Title

Variation of the Virus-Related Elements within Syntenic Genomes of the Hyperthermophilic Archaeon *Aeropyrum*

Running title

Genomic variation in *Aeropyrum*

Authors

Takashi Daifuku, Takashi Yoshida, Takayuki Kitamura, Satoshi Kawaichi, Takahiro Inoue, Keigo Nomura, Yui Yoshida, Sotaro Kuno and Yoshihiko Sako*.

Affiliation

Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan.

***Corresponding author**

Yoshihiko Sako

Postal address: Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan. Phone: (+81) 75-753-6217. Fax: (+81) 75-753-6226. Email: sako@kais.kyoto-u.ac.jp

Journal section

Evolutionary and genomic microbiology

ABSTRACT

The increasing number of genome sequences for archaea and bacteria show their adaptation to different environmental conditions at the genomic level. *Aeropyrum* spp. are aerobic and hyperthermophilic archaea. *Aeropyrum camini* was isolated from a deep-sea hydrothermal vent, and *Aeropyrum pernix* was isolated from a coastal solfataric vent. To investigate the adaptation strategy in each habitat, we compared the genomes of the two species. Shared genome features were small genome size, high GC content, and a large portion of orthologous genes (86-88%). 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 repeats (CRISPR) elements. Spacer sequences derived from the *A. camini* CRISPR showed significant match with proto-spacers of the two proviruses infecting *A. pernix*, indicating *A. camini* interacted with viruses closely related to APSV1 and APOV1. Further, a significant fraction of the non-orthologous

37 genes (41-45%) were proviral genes or ORFans probably originating from viruses.

38 Although the genomes of *A. camini* and *A. pernix* were conserved, we observed

39 non-synteny that was primarily attributed to virus-related elements. Our findings

40 indicated the genomic diversification of *Aeropyrum* spp. are substantially caused by

41 viruses.

INTRODUCTION

Between closely related organisms, gene repertoires and genome organizations differ depending on the ecological characteristics of each habitat. For example, whole-genomic comparisons of the cyanobacterial *Prochlorococcus* spp., with physiological features relevant to the different ecological niches within a stratified oceanic water column, show gross signatures of this niche differentiation (1). The members of anaerobic hyperthermophilic archaeal genus *Pyrococcus* adapt to abiotic and biotic environmental conditions through positive gene selection (2). Although genomes of temperate coastal SAR11 isolates are highly conserved in the core genome common to all strains (3) and show synteny (the conservation of DNA sequence and gene order) (4), variations exist among genes for phosphorus metabolism, glycolysis, and C1 metabolism. This suggests the adaptive specialization in nutrient resource utilization is important to niche partitioning (5).

Aeropyrum species are heterotrophic, aerobic, neutrophilic, and hyperthermophilic archaea. The two currently known species, *Aeropyrum pernix* and *Aeropyrum camini* were isolated from geographically distinct locations. The type strain of the type species, *A. pernix* K1, was isolated from a coastal solfataric vent on Kodakara-Jima Island in south-western Japan (6), and 11 additional strains were isolated from a coastal shallow

hydrothermal vent and a coastal hot spring in south-western Japan (7). The complete genome sequence of *A. pernix* K1 was determined and it encodes for cytochrome *c* oxidases (COXs), which are the terminal electron acceptors of the respiratory chain of most aerobic organisms (8). 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, and is recognized as the first aerobic hyperthermophilic archaeon from a deep-sea hydrothermal environment (9). Despite the geographically distinct habitat of *A. pernix* and *A. camini,* they are phylogenetically closely related based on their 16S rRNA gene sequences (99%) and are similar in morphology and growth characteristics, except for some distinguishable physiological properties such as optimum temperature and pH range for growth (9). To examine the genetic differences between *A. camini* isolated from a deep-sea hydrothermal vent and *A. pernix* isolated from a coastal solfataric vent, we report here the complete genome sequence of *A. camini* and compare it to the previously completed genome sequence of *A. pernix*. This comparative genome analysis shows that the genomic variation is partly brought about by viruses.

MATERIALS AND METHODS

78 **Strain and DNA extraction**

79 *A. camini* SY1 was obtained from Deutsche Sammlung von Mikroorganismen und
80 Zellkulturen (DSMZ, Braunschweig, Germany) as DSMZ 16960. *A. camini* cells were
81 grown in a cotton-plugged 2,000 ml Erlenmeyer flask containing 500 ml JXTm medium
82 (7) using an air-batched rotary shaker (RGS-32.TT; Sanki Seiki, Osaka, Japan) at 120
83 rpm. The pH of the medium was 8.0 and the incubation temperature was 85°C. Cells in
84 the mid-exponential phase were harvested by centrifugation at 10,000×g for 1 min at
85 4°C. Cell pellets were stored at -80°C. DNA was extracted using a Wizard genomic
86 DNA purification system (Promega, Madison, WI) according to the manufacturer's
87 instructions. DNA was further purified using phenol/chloroform/isoamyl alcohol
88 (25:24:1) treatment and precipitated with 2-propanol. The DNA was dissolved in 100 μl
89 distilled deionized water.

90

91 **Genome sequencing, functional annotation, and comparative genomics**

92 The genome of *A. camini* SY1 was sequenced using a Roche 454 GS FLX Titanium
93 pyrosequencing platform (Roche Diagnostics, Burgess Hill, West Sussex, UK) with an 8
94 kb paired-end library. The GS FLX sequencing (1/4 plate) resulted in the generation of
95 about 116 Mb sequences with an average read length of 342 b, providing approximately

73-fold coverage of the genome. Reads were assembled onto a scaffold including large 10 contigs (> 500 b) using a GS *De Novo* Assembler ver. 2.3. The gaps between the contigs were filled by sequencing PCR products using a 3130 Capillary Sequencer (Applied Biosystems, Foster City, CA). The genome sequence was automatically annotated with the Microbial Genome Annotation Pipeline ver. 2.02 (10). 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. We modified the position of the start codon of ORFs with no RBS according to orthologous counterpart encoded on *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) using RPS-BLAST (11) 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 (12). We 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 hit (RBH), 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 (13). Overall similarity between genomes was generated with

114 the genome-to-genome distance calculator (GGDC) (14). This system calculates the
115 genomic distance and estimates DNA-DNA hybridization (DDH) values from a set of
116 formulas (1, HSP (high-scoring segment pairs) length / total length; 2, identities / HSP
117 length; and 3, identities / total length). Synteny plots were generated as alignments of
118 the complete genome nucleotide sequences using MUMMER 3.0 (15) and Mauve 2.3.1
119 (16). Insertion sequence (IS) elements were identified using the ISfinder database (17).
120 Multiple copies less than 600 bp long flanked by inverted repeats were identified as
121 miniature inverted-repeat transposable elements (MITEs) using the Einverted program
122 from EMBOSS (18).
123
124 **CRISPR analysis**
125 CRISPR (clustered regularly interspaced short palindromic repeats) elements and
126 spacers were identified using the CRISPRFinder (19) with manual validation. The
127 spacer sequences were clustered using CD-HIT-EST (20) with a local sequence identity
128 threshold of 90%, an alignment coverage threshold for a shorter sequence of 60%, and a
129 word size set at 7. Two available viral metagenomes from Yellowstone hot springs (21)
130 and from the Juan de Fuca ridge (22) were retrieved from GenBank trace archive and
131 from the CAMERA database (23), respectively. A similarity search of spacer sequences

was performed against the NCBI non-redundant (nr) database and the viral metagenomes using BLASTN (24) 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 (25). Putative orthologous genes were identified as RBH using BLASTP (26) 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 non-orthologous genes against their own proteomes using BLASTP (26) 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 using BLASTP (26) 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 (27). Genes were compared to the nr database using BLASTP (26) with an e-value of 10^{-5} and default parameters. Each gene whose top non-identical hit was not a gene of a member of the order Desulfurococcales had a normalized bit score (BLAST bit score to homologue divided by BLAST bit score to self) more than 25% greater than

the best hit to a Desulfurococcales gene, and had a bit score greater than 67 was flagged as a putative inter-order HGT gene. The donor species were assigned according to the top non-identical 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 extra-chromosomal elements. The general features of the circular chromosome were compared with those of *A. pernix* (Table 1). The chromosomes were similar in size (*A. camini*: 1,595,994 bp and *A. pernix*: 1,669,696 bp) and in %G+C content (*A. camini*: 56.7% and *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

168 for all 20 amino acids. A similar number of ORFs were identified (*A. camini*: 1,645 and
169 *A. pernix*: 1,700). Of all the ORFs, 70.6% and 70.9% were classified by COG in *A.*
170 *camini* and *A. pernix*, respectively. Although most archaeal genes are predicted to use an
171 AUG start codon, a large percentage of the predicted start codons were GUG (*A.*
172 *camini*: 27% and *A. pernix*: 30%) or UUG (*A. camini*: 41% and *A. pernix*: 38%). Similar
173 values in start codon usage were obtained from archaeal *H. butylicus* (28).
174 Their orthologous genes were identified using the RBH approach. Each of the
175 genomes carried 1,455 (86-88%) orthologous genes (Table 2). Genes involved in the
176 Embden-Meyerhof pathway and the tricarboxylic acid cycle were conserved in both
177 genomes (data not shown). The closest relative of *Aeropyrum* spp. is a
178 peptide-fermenting, sulfur-reducing, and hyperthermophilic archaeon *H. butylicus* (29) ;
179 *A. camini* and *H. butylicus* shared 772 (46-47%) orthologous genes and *A. pernix* and *H.*
180 *butylicus* shared 769 (45-46%) orthologous genes. The functional distribution of
181 non-orthologous genes between *Aeropyrum* spp. and *H. butylicus* were inspected (Fig.
182 1). A COG category with the greatest number of the non-orthologous genes was energy
183 production and conversion (C) except for two categories of general function prediction
184 only (R) and function unknown (S). This was consistent with the fact that *Aeropyrum*
185 spp. are aerobic, whereas *H. butylicus* is an anaerobic sulfur-reducer. *Aeropyrum* spp.

186 contained genes encoding COXs and *H. butylicus* contained genes encoding a sulfur
187 reductase instead of COXs (data not shown). Genome variation between *A. camini* and
188 *A. pernix* was described below in detail.
189 The *A. pernix* genome harbours at least two *oriC* sites on non-coding regions
190 containing crenarchaeal origin recognition boxes (ORBs), the binding sites for
191 Orc1/Cdc6 proteins, and *ori*-specific uncharacterized motifs (UCMs) (30). In the *A.*
192 *camini* genome, we predicted two *oriC* sites on non-coding regions located between
193 ACAM_0493 and ACAM_0494, where four copies of ORB and an UCM were present,
194 and ACAM_1253 and ACAM_1254, where an UCM was present (Fig. 2B and C). Both
195 *oriC* sites coincided with two GC disparity minima described by a Z-curve analysis (Fig.
196 2A).
197
198 **Genome phylogenetics**
199 DDH values estimated by DDGC three formulas were 63.57, 18.86, and 51.97,
200 respectively. Given that the DDH values for species delineation cut-off are above 70
201 (31), these data were comparable to the previous report that *A. camini* is a different
202 species from *A. pernix* (9). GSS based on orthologous genes was 0.87 and nucleotide
203 identity was 73.2-76.6% with a range of 86.2-90.2% of the two chromosomes,

204 indicating the close relationship of *A. camini* and *A. pernix*. Genome synteny decreases
205 with phylogenetic distance, although this relationship varies depending on the group
206 examined (32, 33). Next, we analyzed the degree of genome synteny between *A. camini*
207 and *A. pernix*.
208
209 **Genome synteny between *A. camini* and *A. pernix***
210 There were no large-scale rearrangements in the nucleotide alignment of *A. camini*
211 and *A. pernix* chromosomes, confirming the close relationship of them (Fig. 3).
212 Comparisons of closely related archaeal and bacterial genomes generally show
213 disruptions of synteny with a characteristic X-shape pattern in the dot-plots (34). The
214 factors that affect genome rearrangements are not well understood, but presumably may
215 be associated with the state of recombination systems and the abundance of mobile
216 elements in the respective genomes (35). It is suggested that the low frequency of
217 recombination in *Corynebacterium* spp. is likely due to the absence of RecBCD, a well
218 characterized recombinational enzyme complex in bacteria (36). The RecBCD system
219 was missing in archaea (37) including *Aeropyrum* spp. Thermoacidophilic archaea
220 *Sulfolobus* spp. show poor genome synteny owing to genome rearrangements induced
221 by a large number of mobile elements like IS elements (34-201 IS elements) and MITEs

(61-143 MITEs) (38). 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. Further, hyperthermophilic organisms are highly specialized in the narrow range of habitat and isolated from one another by geographic barriers (39). *Aeropyrum* spp., therefore, can be defined as specialists in the concept of specialists as opposed to generalists, where specialists often have small genomes encoding genes essential for cell maintenance, while most generalists have large genomes encoding additional genes for signal transduction or metabolism allowing survival in variable environments (35, 40). In the highly 'specialized' small genome of *Aeropyrum*, the disruption of gene regulation derived from synteny breaks may be limited due to elimination of individuals associated with reduced fitness.

Virus-related elements

Although both genomes showed synteny, we observed some synteny disruptions. The most prominent were contained in virus-related elements. First, *A. pernix* contains two proviral regions that were induced under suboptimal conditions (41). Both viruses containing circular double-stranded (ds) DNA genomes were isolated and named as

240 *Aeropyrum pernix* spindle-shaped virus 1 (APSV1) and *Aeropyrum pernix* ovoid virus 1
241 (APOV1) (41). The proviral sequences were absent from the *A. camini* genome at the
242 conserved tRNA sequences homologous with *attP* sites, the recombination sites for
243 viruses (Fig. 3), although we could not rule out the possibility that *A. camini* was cured
244 of the proviruses in isolation step repeated at least three times (9). A translocated
245 inversion of 2 kbp sequence was identified upstream of the integrated APSV1 genome.
246 The inversion might be caused by 12 bp inverted repeat observed in that region.
247 Second, synteny disruptions were observed in the CRISPR elements (Fig. 3). The
248 CRISPR system is a recently recognized defense mechanism in archaea and bacteria
249 against foreign DNA such as viruses and plasmids (42). CRISPR allows cells to
250 specifically recognize and destroy target sequences using spacers derived from invaders
251 and, in many respects, parallels the function of the eukaryotic RNA interference system
252 (43). CRISPR spacers effectively act as libraries of previous invasion by
253 extra-chromosomal elements. In practice, host-virus interactions are investigated by the
254 analysis of CRISPR spacers in the natural cyanobacterial community (44). *A. camini*
255 contained four CRISPR loci (Aca_1-4) composed of 14, 15, 27, and three repeat-spacer
256 units, respectively (Table 3). Aca_1 and Aca_3 were interrupted by genes which did not
257 show similarity to any other available protein sequences. According to the CRISPRdb

database, the *A. pernix* genome carried three CRISPR loci (Ape_1-3) composed of 26, 41, and 18 repeat-spacer units, respectively (Table 3). Each non-coding sequence upstream of the first CRISPR of all CRISPR elements was AT-rich (%G+C content ranging from 31.5 to 52.0% lower than that of each genome sequence) and included a RBS, TATA box, and B recognition element. Therefore, the sequences were considered to be leader sequences which are transcription initiation sites for the CRISPR (Fig. 4 empty boxes) (45). CRISPR-associated (*cas*) genes were adjacent to Aca_1, Aca_3, and Ape_2 (Fig. 4), but not to the others. The CRISPR/Cas types are classified on the basis of their repeat sequences, leader sequences, and *cas* genes (46, 47). In a previous report, Ape_2 CRISPR/Cas system was determined as the DNA targeting subtype I-A (46). A subtype I-A CRISPR/Cas system homologous to Ape_2 CRISPR/Cas system was identified in *A. camini* (Aca_3 CRISPR/Cas system) (Fig. 4). 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 CRISPR/Cas system notwithstanding the presence in Aca_1 CRISPR/Cas system of *cas3* peculiar to the type I CRISPR/Cas system (Fig. 4).

Fifty-nine and 85 CRISPR spacers were retrieved from *A. camni* and *A. pernix*, respectively, and no significant matches were found among them. When all 144 spacers

were compared to the NCBI nr nucleotide database, three spacers (two 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 3). This strongly suggested the *Aeropyrum* 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 in its genome. In general, single nucleotide mutation of targeted sequences can render CRISPR/Cas ineffectual (48, 49). *A. pernix* may avoid destroying its own genome due to 5 nucleotide discrepancies between Ape_1_4 spacer and the proviral sequence (Table 4). Of three provirus-matching spacers in Aca_3, two spacers showed synonymous and non-synonymous substitutions compared with their corresponding putative proto-spacers in proviruses (Table 4), indicating *A. camini* interacted with viruses which 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 in the Yellowstone hot springs (21) and the Juan de Fuca ridge (22) other than the APSV1 and APOV1 genomes. It is noteworthy that none of the CRISPR spacers matched the non-orthologous genes of *Aeropyrum* spp. which are extrachromosomal elements in most cases (described below). In this research, we

293 examined only 144 spacers collected from two *Aeropyrum* spp. More CRISPR spacers
294 might enable us to identify the spacers matched with non-orthologous genes.
295
296 **Non-orthologous genes in *A. camini* and *A. pernix***
297 Along with the virus-related elements that primarily contributed to the synteny
298 disruptions, non-orthologous genes were located on non-syntenic regions scattered over
299 the genomes. In *A. camini* genome, on the other hand, 56 non-orthologous genes (29%)
300 localized at 13-22 kbp, 314-331 kbp, 407-411 kbp, and 687-715 kbp. In *A. pernix*
301 genome, except for the proviral regions, 73 non-orthologous genes (30%) localized at
302 190-211 kbp, 284-305 kbp, 726-764 kbp, 1279-1286 kbp, and 1599-1644 kbp.
303 Of these, notable metabolic pathways missed in *A. camini* on the 726-764 kbp and
304 1279-1286 kbp non-syntenic regions in *A. pernix*, reflecting the smaller genome of *A.*
305 *camini* than that of *A. pernix*. L-rhamnose is a common component of the cell wall in
306 bacteria (50), and is also found in cytoplasmic membrane of methanogenic archaeon
307 *Methanospirillum hungatei* (51). *RmlABCD* genes involved in a nucleotide-activated
308 L-rhamnose (dTDP-L-rhamnose) synthesis pathway were identified (APE_1178-1181)
309 on the 726-764 kbp non-syntenic region in *A. pernix*. Cobamides (e.g., coenzyme B_{12})
310 are unique for their structural complexity, and archaea synthesize them through

311 salvaging cobinamide from the environment (52). Clustered genes involved in the
312 cobinamide salvaging pathway were found on the 1279-1286 kbp non-syntenic region
313 in *A. pernix*. These facts implied that *A. camini* may not be able to synthesize
314 L-rhamnose and cobamides.
315 A report shows geographical distribution of gene contents (e.g., mobile elements)
316 among *Sulfolobus islandicus* strains from hot springs separated by distance (53). This
317 suggests that the variation of metabolic pathways in *Aeropyrum* implies their locality,
318 although the pathways are not necessarily responsible for environmental adaptation.
319 Meanwhile, genetic islands are found within genomes of *S. islandicus* strains from a
320 single hot spring (54). The variation among strains (*A. camini* and *A. pernix*,
321 respectively) might be found in future analysis of more *Aeropyrum* spp. genomes.
322 Of the all non-orthologous genes, paralogous genes were identified (*A. camini*: five
323 genes and *A. pernix*: 16 genes) in the range of 3-7% by searching against their own
324 proteomes (Tables 2 and S1). In *A. pernix* genome, eight paralogous genes were
325 annotated as hypothetical proteins with no conserved domains; however, these
326 nucleotide sequences contained the MITEs noted above. The other genes were classified
327 into ORFans (*A. camini*: 86 genes and *A. pernix*: 31 genes) which did not show
328 similarity to any other available protein sequences in the nr database, HGT genes (*A.*

329 *camini*: 22 genes and *A. pernix*: 45 genes) and proviral genes (*A. pernix*: 70 genes)
330 (Tables 2 and S1). HGT events are likely to occur among organisms with similar life
331 styles and habitats in particular among archaeal and bacterial hyperthermophiles (27).
332 The donors of the HGT genes identified in the *Aeropyrum* genomes were thermophiles
333 or derived from environmental sequences collected from the thermophilic environment
334 in the range of 82-84% (Table S2). These were compatible with the concept that
335 *Aeropyrum* spp. are specialized in the thermophilic environment. The unclassified genes
336 in the analysis above were further inspected (Table S1).
337 Surveys for viral metagenomes suggest the diversity of viral sequences is vast and
338 remains largely unexplored (55). Therefore, it seems plausible that a major fraction of
339 archaeal and bacterial ORFans are derived from the poorly explored but vast viral gene
340 pool; although it is impossible to rule out that ORFans have homologs in multiple
341 genomes that avoid detection because of their rapid evolution (35). ORFans probably
342 derived from viruses and proviral genes accounted for 41-45% of non-orthologous
343 genes. From environmental samples collected at the coastal Yamagawa hot spring in
344 Ibusuki, Japan, two viruses infecting *A. pernix* were isolated: a dsDNA virus,
345 *Aeropyrum pernix* bacilliform virus 1, APBV1 (56), and a single-stranded *Aeropyrum*
346 coil-shaped virus, ACV (57). *A. camini* could not be infected by ACV (57) and its

susceptibility to the infection by APBV1 was not tested (56). Morphologically diverse virus-like particles were also observed in Yamagawa hot spring (56). Our 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.

Here we show that *Aeropyrum* may be specialized in aerobic and thermophilic environment, and accordingly, possess small and conservative genomes; nevertheless, *Aeropyrum* interact with diverse viruses and their genomic diversification are substantially caused by the viruses.

ACKNOWLEDGEMENT

This work was supported by Grant-in-Aid for Science Research (no. 20248023) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

FIGURE LEGENDS

Figure 1. 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.

Figure 2. 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 non-conserved bases.

Figure 3. 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 gray boxes, filled boxes, and thin lines, respectively, on the two nucleotide alighnments.

Figure 4. Schematic representation 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.

REFERENCES

1. **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.

2. **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.

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

4. **Bentley SD, Parkhill J.** 2004. Comparative genomic structure of prokaryotes. Annu. Rev. Genet. **38:**771-792.

5. **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.

6. **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.

7. **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.

8. **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* K1. DNA Res. **6:**83-101.

9. **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.

10. **Sugawara H, Ohyama A, Mori H, Kurokawa K.** 2009. Microbial Genome Annotation Pipeline (MiGAP) for diverse users. Software Demonstrations S001-1-2. 20th Int. Conf. Genome Inform. (GIW2009) Posters and Software Demonstrations, Yokohama, Japan.

11. **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. Nucleic Acids Res. **30:**281-283.

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

13. **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.

14. **Auch AF, von Jan M, Klenk HP, 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.

15. **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.

16. **Darling AE, Mau B, Perna NT**. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one **5**:e11147.

17. **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.

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

19. **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.

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

461 21. **Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M,**
462 **Mead D.** 2008. Assembly of viral metagenomes from yellowstone hot springs.
463 Appl. Environ. Microbiol. **74:**4164-4174.

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

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

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

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

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

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

479 28. **Brügger K, Chen L, Stark M, Zibat A, Redder P, Ruepp A, Awayez M, She**
480 **Q, Garrett RA, Klenk HP.** 2007. The genome of *Hyperthermus butylicus*: a
481 sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to
482 108°C. Archaea **2:**127-135.

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

30. **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.

31. **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.

32. **Rocha EPC.** 2006. Inference and analysis of the relative stability of bacterial chromosomes. Mol. Biol. Evol. **23:**513-522.

33. **Tamames J.** 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. **2:**RESEARCH0020.

34. **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.

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

36. **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.

37. **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.

38. **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.

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

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

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

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

524 43. **Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV.** 2006. A
525 putative RNA-interference-based immune system in prokaryotes: computational
526 analysis of the predicted enzymatic machinery, functional analogies with
527 eukaryotic RNAi, and hypothetical mechanisms of action. Biol. Direct **1:**7.

528 44. **Kuno S, Yoshida T, Kaneko T, Sako Y.** 2012. Intricate interactions between the
529 bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic
530 elements, revealed by diversified clustered regularly interspaced short
531 palindromic repeat (CRISPR) signatures. Appl. Environ. Microbiol.
532 **78:**5353-5360.

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

535 46. **Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath**
536 **P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin**
537 **EV.** 2011. Evolution and classification of the CRISPR-Cas systems. Nat. Rev.
538 Microbiol. **9:**467-477.

539 47. **Lillestøl RK, Shah SA, Brügger K, Redder P, Phan H, Christiansen J,**
540 **Garrett RA.** 2009. CRISPR families of the crenarchaeal genus *Sulfolobus*:
541 bidirectional transcription and dynamic properties. Mol. Microbiol. **72:**259-272.

48. **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.

49. **Deveau H, Barrangou R, Garneau JE, Labonté 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.

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

51. **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.

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

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

54. **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 biology **10:**e1001265.

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

56. **Mochizuki T, Yoshida T, Tanaka R, Forterre P, Sako Y, Prangishvili D.** 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.

57. **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.

573 **TABLES**

Table 1. Genome statistics of *Aeropyrum*.

	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	50 (2.95%)	50 (2.86%)
No. of ORFs	1645 (97.1%)	1700 (97.1%)
Genes assigned to COGs	1162 (70.6%)	1205 (70.9%)

574

575

576

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

	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

577

Table 3. Characteristics of the CRISPR elements of *A. camini* and *A. pernix*.

Species	CRISPR loci	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	-	-

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

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

580

581

Table 4. 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)	. . . . .	. . . . .

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

583 [b] A reverse complementary sequence is shown.

584 [c] Identical nucleotides and amino acids are indicated by dots. Synonymous and non-synonymous substitutions are shown in bold and italic letters, respectively.

Daifuku et al. Fig. 1.

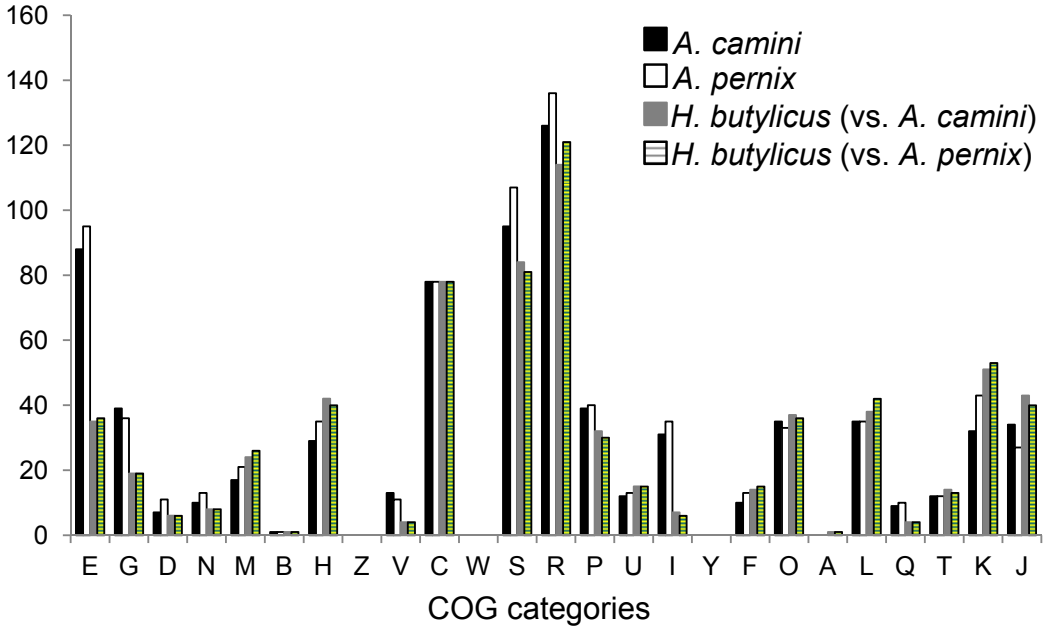

Figure 1. Numbers of non-orthologous genes between *Aeropyrum* spp. and *H. butylicus* assigned to COG functional categories. The one letter code for COG categories is the following: E, amino acid transport and metabolism; G, carbohydrate transport and metabolism; D, cell division and chromosome partitioning; 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.

Daifuku et al. Fig. 2.

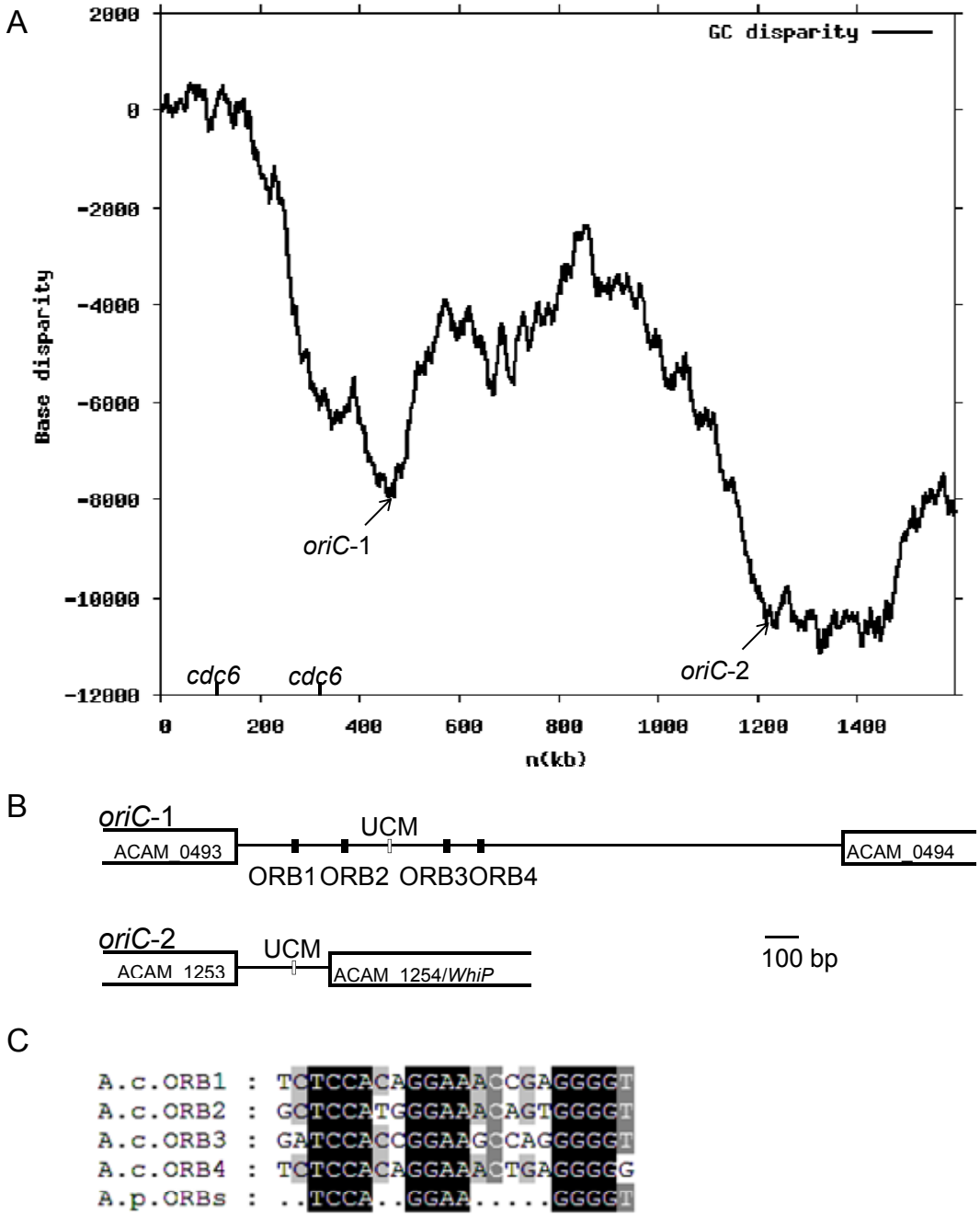

Figure 2. Prediction of *A. camini* replication origins. (A) The GC disparity curve for the *A. camini* genome. In the genome map, predicted *oriC* and *cdc6* genes are shown. (B) The structure of the predicted *oriC* region is shown. ORB elements, UCMs, and ORFs flanking the *oriC* site are shown as 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 non-conserved bases.

Daifuku et al. Fig. 3.

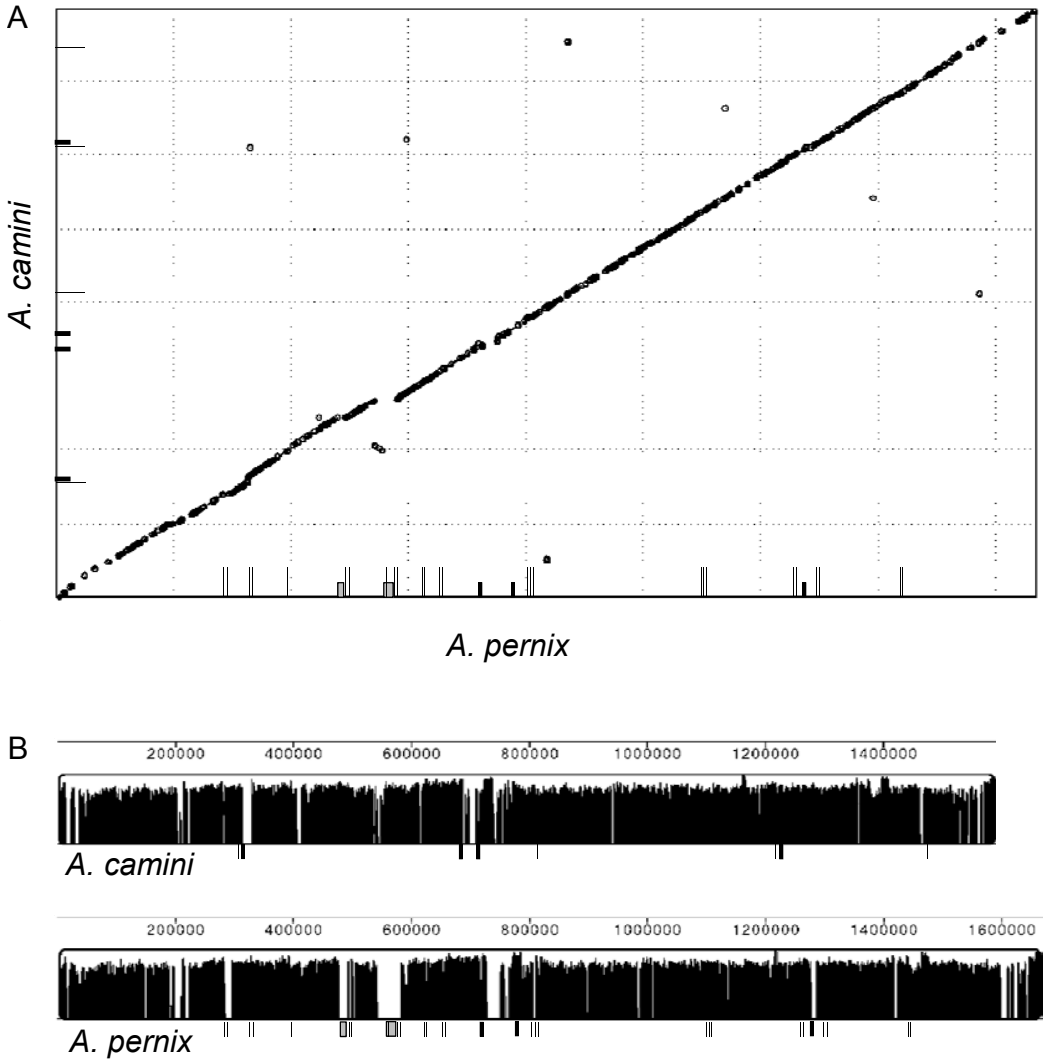

Figure 3. Comparison of the chromosomes of *A. camini* and *A. pernix*. (A) MUMMER nucleotide alignment, where dots indicate similar sequences shared by the two species. (B) Mauve nucleotide alignment, where the height of plots is proportional to the level of sequence identity in that region. Proviral regions, CRISPR elements, and MITEs are shown on the map in gray boxes, filled boxes, and thin lines, respectively, on the two nucleotide alighnments.

Daifuku et al. Fig. 4.

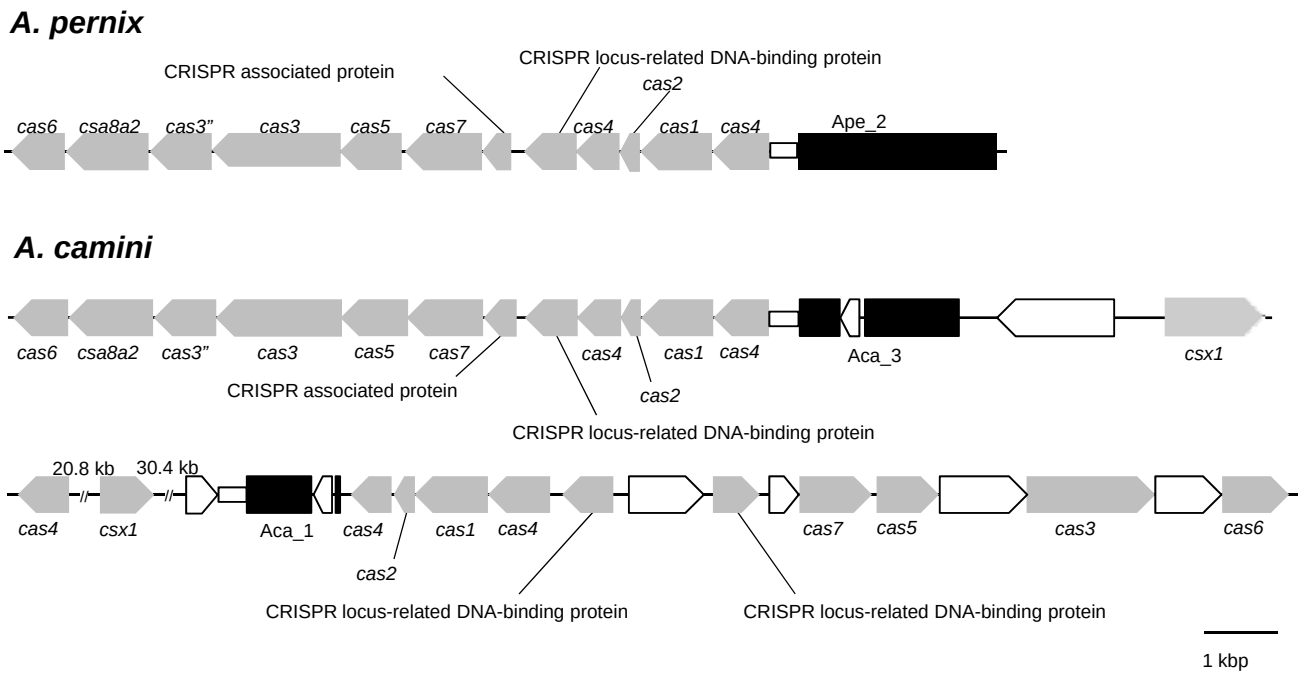

Figure 4. Schematic representation of the Ape_2, Aca_3, and Aca_1 CRISPR/Cas systems. ORFs, leader sequences, and CRISPRs are shown as arrows, empty boxes, and filled boxes, respectively. *Cas* genes are indicated in gray.

1 Supplemental tables

Table S1. Non-orthologous genes between *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]
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
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]
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]
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]
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

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
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
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
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
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
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
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
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]

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
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
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	HGT gene[a]

						of eukaryotic chaperone, SACSIN	
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
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
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
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]
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
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
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
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
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	-	-	-552	ORFan
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
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]
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.

Table S2. 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*	-
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*	+
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.

2