206858..208840	+	660	−	−	−	orthologous gene[a]
APE_0283a	209347..209469	−	40	−	−	−	paralogous gene
APE_0287	209991..210578	+	195	−	−	−	HGT gene
APE_0288	210667..211257	+	196	1846	K	MarR, Transcriptional regulators	HGT gene
APE_0290a	212077..212292	−	71	−	−	−	orthologous gene[a]
APE_0297	215139..215558	−	139	−	−	−	orthologous gene[a]
APE_0300.1	216356..217615	−	419	1123	R	ATPase components of various ABC−type transport systems, contain duplicated ATPase	paralogous gene[a]
APE_0301.1	217622..218590	−	322	0444	E/P	DppD, ABC−type dipeptide/oligopeptide/nickel transport system, ATPase component	paralogous gene[a]

Table 2-7. Continued.

APE_0302.1	218587..219492	−	301	1173	E/P	DppC, ABC−type dipeptide/oligopeptide/nickel transport systems, permease components	HGT gene
APE_0303	219506..220534	−	342	0601	E/P	DppB, ABC−type dipeptide/oligopeptide/nickel transport systems, permease components	HGT gene
APE_0304.1	220551..222950	−	799	0747	E	DdpA, ABC−type dipeptide transport system, periplasmic component	HGT gene
APE_0325	235836..236171	+	111	−	−	−	ORFan
APE_0334	240475..240819	−	114	−	−	−	orthologous gene[a]
APE_0413	284391..287510	−	1039	0553	K/L	HepA, Superfamily II DNA/RNA helicases, SNF2 family	depleted in *A. camini*[a]
APE_0414.1	287521..288069	−	182	−	−	−	ORFan
APE_0415	288073..291639	−	1188	1483	R	Predicted ATPase	depleted in *A. camini*[a]
APE_0416.1	291643..294657	−	1004	1743	L	Predicted Zn−ribbon RNA−binding protein	depleted in *A. camini*[a]
APE_0416a	295342..295518	+	58	−	−	−	ORFan
APE_0433a	304995..305171	−	58	−	−	−	depleted in *A. camini*[a]
APE_0470a	326778..327029	+	83	−	−	−	paralogous gene

Table 2-7. Continued.

APE_0471b	328673..328912	−	79	−	−	−	HGT gene[a]
APE_0471c	328922..329092	−	56	−	−	−	ORFan
APE_0472	329133..329651	−	172	5378	R	Predicted nucleotide−binding protein	HGT gene
APE_0472a	329606..329833	−	75	−	−	−	HGT gene
APE_0472c	330150..330287	+	45	−	−	−	paralogous gene
APE_0688	460283..461257	−	324	−	−	−	HGT gene
APE_0708a	474062..474271	+	69	−	−	−	ORFan
APE_0716.1	478168..479115	+	315	4342	S	Uncharacterized protein conserved in archaea	proviral gene
APE_0718	479413..479799	−	128	−	−	−	proviral gene
APE_0718a	479786..480055	−	89	−	−	−	proviral gene
APE_0720	480059..480400	−	113	−	−	−	proviral gene
APE_0720a	480407..480652	−	81	1414	K	IclR, Transcriptional regulator	proviral gene
APE_0722	480774..481433	+	219	−	−	−	proviral gene
APE_0722a	481345..481620	−	91	−	−	−	proviral gene
APE_0722b	481978..482151	−	57	−	−	−	proviral gene
APE_0722c	482257..482580	−	107	−	−	−	proviral gene
APE_0725.1	482633..484669	−	678	−	−	−	proviral gene
APE_0727	484760..485614	−	284	−	−	−	proviral gene
APE_0728	485768..486436	−	222	−	−	−	proviral gene
APE_0728a	486458..486715	−	85	−	−	−	proviral gene

Table 2-7. Continued.

APE_0728b	486804..487088	+	94	—	—	—	proviral gene
APE_0730	487176..487541	—	121	—	—	—	proviral gene
APE_0730a	487552..488331	—	259	—	—	—	proviral gene
APE_0731	488396..489571	—	391	—	—	—	proviral gene
APE_0734	489628..490233	+	201	—	—	—	proviral gene
APE_0735.1	490226..490624	+	132	—	—	—	proviral gene
APE_0736	490641..491384	+	247	—	—	—	proviral gene
APE_0737	491368..491670	+	100	—	—	—	proviral gene
APE_0745.1	494656..496827	+	723	0467	T	RAD55, RecA—superfamily ATPases implicated in signal transduction	paralogous gene[a]
APE_0760.1	502217..502624	+	135	2250	S	Uncharacterized conserved protein related to C—terminal domain of eukaryotic chaperone, SACSIN	HGT gene[a]
APE_0761.1	502674..502949	+	91	—	—	—	paralogous gene[a]
APE_0762.1	503018..503269	—	83	—	—	—	paralogous gene
APE_0816a.1	542538..542666	—	42	—	—	—	paralogous gene
APE_0818a	544129..544380	+	83	—	—	—	proviral gene
APE_0820.1	544519..544908	+	129	—	—	—	proviral gene
APE_0821	544909..545889	+	326	—	—	—	proviral gene
APE_0824	545948..546649	—	233	0501	O	HtpX, Zn—dependent protease with chaperone function	proviral gene

Table 2-7. Continued.

APE_0825.1	547079..548005	+	308	—	—	—	proviral gene
APE_0826	548559..549350	—	263	—	—	—	proviral gene
APE_0826a	549567..549731	+	54	—	—	—	proviral gene
APE_0826b	549704..549979	+	91	—	—	—	proviral gene
APE_0830	550445..551401	+	318	—	—	—	proviral gene
APE_0832.1	551405..551809	+	134	—	—	—	proviral gene
APE_0833.1	551865..552668	+	267	—	—	—	proviral gene
APE_0836	552807..553472	+	221	—	—	—	proviral gene
APE_0837	553499..554215	+	238	—	—	—	proviral gene
APE_0840	554231..555496	+	421	—	—	—	proviral gene
APE_0840a	555521..555811	+	96	—	—	—	proviral gene
APE_0843.1	555824..558604	+	926	—	—	—	proviral gene
APE_0847.1	558639..559010	+	123	—	—	—	proviral gene
APE_0848.1	559007..559300	+	97	—	—	—	proviral gene
APE_0850	559331..559648	+	105	—	—	—	proviral gene
APE_0850a	559676..559942	+	88	—	—	—	proviral gene
APE_0852.1	559993..561267	+	424	—	—	—	proviral gene
APE_0855.1	561264..561488	+	74	—	—	—	proviral gene
APE_0856	561510..562358	+	282	—	—	—	proviral gene
APE_0858	562446..564050	+	534	—	—	—	proviral gene
APE_0859	564324..564830	+	168	—	—	—	proviral gene

Table 2-7. Continued.

APE_0860	564811..565827	+	338	—	—	—	proviral gene
APE_0862.1	565904..566224	+	106	—	—	—	proviral gene
APE_0864.1	566269..566676	+	135	—	—	—	proviral gene
APE_0865.1	566712..567041	+	109	—	—	—	proviral gene
APE_0867.1	567135..568112	+	325	—	—	—	proviral gene
APE_0867a	568103..568360	+	85	—	—	—	proviral gene
APE_0867b	568627..568878	+	83	—	—	—	proviral gene
APE_0870.1	568890..569357	+	155	—	—	—	proviral gene
APE_0871.1	569321..570364	—	347	—	—	—	proviral gene
APE_0871a.1	570844..571044	+	66	—	—	—	proviral gene
APE_0872.1	571041..572450	+	469	0270	L	Dcm, Site—specific DNA methylase	proviral gene
APE_0874.1	572452..573504	—	350	—	—	—	proviral gene
APE_0875.1	573589..574260	+	223	1194	L	MutY, A/G—specific DNA glycosylase	proviral gene
APE_0878	574307..574792	+	161	—	—	—	proviral gene
APE_0879.1	574798..576231	+	477	—	—	—	proviral gene
APE_0880	576234..577838	+	534	—	—	—	proviral gene
APE_0880a	577907..578191	—	94	—	—	—	proviral gene
APE_0880b	578331..578576	+	81	—	—	—	proviral gene
APE_0880c	578583..578834	+	83	—	—	—	proviral gene
APE_0883	578834..579262	+	142	—	—	—	proviral gene
APE_0883a	579279..579509	+	76	—	—	—	proviral gene

Table 2-7. Continued.

APE_0883b	579522..579803	+	93	—	—	—	proviral gene
APE_0885.1	580035..580361	+	108	—	—	—	proviral gene
APE_0885a	580367..580612	+	81	—	—	—	proviral gene
APE_0885b	581194..581448	—	84	—	—	—	paralogous gene
APE_0954a	624975..625256	+	93	—	—	—	paralogous gene
APE_0996a	653484..653780	—	98	—	—	—	paralogous gene
APE_1041	670773..671204	+	143	2426	S	Predicted membrane protein	orthologous gene[a]
APE_1061.1	681316..682329	—	337	0492	O	TrxB, Thioredoxin reductase	HGT gene
APE_1169a	726552..726815	—	87	—	—	—	orthologous gene[a]
APE_1177.1	728049..729443	+	464	—	—	—	orthologous gene[a]
APE_1178	730149..730712	—	187	1898	M	RfbC, dTDP—4—dehydrorhamnose 3,5—epimerase and related enzymes	depleted in *A. camini*[a]
APE_1179.1	730716..731618	—	300	1091	M	RfbD, dTDP—4—dehydrorhamnose reductase	depleted in *A. camini*[a]
APE_1180	731627..732619	—	330	1088	M	RfbB, dTDP—D—glucose 4,6—dehydratase	depleted in *A. camini*[a]
APE_1181	732632..733699	—	355	1209	M	RfbA, dTDP—glucose pyrophosphorylase	paralogous gene[a]
APE_1182	734385..735953	—	522	3379	S	Uncharacterized conserved protein	paralogous gene[a]
APE_1183	736704..737633	—	309	—	—	—	HGT gene
APE_1184	738215..739288	—	357	1216	R	Predicted glycosyltransferases	depleted in *A. camini*[a]

Table 2-7. Continued.

APE_1186.1	739270..740559	−	429	−	−	−	HGT gene[a]
APE_1187.1	740556..742859	−	767	−	−	−	ORFan
APE_1188	742865..744016	−	383	0438	M	RfaG, Glycosyltransferase	HGT gene
APE_1189.1	744013..744864	−	283	0463	M	WcaA, Glycosyltransferases involved in cell wall biogenesis	HGT gene
APE_1190.1	744904..745524	−	206	−	−	−	orthologous gene[a]
APE_1191	745695..746786	−	363	0438	M	RfaG, Glycosyltransferase	HGT gene
APE_1192	746791..747663	−	290	1215	M	Glycosyltransferases, probably involved in cell wall biogenesis	HGT gene
APE_1193.1	747756..748187	−	143	−	−	−	ORFan
APE_1209	758678..759475	−	265	−	−	−	ORFan
APE_1209a	759592..759789	+	65	−	−	−	HGT gene[a]
APE_1209b	759849..760274	−	141	1848	R	Predicted nucleic acid−binding protein, contains PIN domain	HGT gene
APE_1209c.1	760255..760476	−	73	−	−	−	paralogous gene
APE_1209d	760551..760721	+	56	−	−	−	ORFan
APE_1209e	760854..761129	+	91	1960	I	CaiA, Acyl−CoA dehydrogenases	HGT gene[a]
APE_1209f	761133..761231	+	32	−	−	−	orthologous gene[a]
APE_1210.1	761487..763202	+	571	−	−	−	ORFan
APE_1211.1	763296..763670	+	124	−	−	−	ORFan

Table 2-7. Continued.

APE_1245.1	791720..792700	−	326	1063	E/R	Tdh, Threonine dehydrogenase and related Zn−dependent dehydrogenases	HGT gene
APE_1275	806006..807409	+	467	3033	E	TnaA, Tryptophanase	HGT gene
APE_1275a	807504..807785	+	93	−	−	−	HGT gene[a]
APE_1275b	807668..807964	+	98	−	−	−	paralogous gene[a]
APE_1275c	808068..808184	+	38	−	−	−	paralogous gene[a]
APE_1276	808208..809323	+	371	−	−	−	paralogous gene
APE_1277	809320..809886	−	188	−	−	−	ORFan
APE_1278	809861..810268	−	135	−	−	−	orthologous gene[a]
APE_1339.1	848760..849434	+	224	−	−	−	orthologous gene[a]
APE_1343a.1	854634..854909	+	91	−	−	−	HGT gene[a]
APE_1408	896868..897626	−	252	−	−	−	ORFan
APE_1409.1	897633..898628	−	331	−	−	−	ORFan
APE_1409a	898978..899073	−	31	−	−	−	ORFan
APE_1473a	938145..938453	−	102	−	−	−	HGT gene
APE_1477	938555..939481	−	308	2431	S	Predicted membrane protein	orthologous gene[a]
APE_1478.1	939776..940441	+	221	1808	S	Predicted membrane protein	orthologous gene[a]
APE_1552	979949..980908	−	319	−	−	−	orthologous gene[a]
APE_1555.1	980764..981771	−	335	0038	P	EriC, Chloride channel protein EriC	orthologous gene[a]
APE_1558	983868..984707	+	279	2810	V	Predicted type IV restriction endonuclease	HGT gene

Table 2-7. Continued.

APE_1558a	984884..985168	+	94	—	—	—	ORFan
APE_1558b	985215..985499	+	94	2026	J/D	RelE, Cytotoxic translational repressor of toxin—antitoxin stability system	HGT gene[a]
APE_1558c	985687..985860	—	57	—	—	—	ORFan
APE_1574.1	996232..996798	+	188	—	—	—	ORFan
APE_1586a	1007461..1007607	—	48	—	—	—	ORFan
APE_1586b	1007647..1007754	—	35	—	—	—	paralogous gene[a]
APE_1588	1007789..1008187	—	132	5573	R	Predicted nucleic—acid—binding protein, contains PIN domain	HGT gene
APE_1588a	1008527..1008760	+	77	0574	G	PpsA, Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	paralogous gene[a]
APE_1588b	1008781..1008873	+	30	—	—	—	paralogous gene[a]
APE_1594a	1011725..1012030	+	101	—	—	—	ORFan
APE_1708	1075942..1078317	—	791	1196	D	Smc, Chromosome segregation ATPases	orthologous gene[a]
APE_1804	1135577..1136833	—	418	—	—	—	paralogous gene
APE_1882a	1194127..1194252	+	41	—	—	—	HGT gene[a]
APE_1907	1208353..1209093	—	246	1681	N	FlaB, Archaeal flagellins	paralogous gene[a]
APE_1921	1215480..1215983	—	167	—	—	—	HGT gene

Table 2-7. Continued.

APE_1929.1	1219285..1219953	−	222	3780	L	DNA endonuclease related to intein−encoded endonucleases	HGT gene
APE_1979a	1253505..1253723	−	72	−	−	−	orthologous gene[a]
APE_1995.1	1259549..1260130	+	193	−	−	−	orthologous gene[a]
APE_2029.1	1278941..1280023	+	360	2038	H	CobT, NaMN:DMB phosphoribosyltransferase	depleted in *A. camini*[a]
APE_2032.1	1280039..1281112	+	357	1865	S	Uncharacterized conserved protein	depleted in *A. camini*[a]
APE_2034.1	1281109..1281660	+	183	2266	H	GTP:adenosylcobinamide−phosphate guanylyltransferase	depleted in *A. camini*[a]
APE_2035.1	1281627..1282691	+	354	0079	E	HisC, Histidinol−phosphate/aromatic aminotransferase and cobyric acid decarboxylase	depleted in *A. camini*[a]
APE_2037.1	1282673..1283458	+	261	0368	H	CobS, Cobalamin−5−phosphate synthase	HGT gene[a]
APE_2039.1	1283451..1284413	+	320	1270	H	CbiB, Cobalamin biosynthesis protein CobD/CbiB	depleted in *A. camini*[a]
APE_2041.1	1284410..1285366	+	318	0367	E	AsnB, Asparagine synthase	HGT gene
APE_2042.1	1285370..1286068	+	232	2102	R	Predicted ATPases of PP−loop superfamily	HGT gene
APE_2065.1	1300336..1300590	−	84	−	−	−	paralogous gene

Table 2-7. Continued.

APE_2154b	1365604..1365837	+	77	1122	P	CbiO, ABC−type cobalt transport system, ATPase component	orthologous gene[a]
APE_2164.1	1374097..1374468	+	123	0003	D	ArsA, Predicted ATPase involved in chromosome partitioning	HGT gene[a]
APE_2176a	1380995..1381177	−	60	2443	U	Sss1, Preprotein translocase subunit Sss1	orthologous gene[a]
APE_2185.1	1385358..1385795	−	145	−	−	−	orthologous gene[a]
APE_2206.1	1399273..1400127	−	284	−	−	−	orthologous gene[a]
APE_2207.1	1400130..1401128	−	332	1955	N/U	FlaJ, Archaeal flagella assembly protein J	orthologous gene[a]
APE_2239.1	1418407..1419441	−	344	1064	R	AdhP, Zn−dependent alcohol dehydrogenases	paralogous gene[a]
APE_2240	1419723..1420697	+	324	2159	R	Predicted metal−dependent hydrolase of the TIM−barrel fold	HGT gene
APE_2242.1	1420761..1421600	+	279	−	−	−	HGT gene
APE_2242b	1422029..1422199	−	56	−	−	−	paralogous gene[a]
APE_2256.1	1430398..1431318	−	306	4006	S	Uncharacterized protein conserved in archaea	depleted in *A. camini*[a]
APE_2265a	1438193..1438303	+	36	−	−	−	orthologous gene[a]
APE_2284a	1459592..1459828	−	78	3350	S	Uncharacterized conserved protein	depleted in *A. camini*[a]
APE_2326.1	1487204..1487512	+	102	−	−	−	ORFan

Table 2-7. Continued.

APE_2356.1	1503290..1503682	−	130	1585	O/U	Membrane protein implicated in regulation of membrane protease activity	ORFan
APE_2380.1	1514215..1514454	−	79	2031	I	AtoE, Short chain fatty acids transporter	HGT gene
APE_2480a	1575939..1576145	−	68	2888	J	Predicted Zn−ribbon RNA−binding protein with a function in translation	depleted in *A. camini*[a]
APE_2520.1	1599132..1599845	−	237	3473	Q	Maleate cis−trans isomerase	HGT gene
APE_2521.1	1599950..1601233	+	427	0683	E	LivK, ABC−type branched−chain amino acid transport systems, periplasmic component	HGT gene
APE_2522.1	1601330..1602199	+	289	0559	E	LivH, Branched−chain amino acid ABC−type transport system, permease components	HGT gene
APE_2523.1	1602265..1603194	+	309	4177	E	LivM, ABC−type branched−chain amino acid transport system, permease component	HGT gene
APE_2524.1	1603191..1603910	+	239	0411	E	LivG, ABC−type branched−chain amino acid transport systems, ATPase component	HGT gene

Table 2-7. Continued.

APE_2528.1	1604644..1606692	+	682	0145	E/Q	HyuA, N−methylhydantoinase A/acetone carboxylase, beta subunit	paralogous gene[a]
APE_2530.1	1606695..1608353	+	552	0146	E/Q	HyuB, N−methylhydantoinase B/acetone carboxylase, alpha subunit	paralogous gene[a]
APE_2538	1613261..1614940	−	559	1757	C	NhaC, Na^+/H^+ antiporter	HGT gene[a]
APE_2567	1630773..1631426	−	217	−	−	−	ORFan
APE_2577.1	1637549..1638424	+	291	1533	L	SplB, DNA repair photolyase	HGT gene
APE_2580	1640463..1641254	+	263	1853	R	Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family	orthologous gene[a]
APE_2581	1641439..1642656	+	405	0095	H	LplA, Lipoate−protein ligase A	HGT gene
APE_2583.1	1642673..1643947	−	424	1301	C	GltP, Na^+/H^+−dicarboxylate symporters	depleted in *A. camini*[a]
APE_2604a.1	1657630..1658052	−	140	0365	I	Acs, Acyl−coenzyme A synthetases/AMP−(fatty) acid ligases	HGT gene
APE_2616	1665993..1666418	−	141	1848	R	Predicted nucleic acid−binding protein, contains PIN domain	paralogous gene
APE_2616a	1666446..1666640	−	64	−	−	−	paralogous gene
APE_2617.1	1666868..1667311	−	147	−	−	−	HGT gene
APE_2617a	1667286..1667546	−	86	−	−	−	HGT gene[a]

Table 2-7. Continued.

APE_2617b	1667914..1668156	−	80	1848	R	Predicted nucleic acid−binding protein, contains PIN domain	HGT gene[a]
APE_2617c	1668321..1668539	−	72	−	−	−	paralogous gene
APE_2617d	1668839..1669138	−	99	−	−	−	HGT gene[a]
APE_2617e.1	1669179..1669421	−	80	2442	S	Uncharacterized conserved protein	HGT gene[a]

[a] Identified by inspecting the distribution of homologs in crenarchaeal genomes.

Surveys for viral metagenomes suggest that the diversity of viral sequences is vast and remains largely unexplored (Edwards and Rohwer 2005). Therefore, it seems plausible that a major fraction of archaeal and bacterial ORFans are derived from the poorly explored but vast viral gene pool, although it is impossible to rule out that ORFans have homologs in multiple genomes that avoid detection because of their rapid evolution (Koonin and Wolf 2008). ORFans probably derived from viruses and proviral genes accounted for 41 to 45% of nonorthologous genes. Two viruses infecting *A. pernix* were isolated from environmental samples collected at the coastal Yamagawa hot spring in Ibusuki, Japan: a dsDNA virus, *Aeropyrum pernix* bacilliform virus 1 (APBV1) (Mochizuki et al. 2010), and the single-stranded *Aeropyrum* coil-shaped virus (ACV) (Mochizuki et al. 2012). *A. camini* could not be infected by ACV (Mochizuki et al. 2012), and its susceptibility to infection by APBV1 was not tested (Mochizuki et al. 2010). Morphologically diverse virus-like particles were also observed at the Yamagawa hot spring (Mochizuki et al. 2010). This analysis showed that most CRISPR spacers in *A. camini* and *A. pernix* lacked similarity to any other nucleotide sequences in the database. These data indicated that *Aeropyrum* spp. were challenged by diverse and uncharacterized viruses.

The variable gene component is responsible for expanding physiological and ecological capabilities of microorganisms (Gogarten and Townsend 2005), most notably antibiotic resistance among bacterial pathogens (Dobrindt et al. 2004). Although the variable genes

in *Aeropyrum* spp. were mostly derived from viruses with unknown functions, they are potentially responsible for the acquisition of new functions. If every unique microbial strain contains a different set of variable genes, the size of the pan-genome (the total set of genes found in all strains) (Medini et al. 2005) becomes vast in a large microbial population. Each *Aeropyrum* spp. strain appears to share conserved small genomes encoding genes required for cell maintenance and, at the same time the *Aeropyrum* population's pan-genome may be extended by viruses to give a significant genetic reservoir exploited for adaptive purposes, resulting in the increased fitness of the population in changeable extreme environments.

Here I show that *Aeropyrum* spp. may be specialized in aerobic and thermophilic environments and accordingly possess small and conservative genomes; nevertheless, *Aeropyrum* spp. interact with diverse viruses, and their genomic diversification is substantially caused by viruses.

Chapter 3

Comparative genomic analysis of the thermophilic crenarchaea

Introduction

Microorganisms are classified into two categories, generalist or specialist, depending on their life cycle strategies (Newton et al. 2010). Generalists inhabit in broad ecological niches, employ flexible mechanisms for energy and carbon acquisition, and possess large genomes encoding variable metabolic genes or transcriptional regulators (Newton et al. 2010; Koonin and Wolf 2011). Meanwhile, specialists survive in relatively constant environments, utilize a restricted sort of resources, and possess small genomes encoding only essential genes (Parter et al. 2007; Koonin and Wolf 2011). The small genomes of the specialists are believed to be achieved by genome streamlining (Giovannoni et al. 2005). Genome streamlining is the theory that they are expected to minimize genomes due to the unnecessary cost of replicating DNA with no adaptive value (Giovannoni et al. 2014). For example, abundant marine alpha-proteobacterium *Pelagibacter ubique* owns the smallest genome among free-living organisms and is likely to lose non-essential genes (Giovannoni et al. 2005).

In chapter 2, despite their geographically and environmentally different habitat, streamlined genomes of *A. camini* and *A. pernix* share a large proportion of orthologous genes and show high synteny (Daifuku et al. 2013). Their genomic variation is, however, observed at virus-related

elements like proviral regions, defense system against foreign genetic elements (FGEs) (i.e. CRISPR), and ORFans probably derived from viruses (Daifuku et al. 2013). Although they own streamlined syntenic genomes, their genomic diversification is substantially exerted by viruses (Daifuku et al. 2013).

The members of the genus *Aeropyrum* belong to the archaeal phylum crenarchaeota. The crenarchaeota comprises of thermophilic and hyperthermophilic organisms. They inhabit solfataric hot spring or marine hydrothermal vent (Garrity and Holt 2001). The majority of them grow optimally at temperature > 80°C (Garrity and Holt 2001). In chapter 3, I performed a comparative genomic analysis of closely related crenarchaea to inspect the hypothesis that crenarchaea as well as *Aeropyrum* spp. are specialized in their own habitat with the small and conservative genomes; nevertheless their genomic diversification is driven by FGEs including viruses.

Materials and Methods

Comparative genomics

Fully sequenced crenarchaeal genomes were downloaded from the RefSeq database (10 July 2014) (Pruitt et al. 2007). Data for isolation sites were obtained from the Integrated Microbial Genomes database (Markowitz et al. 2008). Putative orthologous genes were identified as RBHs among all

pairs of genomes in the same genus by using BLASTP (Altschul et al. 1997) with a coverage threshold of 50% for both gene sequences and an E value threshold of 10^{-6} at an effective database size of 10^{7}. GSS was calculated as described in chapter 2. To evaluate the conservation of genome structure, Synteny Index (SI) between two genomes was calculated according to the method described elsewhere (Novichkov et al. 2009). Briefly, for each orthologous gene pair between two genomes, the maximum number of orthologous gene pairs was counted in a sliding window of seven genes. The orthologous gene pair with more than five orthologous gene pairs in the flanking regions were determined to be located on the synteny region. SI was calculated as follows: SI = (Nb - Ns)/Nb, where Nb is the total number of RBHs and Ns is the number of RBHs on the synteny region. SI ranges from 0 to 1 and decrease with the syntenic disruptions. Genes potentially derived from viruses are identified by searching nonorthologous genes against RefSeq-viral database (Pruitt et al. 2007) by using BLASTP with an E value threshold of 10^{-6} at an effective database size of 10^{7}. ORFans were identified as described in chapter 2.

Bioinformatic tools

Synteny plots were generated as alignments of the nucleotide sequences by using MUMMER 3.0 (Kurtz et al. 2004). IS elements were identified by using the ISfinder database (Siguier et al. 2006) as described in chapter 2. Geographic distances between isolation sites were calculated at the Geospatial Information Authority of Japan web site

(http://vldb.gsi.go.jp/sokuchi/surveycalc/surveycalc/bl2stf.html) by using the Geodetic Reference System 1980 (GRS80).

Results and Discussion

Phylogenetic distance and genomic synteny

GSSs and SIs were calculated between genomes in the same genus (240 pairs of genomes). GSSs ranged from 0.60 to 1.0 and SIs ranged from 0 to 0.40 (Table 3-1). For example, the following three distinguishable patterns are shown (Fig. 3-1): (A) high synteny with rare transposition of individual genes between *A. pernix* and *A. camini* (GSS = 0.87 and SI = 0.015); (B) synteny within limited regions between *Pyrobaculum neutrophilus* and *Pyrobaculum* sp. 1860 (GSS = 0.73 and SI = 0.18); (C) extensive decay of synteny between *Sulfolobus solfataricus* 98/2 and *Sulfolobus tokodaii* (GSS = 0.61 and SI = 0.39). I found a statistically significant negative correlation between GSSs and SIs ($R^2 = 0.96$, $P < 0.01$, Fig. 3-2). The majority of the data points (229 pairs of GSS and SI) fell inside the 95% prediction interval; the compared data between the members of the genus *Sulfolobus* (209 pairs), *Pyrobaculum* (18 pairs), *Desulfurococcus* (1 pair), and *Thermoproteus* (1 pair). The minority of the data points (11 pairs of GSS and SI), however, fell outside the prediction interval. Of these, one data point, which is attributed to the comparison between *S. solfataricus* P2 and *S. tokodaii*, fell up outside the prediction interval. Another ten data points fell down outside the

prediction interval; the compared data between the members of the genus *Aeropyrum* (1 pair), *Staphylothermus* (1 pair), *Desulfurococcus* (2 pairs), *Metallosphaera* (1 pair), *Pyrobaculum* (3 pairs), *Thermofilum* (1 pair), and *Vulcanisaeta* (1 pair).

Table 3-1. Characteristics of the genomes used in this study.

Genus	No. of genomes	Genome size (Mbp) (min/median/max)	No. of proteins (min/median/max)	No. of IS elements (min/median/max)	No. of orthologous genes (min/median/max)	GSS (min/median/max)	SI (min/median/max)
Desulfurococcales							
Aeropyrum	2	1.6/1.6/1.7	1645/1673/1700	0/1/2	1455/1455/1455	0.87/0.87/0.87	0.02/0.02/0.02
Desulfurococcus	3	1.3/1.4/1.4	1345/1421/1471	2/8/20	1149/1171/1260	0.74/0.74/0.96	0.01/0.08/0.09
Staphylothermus	2	1.6/1.6/1.6	1573/1586/1599	7/9/10	1317/1317/1317	0.90/0.90/0.90	0.03/0.03/0.03
Sulfolobales							
Sulfolobus	21	2.1/2.6/3.0	2146/2608/2978	8/77/378	1587/2121/2432	0.60/0.92/1.00	0/0.07/0.40
Metallosphaera	2	1.8/2.0/2.2	2029/2143/2256	1/4/7	1748/1748/1748	0.76/0.76/0.76	0.06/0.06/0.06
Thermoproteales							
Pyrobaculum	7	1.8/2.1/2.5	1966/2299/2835	0/2/8	1513/1637/2012	0.71/0.74/0.95	0.04/0.20/0.30
Thermofilum	2	1.8/1.8/1.8	1878/1887/1896	2/3/3	1389/1389/1389	0.66/0.66/0.66	0.17/0.17/0.17
Thermoproteus	2	1.8/1.9/1.9	2049/2119/2189	1/2/3	1643/1643/1643	0.74/0.74/0.74	0.19/0.19/0.19
Vulcanisaeta	2	2.3/2.3/2.4	2320/2407/2493	1/5/8	1922/1922/1922	0.84/0.84/0.84	0.06/0.06/0.06

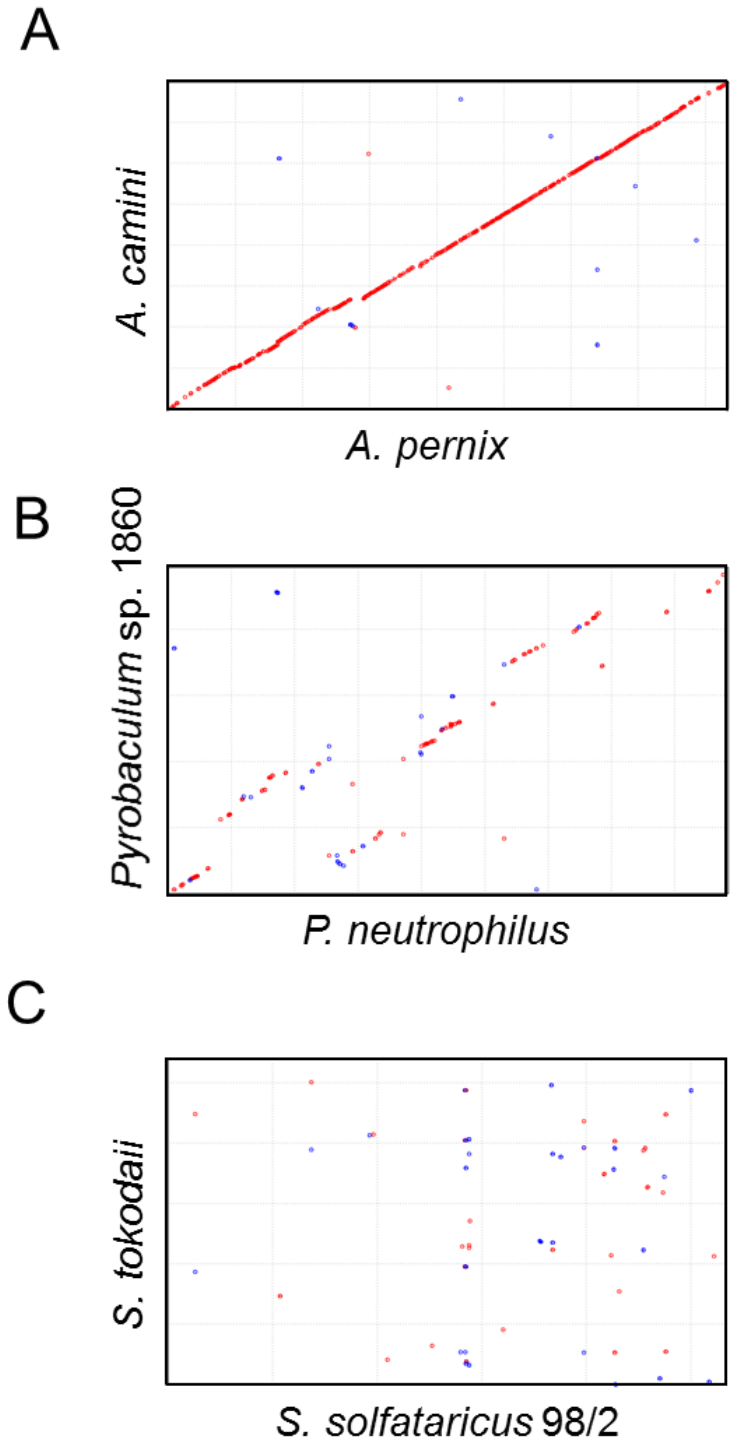

Figure 3-1. Comparison of the chromosomes of crenarchaea. MUMMER nucleotide alignment, where dots indicate similar sequences in the same orientation (red) or reverse orientation (blue), shared by the two species. (A) *A. pernix* and *A. camini* (GSS = 0.87 and SI = 0.015). (B) *P. neutrophilus* and *Pyrobaculum* sp. 1860 (GSS = 0.73 and SI = 0.18). (C) *S. solfataricus* 98/2 and *S. tokodaii* (GSS = 0.61 and SI = 0.39).

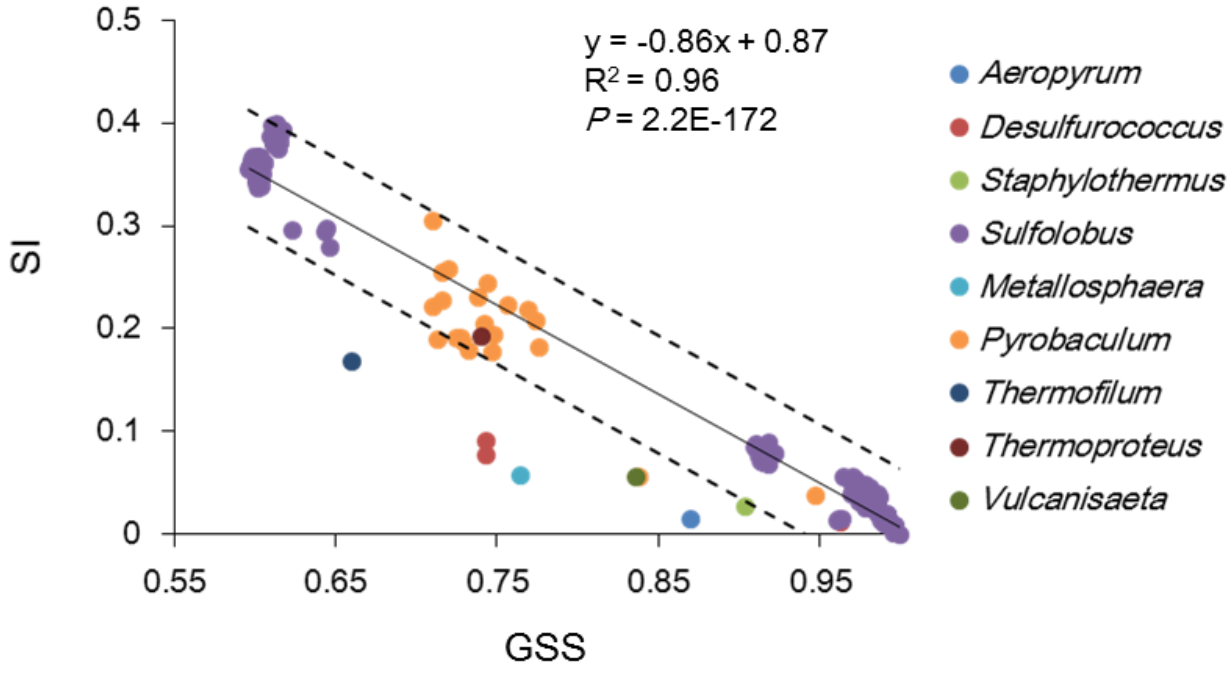

Figure 3-2. Relationship between GSSs and SIs. The equation for the linear regression trend line (y = ax + b), the coefficient of determination (R^2), and the level of significance for the correlation (*P*) are shown. The linear regression trend line and the 95% prediction interval are shown in solid line and dashed lines, respectively. The association between GSSs and SIs is significant (*P* = 2.2E-172).

The conservation of genomic synteny of archaea and bacteria generally decreases with increasing phylogenetic distance (Yelton et al. 2011). Ten pairs of crenarchaea including genus *Aeropyrum*, however, revealed highly syntenic genomes regardless of their phylogenetic distance (Fig. 3-2). One of the factors associated with the synteny disruptions is the abundance of IS (Filée et al. 2007). ISs carry genes encoding the enzymes, the transposases, that catalyze their movement and are generally flanked by terminal inverted repeats (Filée et al. 2007). Multiple copies of IS on a genome can be the basis of homologous recombination (Brügger et al. 2004). The number of IS ranged from 0 to 378 (Table 3-1). *Sulfolobus* spp. (excluding *S. acidocaldarius* strains) possess a large number of ISs (22 to 378). Their genomic synteny collapsed according to their phylogenetic distance (Fig. 3-2). In contrast, crenarchaea excluding *Sulfolobus* spp. possess a small number of IS (equal to or less than 20). Their genomic synteny varied depending on the compared species. For example, the genomic synteny was conserved regardless of the phylogenetic distance between *A. camini* and *A. pernix*, but not between *P. aerophilum* and *P. arsenaticum*. The abundance of IS did not fully explain the degree of the genomic synteny, although the genomic integrity may be partly due to the small amount of IS.

Another factor of the genomic integrity may be the selective constraint whether synteny disruptions are allowed or not. The synteny disruptions are involved in the alteration of the transcriptional architecture (Yoon et al. 2011). The transcriptional changes can be lethal in natural environment and may not be allowed for the crenarchaea which is restricted

to a narrow range of habitat. In contrast, genomic rearrangements may be allowed for the crenarchaea which adapt to variable habitats (Fig. 3-2).

Next, I calculated Conservation Degree (CD) as the distance between each data point and the regression line in Fig. 3-2 in order to measure the degree of syntenic conservation considering the phylogenetic distance. The distance for the data point up and down the regression line were represented by plus and minus, respectively. The CDs decreased with the increasing degree of genomic synteny considering the phylogenetic distance. The CDs ranged from -0.12 to 0.044 for all the data points (Fig. 3-3). Specifically, CDs ranged from -0.12 to -0.044 and -0.041 to 0.044 for the ten data points including *Aeropyrum* spp. observed down outside the prediction interval as described above and for the others, respectively. A statistically significant positive correlation was observed between CDs and the average genome size of the compared pair of genomes ($R^2 = 0.50$, $P < 0.01$, Fig. 3-3) with an exception of the compared data between *D. fermentans* and *D. kamchatkensis* (CD = -0.021 and the average genome size = 1.4 Mbp).

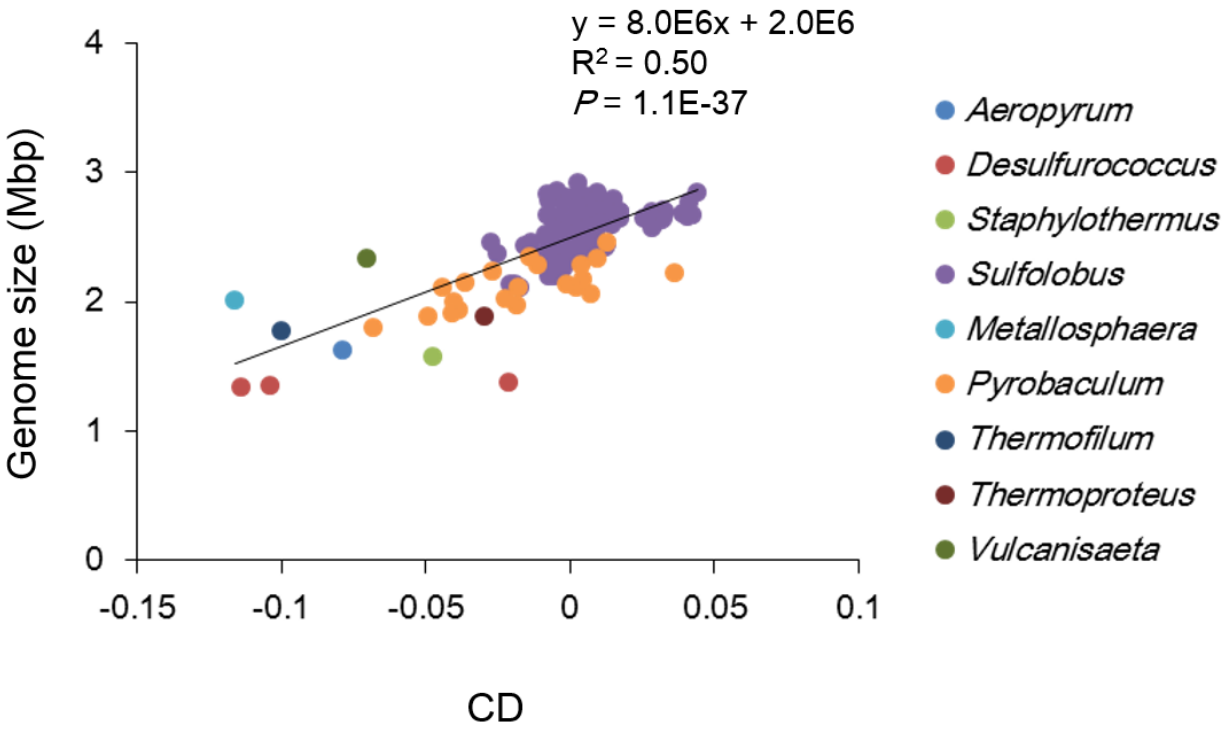

Figure 3-3. Relationship between CDs and average genome size of a compared pair of genomes. The equation for the linear regression trend line (y = ax + b), R^2, and *P* are shown. The linear regression trend line is shown in solid line. CD is positively associated with genome size (*P* = 1.1E-37).

Phylogenetic divergence and geographic distance

Phylogenetic divergence is promoted by the geographic distance for the specialists which does not expand their habitat area easily (Whitaker et al. 2003; Reno et al. 2009). I investigated the relationship between GSSs (phylogenetic distance) and the distance of isolation sites among different species. Both values were statistically correlated in crenarchaea including *Aeropyrum* spp. observed down outside the prediction interval in Fig. 3-2 (R^2 = 0.76, P < 0.01, Fig. 3-4A), suggesting that these crenarchaea are highly restricted in their own habitat as specialists. Meanwhile, I found no significant correlation between GSSs and the distance for the other crenarchaea (R^2 = 0.0014, P > 0.01, Fig. 3-4B), suggesting that the crenarchaea are likely to adapt to relatively variable environment as generalists and disperse easily.

Analysis of protein-coding sequences

The ratio of the orthologous genes between genomes in all encoded genes ranged from 0.62 to 0.87. I found a statistically significant positive correlation between CDs and the ratio of the orthologous genes (R^2 = 0.39, P < 0.01, Fig. 3-5A), indicating that the ancestors of specialists share essential genes. I searched genes derived from viruses among the nonorthologous genes between genomes. The ratio of the genes significantly matched with genes in the RefSeq-viral database in the nonorthologous genes ranged from 0.027 to 0.15. CDs positively correlated with the ratio of the genes matched with RefSeq-viral genes (R^2 = 0.084, P < 0.01, Fig. 3-5B). Viral metagenome

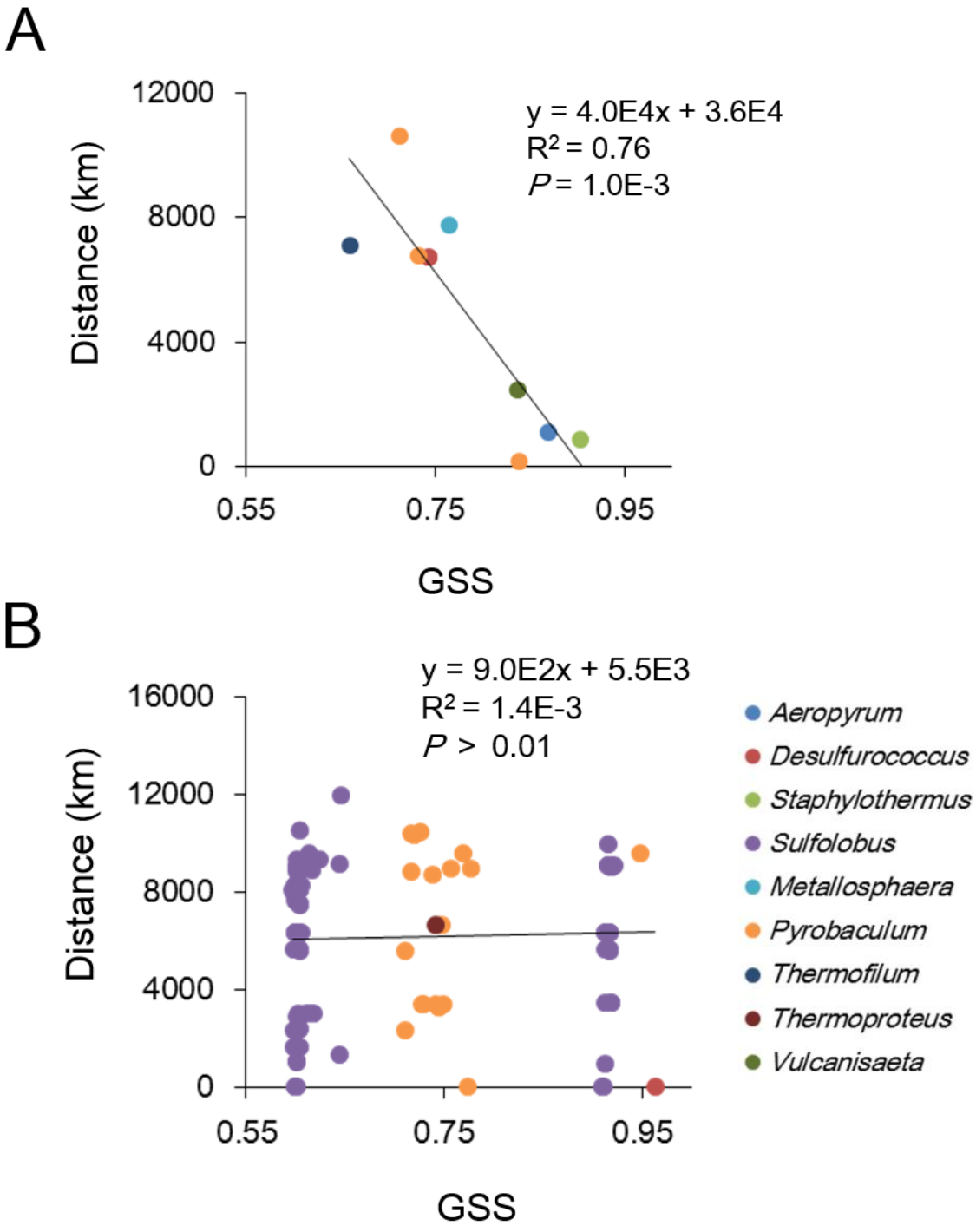

Figure 3-4. Relationship between GSS and distance between isolation sites among species in the same genus. The equation for the linear regression trend line (y = ax + b), R^2, and *P* are shown. The linear regression trend line is shown in solid line. (A) Crenarchaea with conservative genomes including *Aeropyrum* spp. The association between GSS and distance is significant (*P* = 1.0E-3). (B) The other crenarchaea. The association between GSS and distance is not significant (*P* > 0.01).

analyses suggest that the diversity of viral sequences remains largely unexplored (Edwards and Rohwer 2005). Therefore, it appears plausible that a major fraction of archaeal and bacterial ORFans are derived from the poorly explored gene pool (the viral metagenome), although it is impossible to rule out that ORFans have homologs in multiple genomes that avoid detection because of their rapid evolution (Koonin and Wolf 2008). The ratio of the ORFans in the nonorthologous genes ranged from 0.02 to 0.44. CDs negatively correlated with the ratio of the ORFans ($R^2 = 0.20$, $P < 0.01$, Fig. 3-5C).

Aeropyrum spp. are specialists and interact with distinct population of viruses, and their genomic diversification is considerably caused by viruses (Daifuku et al. 2013). The genomic diversification of the crenarchaeal specialists may also be affected by viruses, in spite of their syntenic genomes and their similar gene repertory (Fig. 3-5A). On the other hand, the crenarchaea generalists which possess plastic and relatively large genomes diversify not only by viruses, but also by internal factors such as genomic rearrangements (Fig. 3-2).

Prokaryotes are exposed to FGEs like extracellular membrane vesicles (MVs) and viruses in natural environment. MVs are produced by all three domains of life and contain proteins, DNA, and RNA (Gaudin et al. 2014). Hosts have variety of defense mechanisms like CRISPR elements and toxin-antitoxin systems against FGEs (Makarova et al. 2011b). Viruses develop counter-defense systems (Labrie et al. 2010), while MVs does not. While the generalist genomes of geographically distinct strains of *S.*

acidocaldarius is nearly identical and there may be no geographic barriers between the local populations (Mao and Grogan 2012), the coevolutionary interactions between host and virus may lead to locality of the community (Brockhurst et al. 2007; Kunin et al. 2008; Koskella et al. 2011). Even the crenarchaeal generalists may be restricted to a narrow range of habitat through adaptation to local virus population and transform into specialists.

In conclusion, some crenarchaea (e.g. *Aeropyrum*) possess conservative and small genomes and specialize in their habitat and the other crenarchaea (e.g. *Sulfolobus*) possess large genomes with extensive genomic rearrangements and can adapt to variable habitats. Regardless of the conservative genomes of specialists, their diversification is partly maintained by viruses.

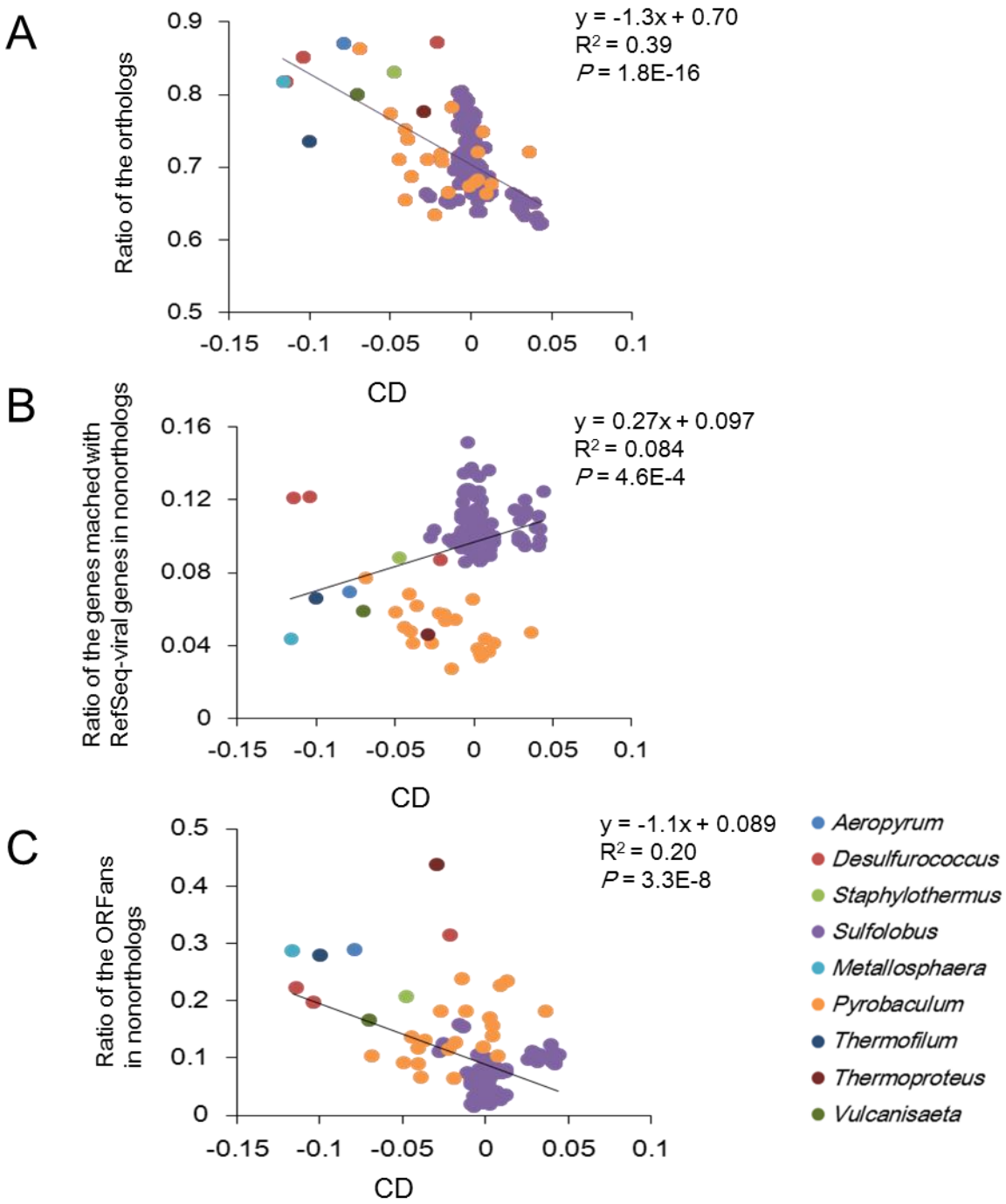

Figure 3-5. CDs are plotted against (A) the ratio of the orthologous genes in all genes, (B) the ratio of the genes mached with RefSeq-viral genes in nonorthologous genes, and (C) the ratio of the ORFans in nonorthologous genes. The equation for the linear regression trend line (y = ax + b), R^2, and *P* are shown. The linear regression trend line is shown in solid line. CDs are negatively correlated with both the ratio of the orthologous genes and the ratio of the ORFans (*P* = 1.8E-16 and *P* = 3.3E-8, respectively). CDs are positively correlated with the ratio of the genes matched with RefSeq-viral genes (*P* = 4.6E-4).

Chapter 4

General overview

The increasing number of genome sequences of archaea and bacteria leads to show their adaptation to different environmental conditions at the genomic level. *Aeropyrum* spp. are aerobic and hyperthermophilic archaea. *A. camini* was isolated from a deep-sea hydrothermal vent, and *A. pernix* was isolated from a coastal solfataric vent. In chapter 2, I compared the genomes of the two species to investigate the adaptation strategy in each habitat. Their shared genome features were a small genome size, a high GC content, and a large portion of orthologous genes (86 to 88%). The genomes also showed high synteny. These shared features may have been derived from the small number of mobile genetic elements and the lack of a RecBCD system, a recombinational enzyme complex. In addition, the specialized physiology (aerobic and hyperthermophilic) of *Aeropyrum* spp. may also contribute to the entire-genome similarity. Despite having stable genomes, interference of synteny occurred with two proviruses, *A. pernix* spindle-shaped virus 1 (APSV1) and *A. pernix* ovoid virus 1 (APOV1), and clustered regularly interspaced short palindromic repeat (CRISPR) elements. CRISPR spacer sequences observed in the *A. camini* showed significant matches with protospacers of the two proviruses found in the genome of *A. pernix*, indicating that *A. camini* interacted with viruses closely related to APSV1 and APOV1. Furthermore, a significant fraction of the

nonorthologous genes (41 to 45%) were proviral genes or ORFans probably originating from viruses. Although the genomes of *A. camini* and *A. pernix* were conserved, I observed nonsynteny regions that were attributed primarily to virus-related elements. These findings indicated that the genomic diversification of *Aeropyrum* spp. is substantially caused by viruses.

The archaeal phylum crenarchaeota is composed of thermophilic or hyperthermophilic organisms. I hypothesized that although crenarchaea as well as *Aeropyrum* spp. interact with distinct community of viruses and their genomic diversification is caused by viruses, although they are highly specialized in narrow range of habitat and possess streamlined genomes. In chapter 3, to test this hypothesis, I performed a comprehensive comparative genomic analysis of crenarchaea (240 pairs of crenarchaea). Genomic sysnteny depended on phylogenetic distance. Crenarchaea including marine hyperthermophilic *Aeropyrum* spp. showed high genomic synteny regardless of phylogenetic distance. The degree of genomic conservation correlated with genome size. The crenarchaea with less synteny disruptions and small genomes (1.4 - 2.3 Mbp) are likely to be isolated by geographic distance, implying that the ancestors of the crenarchaea are highly specialized in their own habitat. On the other hand, the other crenarchaea with plastic and large genomes are likely to be cosmopolitan as generalists. Although the specialists shared a higher ratio of ortholgous, some of their nonorthologous genes were probably derived from viruses. These findings suggested that genomic diversification of the specialists was partly promoted by viruses in spite of their small and conservative genomes.

Acknowledgements

This thesis could not have been completed without help of many people. First of all, I express my sincere gratitude to my supervisor, Professor Yoshihiko Sako, who introduced me into the exciting world of thermophiles and encouraged me throughout the past six years in the laboratory of marine microbiology. I am deeply grateful to associate professor Takashi Yoshida who gave me critical and positive advice. I am also grateful to Professor Shigeki Sawayama for reviewing my thesis.

I thank my colleagues in this laboratory for their support and discussion. Especially, I offer special thanks to Mr. Takayuki Kitamura, Mr. Shin Fujiwara, and Mr. Kimiho Omae for valuable discussion and technical help.

I was supported by a Grant-in-Aid for JSPS Fellows from the Ministry of Education, Culture, Sports, Science, and Technology.

On a personal note, I thank my parents and wife for their support.

References

Alcaraz LD, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L, Olmedo G (2010) Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. BMC Genomics **11:**332.

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

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-3402.

Anderson RE, Brazelton WJ, Baross JA (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol **77:**120-133.

Auch AF, von Jan M, Klenk H, Göker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci **2:**117-134.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709-1712.

Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes Annul Rev Genet **38:**771-791.

Blackwood JK, Rzechorzek NJ, Bray SM, Maman JD, Pellegrini L, Robinson NP (2013) End-resection at DNA double-strand breaks in the three domains of life. Biochem Soc Trans **41:**314-320.

Brockhurst MA, Morgan AD, Fenton A, Buckling A (2007) Experimental coevolution with bacteria and phage. The *Pseudomonas fluorescens* -Φ2 model system. Infect Genet Evol **7:**547-552.

Brügger K, Chen L, Stark M, Zibat A, Redder P, Reupp A, Awayez M, She Q, Garrett RA, Klenk HP (2007) The genome of *Hyperthermus butylicus*: a

sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 ° C. Archaea **2:**127-135.

Brügger K, Torarinsson E, Redder P, Chen L, Garrett RA (2004) Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements. Biochem Soc Trans **32:**179-183.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ (2012) Patterns of gene flow define species of thermophilic archaea. PLoS Biol **10:**e1001265.

Cohan FM, Koeppel AF (2008) The origins of ecological diversity in prokaryotes. Curr Biol **18:**R1024-R1034.

Csűrös M, Miklós I (2009) Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol **26:**2087-2095.

Daifuku T, Yoshida T, Kitamura T, Kawaichi S, Inoue T, Nomura K, Yoshida Y, Kuno S, Sako Y (2013) Variation of the virus-related elements within syntenic genomes of the hyperthermophilic archaeon *Aeropyrum*. Appl Environ Microbiol **79:**5891-5898.

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5:**e11147.

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:**1390-1400.

Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol **2:**414-424.

Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol **3:**504-510.

Escalante-Semerena JC (2007) Conversion of cobinamide into adenosylcobamide in bacteria and archaea. J Bacteriol **189:**4555-4560.

Fiala G, Stetter KO (1986) *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch Microbiol **145:**56-61.

Filée J, Siguier P, Chandler M (2007) Insertion sequence diversity in archaea. Microbiol Mol Biol Rev **71:**121-157.

Gao F, Zhang CT (2008) Ori-Finder: A web-based system for finding *oriC*s in unannotated bacterial genomes. BMC Bioinformatics **9:**79.

Garrity GM, Holt JG (2001) Phylum AI. *Crenarchaeota* phy. nov., pp. 169-210. *In* Boone DR, Castenholz RW, and Garrity GM (ed.), Bergey's Manual of Systematic Bacteriology, vol. 1. Springer, New York, NY.

Gaudin M, Krupovic M, Marguet E, Gauliard E, Cvirkaite−Krupovic V, Le Cam E, Oberto J, Forterre P (2014) Extracellular membrane vesicles harbouring viral genomes. Environ Microbiol **16:**1167-1175.

Giovannoni SJ, Cameron Thrash JC, Temperton B (2014) Implications of streamlining theory for microbial ecology. ISME J **8:**1553-1565.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science **309:**1242-1245.

Giraud MF, Naismith JH (2000) The rhamnose pathway. Curr Opin Struct Biol **10:**687-696.

Goericke RE, Welschmeyer NA (1993) The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep Sea Research (Part I, Oceanographic Research Papers) **40:**2283-2294.

Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol **3:**679-687.

Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res **35:**W52-W57.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappe MS (2012) Streamlining and core genome conservation among highly divergent members of the SAR11 clade. MBio **3:**e00252-12.

Gunbin KV, Afonnikov DA, Kolchanov NA (2009) Molecular evolution of the hyperthermophilic archaea of the *Pyrococcus* genus: analysis of adaptation to different environmental conditions. BMC Genomics **10:**639.

Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, Kosugi H, Hosoyama A, Fukui S, Nagai Y, Nishijima K, Nakazawa H, Takamiya M, Masuda S, Funahashi T, Tanaka T, Kudoh Y, Yamazaki J, Kushida N, Oguchi A, Aoki K, Kubota K, Nakamura Y, Nomura N, Sako Y, Kikuchi H (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* Kl. DNA Res **101:**83-101.

Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res **36:**6688-6719.

Koskella B, Thompson JN, Preston GM, Buckling A (2011) Local biotic environment shapes the spatial scale of bacteriophage adaptation to bacteria. Am Nat **177:**440-451.

Kunin V, He S, Warnecke F, Peterson SB, Martin HG, Haynes M, Ivanova N, Blackall LL, Breitbart M, Rohwer F, McMahon KD, Hugenholtz P (2008) A bacterial metapopulation adapts locally to phage predation despite global dispersal. Genome Res **18:**293-297.

Kuno S, Yoshida T, Kaneko T, Sako Y (2012) Intricate interactions between the bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repreat (CRISPR) signatures. Appl Environ Microbiol **78:**5353-5360.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol **5:**R12.

Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol **8:**317-327.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658-1659.

Lillestøl RK, Redder P, Garrett RA, Brügger KIM (2006) A putative viral defence mechanism in archaeal cells. Archaea **2:**59-72.

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 SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI Yakunin AF, van der Oost J, Koonin EV (2011a) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol **9:**467-477.

Makarova KS, Wolf YI, Snir S, Koonin EV (2011b) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol **193:**6039-6056.

Mao D, Grogan D (2012) Genomic evidence of rapid, global-scale gene flow in a *Sulfolobus* species. ISME J **6:**1613-1616.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nculeic Acids Res **30:**281-283.

Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IMA, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides NC (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res **36:**528-533.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev **15:**589-594.

Mochizuki T, Krupovic M, Pehau-Arnaudet G, Sako Y, Forterre P, Prangishvili D (2012) Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proc Natl Acad Sci USA **109:**13386-13391.

Mochizuki T, Sako Y, Prangishvili D (2011) Provirus induction in hyperthermophilic archaea: characterization of *Aeropyrum pernix* spindle-shaped virus 1 and *Aeropyrum pernix* ovoid virus 1. J Bacteriol **193:**5412-5419.

Mochizuki T, Yoshida T, Tanaka R, Forterre P, Sako Y (2010) Diversity of viruses of the hyperthermophilic archaeal genus *Aeropyrum*, and isolation of the *Aeropyrum pernix* bacilliform virus 1, APBV1 ,the first representative of the family *Clavaviridae*. Virology **402:**347-354.

Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature **393:**464-467.

Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature **420:**806-810.

Nakagawa S, Takai K, Horikoshi K, Sako Y (2004) *Aeropyrum camini* sp. nov., a strictly aerobic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney. Int J Syst Evol Microbiol **54:**329-335.

Nakamura Y, Nishio Y, Ikeo K, Gojobori T (2003) The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. Gene **317:**149-155.

Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE, Howard EC, King E, Oakley CA, Reisch CR, Rinta-Kanto JM, Sharma S, Sun S, Varaljay V, Vila-Costa M, Westrich JR, Moran MA (2010) Genome characteristics of a generalist marine bacterial lineage. ISME J **4:**784-798.

Nomura N, Morinaga Y, Kogishi T, Kim EJ, Sako Y, Uchida A (2002) Heterogeneous yet similar introns reside in identical positions of the

rRNA genes in natural isolates of the archaeon *Aeropyrum pernix*. Gene **295:**43-50.

Novichkov PS, Wolf YI, Dubchak I, Koonin EV (2009) Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol **191:**65-73.

Parter M, Kashtan N, Alon U (2007) Environmental variability and modularity of bacterial metabolic networks. BMC Evol Biol **7:**169.

Podar M, Anderson I, Makarova KS, Elkins JG, Ivanova N, Wall MA, Lykidis A, Mavromatis K, Sun H Hudson ME, Chen W, Deciu C, Hutchison D, Eads JR, Anderson A, Fernandes F, Szeto E, Lapidus A, Kyrpides NC, Saier MH, Richardson PM, Rachel R, Huber Harald, Eisen JA, Koonin EV, Keller M, Stetter KO. (2008) A genomic analysis of the archaeal system *Ignicoccus hospitalis-Nanoarchaeum equitans*. Genome Biol **9:**R158.

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res **35:**D61-D65.

Reno ML, Held NL, Fields CJ, Burker PV Whitaker RJ (2009) Biogeography of the *Sulfolobus islandicus* pan-genome. Proc Natl Acad Sci USA **106:**8605-8610.

Rhodes ME, Spear JR, Oren A, House CH (2011) Differences in lateral gene transfer in hypersaline versus thermal environments. BMC Evol Biol **11:**199.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16:**276-277.

Robinson NP, Bell SD (2007) Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. Proc Natl Acad Sci USA **104:**5806-5811.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan

MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature **424:**1042-1047.

Rocha EPC (2003) Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol **23:**513-522.

Sako Y, Nomura N, Uchida A, Ishida Y, Morii H, Koga Y, Hoaki T, Maruyama T (1996) *Aeropyrum pernix* gen. nov., sp. nov., a novel aerobic hyperthermophilic archaeon growing at temperatures up to 100°C. Int J Syst Bacteriol **46:**1070-1077.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from Yellowstone hot springs. Appl Environ Microbiol **74:**4164-4174.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol **5:**e75.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res **34:**D32-D36.

Sorek R, Kunin V, Hugenholtz P (2008) CRISPR-a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol **6:**181-186.

Sprott GD, Shaw KM, Jarrell KF (1983) Isolation and chemical composition of the cytoplasmic membrane of the archaebacterium *Methanospirillum hungatei*. J Biol Chem **25:**4026-4031.

Sugawara H, Ohyama A, Mori H, Kurokawa K (2009) Microbial Genome Annotation Pipeline (MiGAP) for diverse users, software demonstration S001-1-2. Abstr 20th Int Conf Genome Inform (GIW2009), Yokohama, Japan.

Tamames J. (2001) Evolution of gene order conservation in prokaryotes. Genome Biol **2:**1-0020.

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP,

Trüper HG (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol **37:**463-464.

West NJ, Schönhuber WA, Fuller NJ, Amann RI, Rippka R, Post AF, Scanlan DJ (2001). Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. Microbiology **147:**1731-1744.

Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science **301:**976-978.

Wolf YI, Makarova KS, Yutin N, Koonin EV (2012) Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol Direct **7:**46.

Yelton AP, Thomas BC, Simmons SL, Wilmes P, Zemla A, Thelen MP, Justice N, Banfield JF (2011) A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes. PLoS Comput Biol **7:**e1002230.

Yoon SH, Reiss DJ, Bare JC, Tenenbaum D, Pan M, Slagel J, Moritz RL, Lim S, Hackett M, Menon AL, Adams MWW, Barnebey A, Yannone SM, Leigh JA, Baliga NS (2011) Parallel evolution of transcriptome architecture during genome reorganization. Genome Res **21:**1892-1904.

Zillig W, Holz I, Janekovic D, Klenk HP, Imsel E, Trent J, Wunderl S, Forjaz VH, Coutinho R, Ferreira T (1990) *Hyperthermus butylicus*, a hyperthermophilic sulfur-reducing archaebacterium that ferments peptides. J Bacteriol **172:**3959-3965.

Publication list

1. Daifuku T, Yoshida T, Kitamura T, Kawaichi S, Inoue T, Nomura K, Yoshida Y, Kuno S, Sako Y (2013) Variation of the virus-related elements within syntenic genomes of the hyperthermophilic archaeon *Aeropyrum*. Appl Environ Microbiol **79:**5891-5898.

2. Daifuku T, Yoshida T, Sako Y (2013) Genome variation in the hyperthermophilic archaeon *Aeropyrum*. Mob Genet Elements **3:**e26833.

3. Daifuku T, Yoshida T, Sako Y Comparative genomic analysis of closely related hyperthermophilic crenarchaea. in preparation

4. Inoue T, Yoshida T, Wada K, Daifuku T, Fukuyama K, Sako Y (2011) A simple, large-scale overexpression method of deriving carbon monoxide dehydrogenase II from thermophilic bacterium *Carboxydothermus hydrogenoformans*. Biosci Biotechnol Biochem **75:**1392-1394.

5. Yoneda Y, Yoshida T, Kawaichi S, Daifuku T, Takabe K, Sako Y (2012) *Carboxydothermus pertinax* sp. nov., a thermophilic, hydrogenogenic, Fe(III)-reducing, sulfur-reducing carboxydotrophic bacterium from an acidic hot spring. Int J Syst Evol Microbiol **62:**1692-1697.

6. Ueta M, Wada C, Daifuku T, Sako Y, Bessho Y, Kitamura A, Ohniwa R, Morikawa K, Yoshida H, Kato T, Miyata T, Namba K, Wada A (2013) Conservation of two distinct types of 100S ribosome in bacteria. Genes to Cells **18:**554-574.

7. Yoneda Y, Yoshida T, Daifuku T, Kitamura T, Inoue T, Kano S, Sako Y (2013) Quantitative detection of carboxydotrophic bacteria *Carboxydothermus* in a hot aquatic environment. Fundam Appl Limnol **182:**161-170.

8. Inoue T, Takao K, Yoshida T, Wada K, Daifuku T, Yoneda Y, Fukuyama K, Sako Y (2013) Cysteine 295 indirectly affects Ni coordination of carbon monoxide dehydrogenase-II C-cluster. Biochem Biophys Res Commun **441:**13-17.