3204–3223 *Nucleic Acids Research, 2011, Vol. 39, No. 8* *Published online 15 December 2010*
doi:10.1093/nar/gkq1228

Insights into the evolution of *Archaea* and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group

Takuro Nunoura[1,*], Yoshihiro Takaki[2], Jungo Kakuta[1], Shinro Nishi[2], Junichi Sugahara[3,4], Hiromi Kazama[1], Gab-Joo Chee[2], Masahira Hattori[5], Akio Kanai[3,4], Haruyuki Atomi[6], Ken Takai[1] and Hideto Takami[2]

[1]Subsurface Geobiology & Advanced Research (SUGAR) Project, Extremobiosphere Research Program, Institute of Biogeosciences, [2]Microbial Genome Research Group, Extremobiosphere Research Program, Institute of Biogeosciences, Japan Agency for Marine-Earth Science & Technology (JAMSTEC), 2-15 Natsushima-cho, Yokosuka 237-0061, [3]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, [4]Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa 252-8520, [5]Center for Omics and Bioinformatics, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-no-ha 5-1-5, Kashiwa 277-8561 and [6]Department of Synthetic Chemistry and Biological Chemistry, Graduate School of Engineering, Kyoto University, Katsura, Nishikyo-ku, Kyoto 615-8510, Japan

Received August 25, 2010; Revised November 10, 2010; Accepted November 11, 2010

ABSTRACT

The domain *Archaea* has historically been divided into two phyla, the *Crenarchaeota* and *Euryarchaeota*. Although regarded as members of the *Crenarchaeota* based on small subunit rRNA phylogeny, environmental genomics and efforts for cultivation have recently revealed two novel phyla/divisions in the *Archaea*; the '*Thaumarchaeota*' and '*Korarchaeota*'. Here, we show the genome sequence of Candidatus '*Caldiarchaeum subterraneum*' that represents an uncultivated crenarchaeotic group. A composite genome was reconstructed from a metagenomic library previously prepared from a microbial mat at a geothermal water stream of a sub-surface gold mine. The genome was found to be clearly distinct from those of the known phyla/divisions, *Crenarchaeota* (hyperthermophiles), *Euryarchaeota*, *Thaumarchaeota* and *Korarchaeota*. The unique traits suggest that this crenarchaeotic group can be considered as a novel archaeal phylum/division. Moreover, *C. subterraneum* harbors an ubiquitin-like protein modifier system consisting of Ub, E1, E2 and small Zn RING finger family protein with structural motifs specific to eukaryotic system proteins, a system clearly distinct from the prokaryote-type system recently identified in *Haloferax* and *Mycobacterium*. The presence of such a eukaryote-type system is unprecedented in prokaryotes, and indicates that a prototype of the eukaryotic protein modifier system is present in the *Archaea*.

INTRODUCTION

The *Archaea* have long been presumed to consist of two phyla, the *Crenarchaeota* and *Euryarchaeota*. However, it has been established that diverse uncultivated lineages of *Archaea* inhabit every niche on this planet (1). Recent metagenomic analyses have revealed that two previously uncultivated *Archaea*, the group I marine crenarchaeote Candidatus (Ca.) '*Cenarchaeum symbiosum*' and the hyperthermophilic deeply branching Ca. '*Korarchaeum cryptofilum*', harbor both *Crenarchaeota*- and *Euryarchaeota*-specific genomic traits (2–5). Based on their unique phylogenetic positions and distinct genomic features, it has been proposed that *C. symbiosum* represents a novel phylum/division '*Thaumarchaeota*' (4). The unique genomic features of *K. cryptofilum* also support the proposal of '*Korarchaeota*' whose phylogenetic position had been discussed only based on SSU rRNA gene phylogenetic analysis (5). The proposal of '*Thaumarchaeota*' has

*To whom correspondence should be addressed. Tel: +81 46 867 9707; Fax: +81 46 867 9715; Email: takuron@jamstec.go.jp
Present address:
Gab-Joo Chee, Department of Biochemical Engineering, Dongyang Mirae University, 62-160 Gocheok Guro, Seoul 152-714, Korea

further been supported by the genome sequences of the marine archaeon Ca. '*Nitrosopumilus maritimus*' and the moderately thermophilic archaeon Ca. '*Nitrososphaera gargensis*' (6–9). On the other hand, the phylum '*Nanoarchaeota*', represented by the obligate symbiont Ca. '*Nanoarchaeum equitans*', has been proposed based on SSU rRNA gene phylogeny (10), but a later study using its genomic information suggested that the archaeal group is a fast evolving group within the *Euryarchaeota* (11).

Proteasome-mediated protein degradation coupled with protein modification with ubiquitin (Ub) is one of the hallmarks of eukaryotes (12). In eukaryotes, proteasome-mediated proteolysis is regulated by the Ub system, which is responsible for the conjugation of Ub to target proteins via the function of Ub-activating (E1), Ub-conjugating (E2) and Ub-protein ligating (E3) enzymes (12). Ub, E1 and E2 are members of distinct protein superfamilies that include structurally related proteins termed Ub-like (Ubl), E1-like (E1l) and E2-like (E2l) proteins, respectively. Although only distantly related to their eukaryotic counterparts, Ubl, E1l and E2l proteins are present in prokaryotes (13–15). For simplicity, based on primary structure, we will refer to these proteins as the 'prokaryote-type' Ubl, E1l and E2l proteins. In prokaryotes, some of the prokaryote-type Ubls and E1ls are responsible for sulfur incorporation in the biosynthesis of thiamine, molybdenum/tungstate cofactors and siderophores, while functions of other prokaryote-type proteins remain obscure (13,15). Recently, two proteasome-mediated proteolysis systems utilizing prokaryote-type proteins have been identified; the prokaryotic Ub-like protein (Pup)-proteasome system in *Mycobacterium tuberculosis* and the Ub-like small archaeal modifier proteins (SAMPs)-proteasome system in the halophilic archaeon *Haloferax volcanii* (16–18). In the *Haloferax* system, two prokaryote-type Ubls of the ThiS/MoaD family, which generally had been presumed to contribute in thiamine and molybdenum/tungstate cofactor biosynthesis together with prokaryote-type E1ls, have been shown to be involved in protein degradation via protein conjugation in the absence of E2/E3 homologs (16,18). These studies provided the first evidence that Ub–proteasome protein degradation occurs in *Archaea* and *Bacteria*. As these systems utilize prokaryote-type components, it is of increasing interest whether the origin of the eukaryote-type system resides in the prokaryotes.

The Hot Water Crenarchaeotic Group I (HWCGI) comprises putative thermophiles that have been detected in high-temperature environments such as terrestrial surface and subsurface hot springs, and deep sea hydrothermal environments, but have not yet been cultivated (7,19–22). The phylogroup is known to occupy a relatively deep position within crenarchaeotic lineages but distinct from hyperthermophilic *Crenarchaeota* or *Thaumarchaeota* in SSU rRNA gene phylogenetic analyses (7,21,22). From a geothermal water stream in a subsurface gold mine, we previously found unusual mat formation dominated by uncultured crenarchaeotic lineages including members of HWCGI, and constructed a metagenomic library to elucidate the physiology and genomic traits of these crenarchaeotes (21). Here, we present a composite genome sequence of a member of HWCGI, Ca. '*Caldiarchaeum subterraneum*', from the metagenomic library, and its unique genomic features that are distinct from previously reported archaeal genomes. In particular, the genome has revealed the presence of a eukaryote-type protein modifier system, a trait that had been believed to be inherent in *Eucarya*. The *C. subterraneum* genome harbors unique features that are distinct from previously reported archaeal genomes. The genome set provides clear insight into the biology of the novel deeply branching crenarchaeotic lineage, as well as the evolution of *Archaea* especially in the lineages which include the HWCGI, hyperthermophilic *Crenarchaeota*, *Thaumarchaeota* and *Korarchaeota*.

MATERIALS AND METHODS

Sampling, sample preparation and fosmid library construction

Sampling, DNA isolation and fosmid library construction have been previously described (21). The microbial mat community, in which HWCGI dominated, was taken from a geothermal water stream located at a depth of 320 m from the ground surface from a subsurface mine in Japan. High-molecular DNA up to 50 kb was extracted from microbial mat formation, and fosmid library using pCC1FOS (EPICENTRE, Madison, WI, USA) vector was constructed. Resulting totally 5280 fosmid clones were stored as glycerol stock in 96-well microtiter dishes at −80°C.

Screening for archaeal genome fragments encoding SSU rRNA gene

Genome fragments encoding archaeal SSU rRNA genes in the metagenomic library were reexamined by dot-blot hybridization with a digoxigenin-labeled DNA probe and anti digoxigenin antibody coupled to alkaline phosphatase using a DNA labeling and detection kit (Roche, Basel, Switzerland). SSU rRNA genes amplified from the genome fragments 10-H-8 (HWCGI (*C. subterraneum*); AB201309) and 45-H-12 [HWCGIII (*Nitrosocaldus* sp.); AB201308] obtained previously (7,21) were used as DNA probes. Archaeal SSU rRNA genes in the fosmids acquired by the dot-blot hybridization were amplified by PCR using primers A21F and U1492R (23,24) and directly sequenced from both strands.

Sequencing and enrichments of archaeal genome fragments, and annotation

All fosmid clones in the metagenomic library were extracted from *E. coli* culture, and paired-end sequences of each cloned genomic fragment were sequenced using Big Dye ver. 3.1 sequencing kit (Applied Biosystems, Foster City, CA, USA) in accordance with the manufacturer's recommendations by an ABI3730 DNA sequencer (Applied Biosystems). The end-sequences from cloned genomic fragments were analyzed by BLAST algorithm targeted to NCBI/EMBL/DDBJ database. On the other hand, as a part of metagenomic assessment for the whole microbial community (Takami *et al.*, unpublished data), 151 fosmid clones; 15 clones encoding SSU rRNA gene and 136 clones were randomly selected and sequenced by the whole-genome random-sequencing method described previously using ABI 3730 and the MegaBase 1000 (GE Healthcare, Piscataway, NJ, USA) (25,26).

Fifty-two fosmid clones encoding putative archaeal genome fragments were grouped into four individual pools containing equal weight of 13 fosmids. Each fosmid pool was analyzed in a half plate of the 454 DNA Genome Sequencer 20 (GS20) (Roche) at Takara Bio Inc. (Otsu, Japan). Large contigs obtained by 454 pyrosequencing were analyzed using BLAST algorithm targeted to genomic fragments encoding archaeal SSU rRNA genes reported previously (21), complete sequences of 151 fosmid clones analyzed by Sanger method (Takami *et al.*, unpublished data) and end-sequences of the genome fragments in the metagenomic library. Based on the homology search using BLAST, large scaffolds containing large contigs from 454 sequencing, complete fosmid clone sequences and fosmid-end sequences were manually constructed. In the second round of 454 sequencing, a total of 80 fosmids involving genome fragments extending previously sequenced regions and putative archaeal genome fragments were separated into four groups each containing 20 fosmids. The 20 fosmids in each group were analyzed in a half plate of the 454 GS20. Large contigs obtained from a total of four runs of GS20 were analyzed by BLAST targeting fosmid sequences analyzed by Sanger sequencing and fosmid end-sequences from the metagenomic library. A single large scaffold was manually constructed. Gap-regions in the scaffold were amplified by PCR with appropriate fosmids as templates, and the amplified fragments were analyzed using an ABI 3130xl DNA sequencer. Assembly in overlapping regions and gap regions was accomplished with Sequencher ver. 4.7 software (Gene Codes Corp, Ann Arbor, MI, USA). Finally, the large circular scaffold was constructed by the fosmid clone 10-H-8 (AB201309) reported previously (21), and JFF001_H02 (AP011633), JFF004_H08 (AP011650), JFF011_H10 (AP011675), JFF016_D08 (AP011689), JFF022_F09 (AP011708), JFF029_E04 (AP011723), JFF029_F10 (AP011724), JFF030_F06 (AP011727), JFF037_B02 (AP011745), JFF040_C01 (AP011751), JFF055_C09 (AP011796) analyzed by Sanger method (Takami *et al.*, unpublished data), and JFF001_G10 (AP011862), JFF002_G05 (AP011850), JFF004_B03 (AP011868), JFF005_B08 (AP011872), JFF008_E07 (AP011864), JFF009_A08 (AP011867), JFF009_F01 (AP011875), JFF009_F10 (AP011844), JFF011_A11 (AP011858, AP011859), JFF012_C01 (AP011870), JFF013_A09 (AP011845), JFF015_C06 (AP011842), JFF015_C07 (AP011830), JFF015_E11 (AP011831), JFF017_C01 (AP011851), JFF021_E09 (AP011873), JFF021_G03 (AP011856), JFF022_C07 (AP011838), JFF025_E12 (AP011827), JFF027_H06 (AP011834), JFF028_A01 (AP011854), JFF028_A10 (AP011876), JFF028_E01 (AP011852), JFF029_A12 (AP011865), JFF029_F08 (AP011836), JFF030_C12 (AP011869), JFF030_H11 (AP011855), JFF031_B05 (AP011861), JFF032_D08 (AP011843), JFF033_A05 (AP011857), JFF033_F07 (AP011840), JFF033_G03 (AP011849), JFF034_A01 (AP011853), JFF035_A09 (AP011828), JFF035_E02 (AP011848), JFF036_A12 (AP011839), JFF036_E03 (AP011833), JFF036_H04 (AP011837), JFF039_F10 (AP011846), JFF040_F12 (AP011871), JFF042_C08 (AP011829), JFF049_D05 (AP011863), JFF050_B05 (AP011866), JFF051_A09 (AP011832), JFF051_C10 (AP011826), JFF052_D03 (AP011874), JFF052_E01 (AP011841), JFF052_H05 (AP011847), JFF053_A03 (AP011860) and JFF055_E04 (AP011835) analyzed by the GS20 in this study. Numbers in parentheses following each fosmid clone are accession numbers in DDBJ/EMBL/GenBank database.

The predicted ORFs were initially defined by Glimmer program (http://www.cbcb.umd.edu/software/glimmer/), and putative functions for predicted ORFs were identified by comparing against all non-redundant (NR) sequences deposited in the NCBI database using BLASTP (27). Truncated ORFs and frame shifts found in the initial BLASTP search were confirmed by re-sequencing by the Sanger method. Clusters of Orthologous Groups (COGs) (28), archaeal Clusters of Orthologous Groups (arCOGs) (29) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (30) databases were used for further functional information. For the comparison of genome core genes, publically available archaeal genome sequences in the arCOG database were used, and arCOGs in *K. cryptofilum* were referred to from Elkins *et al.* (5). Assignments of arCOGs for *C. subterraneum* and *N. maritimus* were performed under the following condition; the BLAST E-value threshold was set at 10^{-3}, and the homologous region covers >70% of the hit sequences in arCOGs. Proteins that were putatively separated or fused compared to those in the databaes were manually concatenated or divided, and reexamined. Forty-six tRNA genes were identified by using tRNAscan-SE (31) with *Archaea*-specific search mode and SPLITSX (32) with the following parameters: *–p* 0.55 *–f* 0 *–h* 3. Clusters of regularly interspaced repeats (CRISPR) were identified using the CRISPR Finder (33).

Phylogenetic analyses

The small and large subunit rRNA gene alignments were constructed by ARB software (34). Then, concatenated alignments were constructed using only unambiguously aligned region for phylogenetic analysis. The maximum likelihood tree was computed by using the program package PhyML with HKY85 (35). The support values for the internal nodes were estimated from 100 bootstrap replicates. Protein sequences; RNAP subunits, ribosomal proteins, D-type DNA polymerase (DNAP) small and large subunits and elongation factor II (EFII) were aligned by using CLUSTAL W 1.8 program (36), and ambiguous regions were automatically trimmed according to Gblocks (37,38). Two concatenated alignments were constructed for the phylogenetic analyses of ribosomal proteins (L10, L10e, L11, L13, L14, L15, L15e, L18e, L19e, L2, L22, L3, L30, L44e, L4e, L5, L6, L7Ae, S10, S11, S13, S19, S19e, S2, S27e, S3, S3Ae, S4, S4e, S5, S6e, S7, S8, S8e, S9, S17, S17e, L1, L18, L24, L31e, L32e, S12, S15, L23) and RNAP subunits (RpoA′, RpoA″, RpoB′, RpoB″, RpoD, RpoE′, RpoH and RpoK), and concatenated (SSU+LSU) DNAP. Maximum likelihood trees were constructed using the program package RAxML with WAG+I+G (39). The support values for the internal nodes were estimated from 200 bootstrap replicates. Almost full length of *ef2* sequence from the *Nitrosocaldus* sp. (HWCGIII) was obtained by PCR amplification from the DNA assemblage. A primer set (5′-AATNGCNCAYGTNGAYCAYGGMAARAC-3′, and 5′-GTCTCWGMTGCAGGTATCTC-3′) for the amplification of *ef2* was constructed based on DNA alignments of *ef2* from crenarchaeal lineages including partial *ef2* sequence from the *Nitrosocaldus* sp. (HWCGIII) (31-F-01; GI 106364417) that were obtained from the metagenomic fosmid library used in this study.

Alignments of Ub-like protein family, E1-like protein family, E2-like protein family and JAMM protease family shown in Figure 2 were constructed by ClustalX (40) and edited manually based on the previously reported secondary structures of each protein family (13–15,41–44).

RESULTS

Archaeal diversity within the metagenomic library

As a result of dot blot hybridization and previous PCR screening, a total of 21 and three fosmids-encoding SSU rRNA genes of HWCGI and HWCGIII (Ca. ‘*Nitrosocaldus*’ sp.; SSU rRNA gene similarity between ammonia oxidizing thaumarchaeon Ca. ‘*Nitrosocaldus yellowstonii*’ (21) and the HWCGIII sequences in the metagenomic library [AB201308] was 95%) lineages, respectively, were obtained from the metagenomic library. Among the 21 fosmids-harboring HWCGI SSU rRNA genes, 19 SSU rRNA gene sequences belonged to ribotype I represented by the SSU rRNA gene included in the fosmid clone 10-H-08, while the other two sequences constituted another single ribotype. Here, we named the predominant HWCGI archaeon represented by the 10-H-08 SSU rRNA gene ribotype as Ca. ‘*C. subterraneum*’ (*Caldiarchaeum* type I) (‘*calidus*’ and ‘*subterraneum*’ meaning hot and underground, respectively) and the other minor HWCGI population as ‘*Caldiarchaeum* type II’. Similarity between the two ribotypes of *Caldiarchaeum* SSU rRNA gene sequences was 96.6%. Sixteen of the *C. subterraneum* SSU rRNA genes, each harbored two introns. Three orthologous sequences with 99% similarity were observed among the 16 sequences of the first intron, while five sequences with 95–99% similarity were found for the second intron. No diversity was present among all exon SSU rRNA gene sequences in the *C. subterraneum* SSU rRNA.

Reconstruction of a composite genome

In order to investigate the genomic properties of the metagenomic library, paired- or one-end sequences of the genome fragments were obtained from 3375 fosmid clones, and 151 fosmids (136 randomly selected fosmids and 15 fosmids encoding SSU rRNA gene) were analyzed by Sanger method (Takami *et al.*, unpublished data). Among a total 5965 end-sequences from these cloned fragments, 883 end-sequences (~13.5 % of total end-sequences) displayed highest similarity with sequences derived from *Archaea*. Among these ‘archaeal’ sequences, fosmids were selected for 454 sequencing based on the following two criteria: (i) the presence of paired-ends sequences predicted to encode open reading frames (ORFs) most similar to archaeal sequences; or (ii) the presence of ORFs in either end encoding homologues of archaeal translation, transcription or replication genes. Large contigs obtained by initial 454 sequencing of the 52 fosmids were manually assembled with the sequences from the 151 fosmids described above, two genome fragments-encoding archaeal SSU rRNA genes obtained previously (21) and the end-sequences of all fosmids, followed by a BLAST search. In this step, a scaffold of >1 Mb including the *C. subterraneum* SSU rRNA gene was assembled, but we did not find a large scaffold with other archaeal SSU rRNA genes. For the second round of 454 sequencing, 80 fosmids that met the following criteria were further analyzed: (i) linkage with the scaffold including the *C. subterraneum* SSU rRNA gene sequence; (ii) presence of paired-ends predicted to encode ORFs most similar to archaeal sequences; and (iii) presence of ORFs in either end showing high similarity with archaeal sequences. After the second 454 sequencing, large contigs obtained from 454 sequencing, fosmids analyzed by Sanger method and end-sequences were manually assembled and subjected to BLAST search. As a result, a circular scaffold including complete sequences of 12 fosmid clones analyzed by Sanger sequencing was obtained. The similarities of overlapping regions were generally >99%. Afterwards, gap-regions were obtained by PCR with appropriate fosmid clones as templates, and the amplified fragments were sequenced by Sanger method. Finally, a composite circular genome sequence of *C. subterraneum* (1 680 938 bp) was assembled from a set of 62 complete or partial fosmid sequences (Figure 1). We also obtained 28 complete or partial fosmid sequences derived from *C. subterraneum*, and 10 of them completely overlapped with the composite circular genome. However, 18 sequences harbored distinct insertion (a total of 68 kb)/deletion regions compared to the composite circular genome, or consisted of two genomic regions distantly located on the composite circular genome. The similarities of these regions with the circular genome were >99%. The genomic heterogeneity is likely the result of recombination or rearrangement within a species because we could not obtain any evidence of inter-species genomic recombination in the distinct insertion regions.

General features

The G + C content of the genome from *C. subterraneum* is 51.6%. A single rRNA gene set is identified but rRNA genes do not form an operon structure in the composite genome. Forty-five tRNAs were identified. A total of 1730

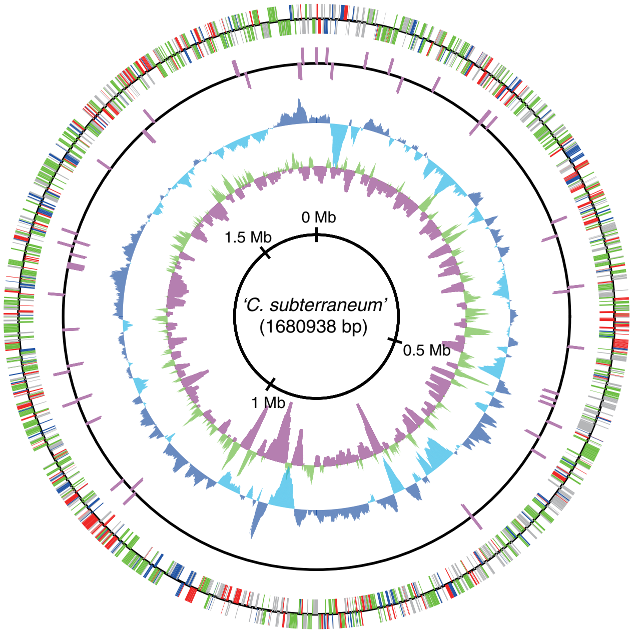


Figure 1. Circular representation of the *C. subterraneum* composite genome. From the inside, the first and second circles show the GC skew (values >0 or <0 are indicated in green and pink, respectively) and the G+C percent content (values greater or smaller than the average percentage in the overall chromosome are shown in blue and sky blue, respectively) in a 10-kb window with 100-bp step, respectively. The third and fourth circles show the presence of RNAs (rRNA and tRNA); CDSs aligned in the clock-wise and counterclock-wise directions are indicated in the upper and lower sides of the circle, respectively. Colors of CDSs indicate their functional categories; red for information storage and processing, green for metabolism, blue for cellular processes and signaling, and gray for poorly characterized function.

predicted ORFs were detected. Among these, 1054 of the predicted protein-encoding sequences (CDSs) could be assigned a function, 352 of the CDSs could be identified as hypothetical conserved proteins and the remaining 324 CDSs did not show significant similarity to any of the amino acid sequences in the protein databases (Supplementary Table S1).

Mobile genes

The genome contains three genes encoding transposases of the IS6 family and one of the IS4 family. Both of these transposase families, originally found in *Bacteria*, are distributed only in the *Euryarchaeota* and not in the *Crenarchaeota* within the archaeal domain (45). Four clustered regularly interspaced short palindromic repeats (CRISPR) and one CRISPR-related gene cluster, presumed to provide resistance against virus infection, are present (46). The genome encodes one prophage-like gene cluster.

DNA replication, repair, cell cycle

Caldiarchaeum subterraneum carries three orc1/cdc6 orthologues and a single minichromosome maintenance protein. The genome encodes multiple DNA-dependent DNA polymerases including two family B type enzymes; the BII type found only in crenarchaeal lineages (47) and the inactivated type (48), and both the small and large subunits of a D-type enzyme (Table 1). Genes for the large and small subunits of replication factor C form a gene cluster. Single genes each encoding the small subunit of primase, sliding clamp (PCNA), ATP dependent ligase, RNase HII, flap endonuclease (FEN1) and ERCC4-like helicase, are present. Genes for one truncated and two complete large subunits of primase are found. Unlike the Hef protein found in *P. furiosus* that consists of ERCC4-like helicase (COG1111) and XPF protein domains (ERCC4-type nuclease), which is the case in most of the euryarchaeotes, both domains are located separately on the genome of *C. subterraneum* as observed in *Thaumarchaeota* and a minority of euryarchaeotes (8,49,50). The ERCC4-like helicase domain (COG1111) is absent from the genome of *Korarchaeota* (8). Both topoisomerase IA and IB were found in *C. subterraneum* as in the case of *Thaumarchaeota* (8) (Table 1). One reverse gyrase gene, which had been considered a genomic signature for hyperthermophiles, but now also detected in thermophiles, is observed (51–53). Genes for chromatin-associated proteins, two Alba and one histone, are present. The archaeon possesses genes for euryarchaeal chromosome segregation proteins including SMC family ATPase, chromosome segregation and condensation protein B and kleisin family Rec8/ScpA/Scc1-like protein (chromosome segregation and condensation protein A) in a single, operon-like structure. The genome harbors one gene for the cell division protein FtsZ. Among the newly identified crenarchaeal cell division proteins CdvA, CdvB and CdvC that have been identified in *Thaumarchaeota* and hyperthermophilic *Crenarchaeota* (with the exception of the *Thermoproteales*), CdvB and CdvC are present but a gene for CdvA is absent in *C. subterraneum* (8,54).

The genome contains genes for double-strand-break repair, direct repair, base excision repair and nucleotide excision repair including photolyase and family Y DNA polymerase, which have previously been found only in *Sulfolobales* among the hyperthermophilic crenarchaeotes (55,56). However, XPB helicase for excision repair, mismatch detection proteins MutS and MutL, mismatch glycosylase MIG and bacterial nucleotide excision repair protein UvrABC are absent.

Translation and transcription

Forty-six tRNAs corresponding to all 61 sense codons and one initiator codon can be identified. Thirteen tRNAs are predicted to be intron-containing tRNAs and three out of the 13 harbor multiple introns (tRNA^{Leu} UAA, tRNA^{Gln} CUG, tRNA^{Thr} GGU). The introns are located not only at anticodon loop regions (canonical position) but also various non-canonical positions (D-arm, V-arm and T-arm), as observed in other crenarchaeal species (57,58). The BHB structure, a well-known motif of archaeal tRNA splicing, is found at exon–intron junctions of tRNA and the corresponding heterotetrameric splicing endonuclease can be identified. Aminoacyl tRNA

Table 1. Distribution patterns of representative components for DNA replication/repair, cell division, translation and transcription among *Crenarchaeota*, *Euryarchaeota*, *Thaumarchaeota*, *Korarchaeota* and *C. subterraneum*

	C. subterraneum	*Crenarchaeota*	*Euryarchaeota*	*Thaumarchaeota*	*Korarchaeota*
Major DNA polymerases[a]	BII, D	BI, BII	BI, D	BII, D	BI, BII, D
Chromosome segregation ATPase	+	–	+	+	+
ERCC4 like helicase (COG01111)	+	–		+	–
Topoisomerase I	IA, IB	IA	IA	IA[b], IB	IA
FtsZ	+	–	+	+	+
Hisotne	+	–[c]	+	+	+
RNA polymerase RpoA	fusion	split	split	fusion	fusion
RNA polymerase RpoB	fusion	split	split/fusion[d]	fusion	fusion
RNA polymerase RPB8	–	+	–	–	+
Ribosomal protein S25, S26, S30	+	+	–	+	+
Ribosomal protein L14e, 34e	+	+	+ (some)	–	+
Ribosomal protein L13e	–	+	–	(+)[e]	+
Ribosomal protein LXa	–	+	+ (most)	–	–
Ribosomal protein L39e	–	+	+	+	–

+, present; –, absent.

[a] Characterization of DNA polymerase is based on Ref. (47).

[b] Only C-terminal domain is found in *C. symbiosum* and *N. maritimus*.

[c] Only found in *Thermofilum pendens* and *Caldivirga maquilingensis*.

[d] Fusion form is observed in *Thermococcales* and *Thermoplasmatales*.

[e] Only found in *N. gargensis*.

synthetases for all of the amino acids are encoded in the genome except for the enzyme for glutaminyl tRNA synthesis, however, glutaminyl tRNA formation is likely dependent on heterodimeric glutamyl-tRNA amidotransferase (GatD and GatE). A selenocysteine incorporation system is lacking, resembling other genomes from crenarchaeal lineages (59).

The archaeal DNA-dependent RNA polymerase in *C. subterraneum* lacks the orthologue of the eukaryotic subunit RPB8 found in the hyperthermophilic crenarchaeotes and *Korarchaeota* (5,60), and possesses all other subunits found in the *Archaea*. RpoA is not fragmented as in eukaryotes, and is similar to those of *Thaumarchaeota* and *Korarchaeota* (Table 1). An ortholog of the eukaryotic RNA polymerase III subunit RPC34 is also found in *C. subterraneum* as in the hyperthermophilic *Crenarchaeota*, *Thaumarchaeota* and some of the *Euryarchaeota* but not the *Korarchaeota* (61). Archaeal homologs related to transcriptional initiation such as transcription factor B (TFB), TATA-binding protein (TBP) and transcription factor E (TFE) are present.

A complete set of 28 archaeal SSU ribosomal proteins are present, including S25e, S26e and S30e, that are absent in the *Euryarchaeota* (4,8,62) (Table 1). A total of 34 LSU ribosomal proteins are present. Although L39e is conserved in the *Euryarchaeota* and hyperthermophilic *Crenarchaeota*, L39e, along with L13e, L35ae, L38e, L41e and LXa (L20a/L18s), was not present on the genome. The absence of L13e is a euryarchaeal feature, and that of L35ae and LXa (L20a/L18a) is common to the *Thaumarchaeota* and *Korarchaeota* (4,5,8,62). The lack of L39e has also been noted in the *Korarchaeota* (4). We observed that L14e and L34e, which are not conserved in the *Thaumarchaeota*, are present on the *C. subterraneum* genome (Table 1).

Energy metabolism

The predicted gene set suggests the potential of chemolithotrophic growth in *C. subterraneum* using hydrogen or carbon monoxide as an electron donor, and oxygen, nitrate or nitrite as an electron acceptor. One Ni–Fe NADP-reducing hydrogenase and one potential aerobic type carbon monoxide dehydrogenase were detected. However, the hydrogenase is phylogenetically similar to those of heterotrophic organisms and potential aerobic type carbon monoxide dehydrogenase lacks biochemical evidence (21). In the respiratory chain, one set of complex II (succinate dehydrogenase), an incomplete complex I (NADH dehydrogenase), cytochrome *b*, rieske protein, heme-copper terminal oxidase, membrane-bound nitrate reductase and periplasmic nitrite reductase are each present. Genes for cytochrome *b*, rieske protein and potential cytochrome *c* are distributed separately on the genome. The subunit II of heme-copper terminal oxidase harbors copper-binding motif residues that are signatures of cytochrome *c* oxidase but not quinone oxidase (63).

Central metabolism

An almost complete Emden-Meyerhof pathway and complete tricarboxylic acid (TCA) cycle are present, but phosphofructokinase that is necessary for glycolysis is missing. ATP citrate lyase and its alternatives such as citryl-CoA synthetase and citryl-CoA lyase (64) are also lacking in the genome. Therefore, the reductive TCA cycle most likely does not function. Genes encoding enzymes for the Calvin–Benson cycle and reductive acetyl-CoA pathway are also not observed. Recently, two carbon assimilation pathways; the 3-hydroxypropionate/4-hydroxybutyrate cycle and the dicarboxylate/4-hydroxybutyrate cycle have been recognized in crenarchaeal lineages. The two cycles utilize distinct carbon dioxide/bicarbonate-fixing pathways to convert acetyl-CoA to succinyl-CoA, but share a common route in converting the succinyl-CoA to two acetyl-CoA molecules (65–69). Enzymes responsible for the conversion of acetyl-CoA and bicarbonate into succinyl-CoA in the 3-hydroxypropionate/4-hydroxybutyrate cycle, methylmalonyl-CoA epimerase, methylmalonyl-CoA mutase and biotin carboxylase and L-chain subunit of acetyl-CoA carboxylase (65,69), are not found in *C. subterraneum* . In contrast, all enzymes converting acetyl-CoA into succinyl-CoA by fixing carbon dioxide and bicarbonate in the dicarboxylate/4-hydroxybutyrate cycle are present. Intriguingly however, although all other enzymes necessary for the regeneration of acetyl-CoA from succinyl-CoA are present, the gene for 4-hydroxybutyryl-CoA dehydratase cannot be found on the genome.

The organism does not have the non-oxidative pentose phosphate pathway that is required for standard pentose/nucleic acid biosynthesis. However, three alternative pathways that replace the non-oxidative pentose phosphate pathway can be identified; the ribulose monophosphate (RuMP) pathway that converts fructose 6-phosphate to ribulose 5-phosphate (70,71), the archaeal 2-deoxyribose 5-phophate aldolase (DERA) pathway that can produce deoxyribose 1-phosphate from glyceraldehyde 3-phosphate and acetaldehyde (72), and the 6-deoxy-5-ketofructose-1-phosphate (DEFP) pathway that supplies 3-dehydroquinate (73,74).

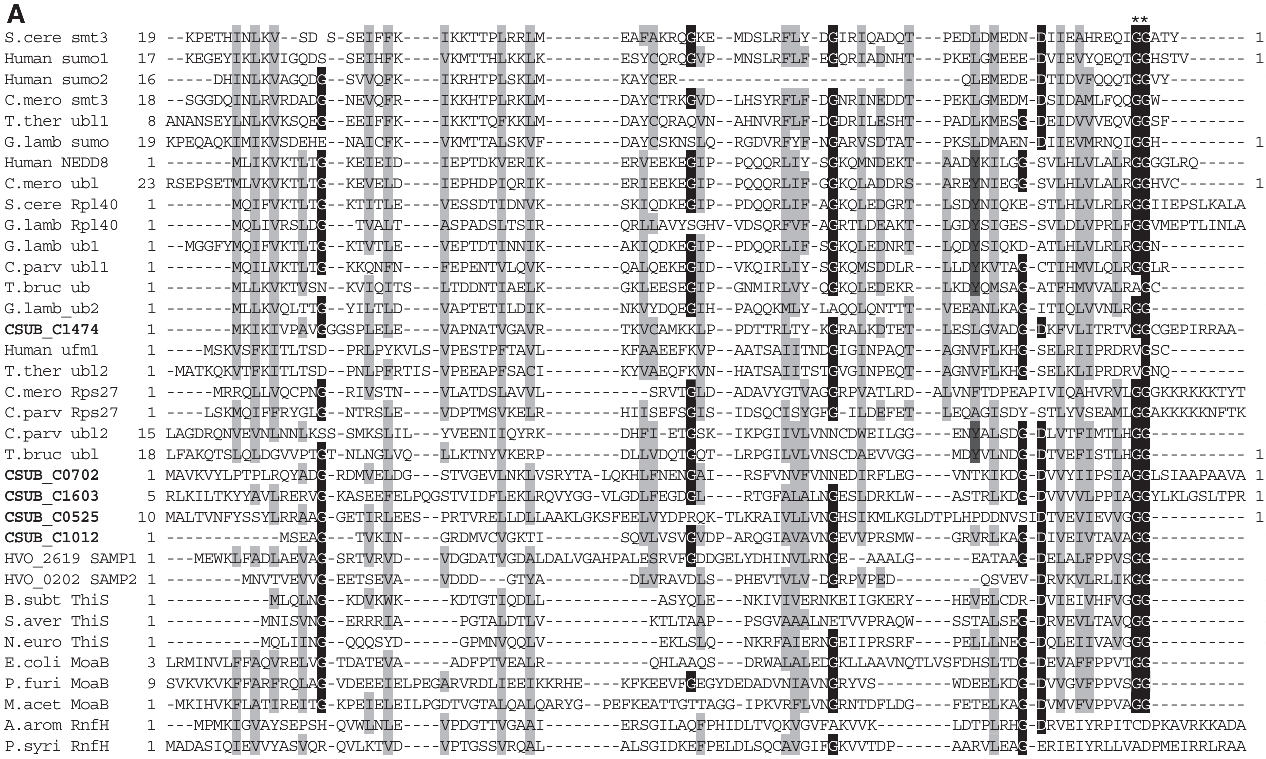
Protein folding and heat shock proteins

The genome possesses gene sets of heat shock proteins such as sHsp, Hsp60, Hsp70 and HtpX. Homologues of Hsp70 related proteins such as DnaJ, DnaK and GrpE have only been found in mesophilic euryarchaeotes and the *Thaumarchaeota* among the *Archaea* (75). Genes for NAC protein, prefoldin, FKBP-type peptidyl-prolyl cis-trans isomerase and thioredoxin are present, but those for Lon and Clp protease are absent.

Ub-like protein modifier system

Among the various unique traits of *C. subterraneum*, an unparalleled finding is the presence of a potential protein-degradation pathway consisting of a eukaryote-type Ub conjugation system associated with proteasome and AAA$^+$ family ATPase. As mentioned above, the structural features of the components of this system clearly distinguishes this system from the prokaryote-type systems recently identified in the *Archaea* and *Bacteria*. In the *H. volcanii* SAMPs-proteasome system, two of five prokaryote-type Ubls (ThiS/MoaD) identified in this haloarchaeon have been shown to conjugate with proteins and function as SAMPs, and conjugation between the SAMPs and a prokaryote-type E1l (MoeB) has been observed (18). *C. subterraneum* possesses four prokaryote-type Ubl (ThiS/MoaD) genes (CSUB_C0702, CSUB_C1012, CSUB_C0525 and CSUB_C1603) along with a molybdenum cofactor/tungstate cofactor biosynthesis pathway including a single prokaryote-type E1l (MoeB) gene (CSUB_C1135). These genes may be involved in a prokaryote-type protein modifier system similar to that found in *H. volcanii* (Figure 2A). Interestingly however, two of the prokaryote-type Ubls (CSUB_C0702 and CSUB_C1603) in *C. subterraneum* have 89 and 12 additional residues following the C-terminal Gly-Gly motif, in contrast to most archaeal prokaryote-type Ubl (MoaD) sequences which terminate after the Gly-Gly sequence (13) (Figure 2A).

In addition to these homologues, the *C. subterraneum* genome harbors an operon-like gene cluster encoding homologues of eukaryote-type Ubl, E1l and E2l (CSUB_C1474, CSUB_C1476 and CSUB_C1475, respectively), suggesting the presence of an unprecedented eukaryote-type Ubl system (Figures 2 and 3). Furthermore, while an apparent homologue of E3 is absent in the genome, a gene for a small Zn finger protein (CSUB_C1477) containing a RING finger motif ($C-X_2-C-X_{11}-C-X_2-C-X_4-H-X_2-C-X_{10}-C-X_2-C$) that mediates the Ub ligase activity of RING-type E3s (76) is also found in the same operon-like gene cluster (Figure 3). Moreover, a gene for RPN11-like protein (RPN11l) (CSUB_C1473), which is the homologue of eukaryotic 26 S proteasome regulatory subunit constituting a part of the proteasome lid sub-complex that catalyzes de-ubiquitination of captured substrates (77,78), is juxtaposed to the operon-like structure in the reverse strand (Figures 2D and 3). The Ubl, E1l and E2l harbor the key residues necessary for their respective functions, and are much more similar to their eukaryotic counterparts than to the prokaryote-type proteins (Figure 2). Ubl found in *C. subterraneum* shares >30–35% identity with the eukaryotic Ub-ribosomal fusion proteins and Ub B, and harbors the Gly–Gly motif found at the C-terminal region of eukaryotic Ub/Ubl (Figure 2A). As nine residues follow the Gly-Gly motif in the *C. subterraneum* Ubl, this suggests that this organism possesses a post-translational modification system, generally presumed to be a trait of the eukaryotic Ub/Ubl system (79). The *C. subterraneum* E1l retains the second-catalytic-cysteine domain involved in Ub-E1 interaction and the adenylation domains found in eukaryote-type E1s (UBA2, UBA3) (80,81) (Figure 2B). The significant eukaryote-type feature in the *C. subterraneum* E1l is the presence of two insertion helices (Asp_{197}–Ser_{208} and Ile_{224}–Leu_{239}) between the Ub-E1 interaction domain and second Mg^{2+}-chelating domain, which are found only in eukaryote-type E1s such as UBA1, UBA2, UBA3 and Aos1 (15) (Figure 2B). The JAMM (JAB1/MPN/Mov34 metalloenzyme) motif is a highly conserved motif found in various metal proteases from all three domains of life (82). The motif is known to be essential for the de-ubiquitination of captured substrate by RPN11 to facilitate their degradation, and is conserved in the RPN11l found in *C. subterraneum* (83). The *C. subterraneum* protein also possesses a C-terminal extension that forms sheet structures, which is a specific characteristic of the eukaryotic RPN11 proteins associated with the proteasome, and not found in archaeal and bacterial JAMM proteins (84) (Figure 2D). However the *C. subterraneum* protein seemingly lacks the central region of the



```
                                                                                                                             **
S.cere smt3    19 --KPETHINLKV--SD S-SEIFFK----IKKTTPLRRLM--------EAFAKRQGKE--MDSLRFLY-DGIRIQADQT---PEDLDMEDN-DIIEAHREQIGGATY------- 101
Human sumo1    17 --KEGEYIKLKVIGQDS--SEIHFK----VKMTTHLKKLK--------ESYCQRQGVP--MNSLRFLF-EGQRIADNHT---PKELGMEEE-DVIEVYQEQTGGHSTV------ 101
Human sumo2    16 -----DHINLKVAGQDG--SVVQFK----IKRHTPLSKLM--------KAYCER-----------------------------QLEMEDE-DTIDVFQQQTGGVY--------  71
C.mero smt3    18 --SGGDQINLRVRDADG--NEVQFR----IKKHTPLRKLM--------DAYCTRKGVD--LHSYRFLF-DGNRINEDDT---PEKLGMEDM-DSIDAMLFQQGGW---------  99
T.ther ubl1     8 ANANSEYLNLKVKSQEG--EEIFFK----IKKTTQFKKLM--------DAYCQRAQVN--AHNVRFLF-DGDRILESHT---PADLKMESG-DEIDVVVEQVGGSF--------  90
G.lamb sumo    19 KPEQAQKIMIKVSDEHE--NAICFK----VKMTTALSKVF--------DAYCSKNSLQ--RGDVRFYF-NGARVSDTAT---PKSLDMAEN-DIIEVMRNQIGGH--------- 102
Human NEDD8     1 -------MLIKVKTLTG--KEIEID----IEPTDKVERIK--------ERVEEKEGIP--PQQQRLIY-SGKQMNDEKT---AADYKILGG-SVLHLVLALRGGGGLRQ-----  81
C.mero ubl     23 RSEPSETMLVKVKTLTG--KEVELD----IEPHDPIQRIK--------ERIEEKEGIP--PQQQRLIF-GGKQLADDRS---AREYNIEGG-SVLHLVLALRGGHVC------- 108
S.cere Rpl40     1 -------MQIFVKTLTG--KTITLE----VESSDTIDNVK--------SKIQDKEGIP--PDQQRLIF-AGKQLEDGRT---LSDYNIQKE-STLHLVLRLRGGIIEPSLKALA  86
G.lamb Rpl40     1 -------MQLIVRSLDG---TVALT----ASPADSLTSIR--------QRLLAVYSGHV-VDSQRFVF-AGRTLDEAKT---LGDYSIGES-SVLDLVPRLFGGVMEPTLINLA  86
G.lamb ub1      1 --MGGFYMQIFVKTLTG--KTVTLE----VEPTDTINNIK--------AKIQDKEGIP--PDQQRLIF-SGKQLEDNRT---LQDYSIQKD-ATLHLVLRLRGGN---------  82
C.parv ubl1     1 -------MQILVKTLTG--KKQNFN----FEPENTVLQVK--------QALQEKEGID--VKQIRLIY-SGKQMSDDLR---LLDYKVTAG-CTIHMVLQLRGGLR--------  78
T.bruc ub       1 -------MLLKVKTVSN--KVIQITS---LTDDNTIAELK--------GKLEESEGIP--GNMIRLVY-QGKQLEDEKR---LKDYQMSAG-ATFHMVVALRAGC---------  78
G.lamb_ub2      1 -------MLLKVQLTTG--YILTLD----VAPTETILDIK--------NKVYDQEGIH--PAQQKMLY-LAQQLQNTTT---VEEANLKAG-ITIQLVVNLRGG----------  76
CSUB_C1474      1 -------MKIKIVPAVGGGSPLELE----VAPNATVGAVR--------TKVCAMKKLP--PDTTRLTY-KGRALKDTET---LESLGVADG-DKFVLITRTVGGCGEPIRRAA-  87
Human ufm1       1 ----MSKVSFKITLTSD--PRLPYKVLS-VPESTPFTAVL--------KFAAEEFKVP--AATSAIITNDGIGINPAQT---AGNVFLKHG-SELRIIPRDRVGSC--------  85
T.ther ubl2      1 -MATKQKVTFKITLTSD--PNLPFRTIS-VPEEAPFSACI--------KYVAEQFKVN--HATSAIITSTGVGINPEQT---AGNVFLKHG-SELKLIPRDRVGNQ--------  88
C.mero Rps27     1 -----MRRQLLVQCPNG--RIVSTN----VLATDSLAVVL-----------SRVTGLD--ADAVYGTVAGGRPVATLRD--ALVNFTDPEAPIVIQAHVRVLGGGKKRKKKTYT  88
C.parv Rps27     1 ----LSKMQIFFRYGLG--NTRSLE----VDPTMSVKELR--------HIISEFSGIS--IDSQCISYGFG-ILDEFET---LEQAGISDY-STLYVSEAMLGGAKKKKKNFTK  89
C.parv ubl2    15 LAGDRQNVEVNLNNLKSS-SMKSLIL---YVEENIIQYRK--------DHFI-ETGSK--IKPGIIVLVNNCDWEILGG----ENYALSDG-DLVTFIMTLHGG---------- 98
T.bruc ubl     18 LFAKQTSLQLDGVVPTGT-NLNGLVQ---LLKTNYVKERP--------DLLVDQTGQT--LRPGILVLVNSCDAEVVGG----MDYVLNDG-DTVEFISTLHGG---------- 102
CSUB_C0702      1 MAVKVYLPTPLRQYADG-RDMVELDG---STVGEVLNKLVSRYTA-LQKHLFNENGAI---RSFVNVFVNNEDIRFLEG----VNTKIKDG-DVVYIIPSIAGGLSIAAPAAVA 118
CSUB_C1603      5 RLKILTKYYAVLRERVG-KASEEFELPQGSTVIDFLEKLRQVYGG-VLGDLFEGDGL----RTGFALALNGESLDRKLW----ASTRLKDG-DVVVVLPPIAGGYLKLGSLTPR 107
CSUB_C0525     10 MALTVNFYSSYLRRAAG-GETIRLEES--PRTVRELLDLLAAKLGKSFEELVYDPRQK-TLKRAIVLLVNGHSIKMLKGLDTPLHPDDNVSIDTVEVIEVVGGG---------- 109
CSUB_C1012      1 -----------MSEAG---TVKIN----GRDMVCVGKTI--------SQVLVSVGVDP-ARQGIAVAVNGEVVPRSMW----GRVRLKAG-DIVEIVTAVAGG----------  71
HVO_2619 SAMP1 1 ---MEWKLFADLAEVAG-SRTVRVD----VDGDATVGDALDALVGAHPALESRVFGDDGELYDHINVLRNGE--AAALG------EATAAG-DELALFPPVSGG----------  87
HVO_0202 SAMP2 1 --------MNVTVEVVG-EETSEVA----VDDD---GTYA---------DLVRAVDLS--PHEVTVLV-DGRPVPED---------QSVEV-DRVKVLRLIKGG----------  66
B.subt ThiS     1 -----------MLQLNG--KDVKWK-----KDTGTIQDLL------------ASYQLE---NKIVIVERNKEIIGKERY----HEVELCDR-DVIEIVHFVGGG----------  66
S.aver ThiS     1 ----------MNISVNG--ERRRIA------PGTALDTLV-----------KTLTAAP---PSGVAAALNETVVPRAQW----SSTALSEG-DRVEVLTAVQGG----------  66
N.euro ThiS     1 ----------MQLIING--QQQSYD------GPMNVQQLV------------EKLSLQ---NKRFAIERNGEIIPRSRF----PELLLNEG-DQLEIIVAVGGG----------  66
E.coli MoaB     3 LRMINVLFFAQVRELVG-TDATEVA-----ADFPTVEALR-----------QHLAAQS---DRWALALEDGKLLAAVNQTLVSFDHSLTDG-DEVAFFPPVTGG----------  88
P.furi MoaB     9 SVKVKVKFFARFRQLAG-VDEEEIELPEGARVRDLIEEIKKRHE----KFKEEVFGEGYDEDADVNIAVNGRYVS--------WDEELKDG-DVVGVFPPVSGG----------  99
M.acet MoaB     1 -MKIHVKFLATIREITG-KPEIELEILPGDTVGTALQALQARYG--PEFKEATTGTTAGG-IPKVRFLVNGRNTDFLDG----FETELKAG-DVMVFVPPVAGG----------  94
A.arom RnfH     1 ---MPMKIGVAYSEPSH-QVWLNLE----VPDGTTVGAAI--------ERSGILAQFPHIDLTVQKVGVFAKVVK--------LDTPLRHG-DRVEIYRPITCDPKAVRKKADA  89
P.syri RnfH     1 MADASIQIEVVYASVQR-QVLKTVD----VPTGSSVRQAL--------ALSGIDKEFPELDLSQCAVGIFGKVVTDP------AARVLEAG-ERIEIYRLLVADPMEIRRLRAA  94
```

Figure 2. Sequence alignments of Ub, E1, E2 (super-) and JAMM family proteins. (**A**) Sequence alignments of eukaryotic and archaeal Ub superfamily proteins; proteins from *Saccharomyces cerevisiae*; S.cere Smt3 (6320718) and S.cere Rpl40 (6322043), from human; Human sumo2 (54792071), Human sumo1 (54792065), Human NEDD8 (5453760) and Human Ufm1 (7705300), from Cy*anidioschyzon merolae*; C.mero smt3 (CME004C), C.mero ubl (CML042C) and C.mero Rps27 (CMN125C), from *Tetrahymena thermophila*; T.ther ubl1 (229594936) and T.ther ubl2 (118367859), from *Cryptosporidium parvum*; C. parv ubl1 (126654302), C.parv Rps27 (66357428) and C.parv ubl2 (66363058), from *Giardia lamblia*; G.lamb sumo (159114790), G.lamb Epl40 (159108136), G.lamb ub1 (159112981), G.lamb ub2 (159111413), from *Trypanosoma brucei*; T.bruc ub (72387960) and T.bruc ubl (72387818), from *C. subterraneum*; eukaryote-type Ubl (CSUB_C1474) and prokaryote-type Ubls (ThiS/MoaD) (CSUB_C0525, CSUB_C0702, CSUB_C1012, CSUB_C1603), from *H. volcanii*; SAMPs, HVO_0202 (302595884) and HVO_2619 (302595883), from *Bacillus subtilis*; B.sub ThiS (CAB13025), from *Streptomyces avermitilis*; S.aver ThiS (BAC73805), from *Nitrosomonas europaea*; N.euro ThiS (CAD84196), from *Escherichia coli*; E.coli MoaB (AAN79339), from *Pyrococcus furiosus*; P.furi MoaB (1VJK_A), from *Methanosarcina acetivorans*; M.acet MoaB (AAM05120), from *Aromatoleum aromaticum*; A.arom NrfH (CAI07579) and from *Pseudomonas syringae*; P.syri NrfH (AAY39230). Asterisks indicate the C-terminal Gly-Gly motif. (**B**) Sequence alignments of adenylation and catalytic cysteine domains in E1 superfamily proteins; proteins from human; Human E1L (23510338), Human sumoE1 (60594167), Human UBA1 (23510338), Human UBA2 (4885649), Human UBA3 (38045942), Human UBA5 (13376212), Human ATG7 (119584500) and Human MOCS3 (7657339), from S*chizosaccharomyces pombe*; S.pomb E1L (162312305) and S.pomb UBA3 (19113852), from *S. cerevisiae*; S.cere Aos1 (6325438), S.cere UBA1 (6322639), S.cere UBA2 (6320598), S.cere ATG7 (6321965), S.cere UBA4 (6321903) and S.cere YgdLl (6322825), from *T. thermophila*; T.ther E1L (118383519), T.ther E1B (118351055), T.ther UBA4 (118351953) and T.ther YgdLl (118400480), from *Trypanosoma cruzi*; T.cruz E1 (71411317), from *Plasmodium yoelii*; P. yoel UBA2 (82595829) and P.uoel MoeB (83315401), from *Trichomonas vaginalis*; T.vagi APG7 (123446747), from *C. subterraneum*; E1l (CSUB_C1476) and MoeB (CSUB_C1135), from *H. volacanii*; HVO_0558 (292654724), *Cupriavidus metallidurans*; C.meta ThiF (4039868), from *Clostridium perfringens*; C.perf (86559649), from *Shewanella* sp. ANA3; S.ANA3 (117676291), from *Rhizobium etli*; R.etli (86359719), from *Anabaena variabilis*; A.vari (ABA25158), from *Polaromonas naphthalenivorans*; P.naph (121605347), from *Nostoc* sp. PCC7120; Nostoc (BAB77147), from *Xanthomonas axonopodis*; X.axon MoeB (21242767), from *E. coli;* E.coli MoeB (1JW9_B) from *C. symbiosum*; C.symb ThiF (ABK78649), from *P. furiosus*; P.furi MoeB (18977661), from *Geobacillus kaustophilus*; G.kaus MoeBl (56419161), *Desulfuromonas acetoxidans*; D.acet ThiF (95930339), from *Desulfovibrio desulfuricans*; D.desu ThiF (78357502), from *Bacteroides thetaiotaomicron*; B.thet (29349047), from *M. tuberculosis*; M.tube Rv (15609475), from *Cytophaga hutchinsonii*; C.hutc (110639176), and from *Bacillus thuringiensis*; B.thur (110639176). Asterisks and plus indicate adenylation active sites and thiolating cysteine, respectively. Mg^{2+} chelating motifs (CxxC) are shown by octothorpes. (**C**) Alignment of E2 superfamily proteins; proteins from human; Human E2A (32967280), Human E2D (5454146), Human E2N (61175265), Human E2G1 (13489085), Human E2G2 (29893557), Human E2K (163660385), Human E2H (4507783), Human E2M (4507791), Human E2J2 (37577124), Human E2J (37577122) and Human Tsg101 (5454140), from *Arabidopsis thaliana*; A.thal E2I (15230881), A.thal E2C (18403097) and A.thal E2J (18401338), from *Chlamydomonas reinhardtii*; C.rein E2K (159463008), from *C. merolae*; C.mero E2D (CMB015C) and C.mero E2N (CMR010C), from *Plasmodium falciparum*; P.fal E2D (124805463), from *S. cerevisiae*; S.cere E2A (6321380), S.cere E2D (6319556), S.cere E2N (6320297), S.cere E2I (6320139), S.cere E2C (6324915), S.cere E2G2 (6323664), S.cere E2K (6320382), S.cere E2H (6579192), S.cere E2M (6323337) and S.cere E2J2 (6320947), from *S. pombe*; S.pomb E2G1 (6323664), from *T. thermophila*; T.ther E2M (118382495), from *T. vaginalis*; T.vagi E2M (123484378), from *G. lamblia*; G. lamb E2D (159111264), from *C. subterraneum*; CSUB_C1475, from *Ruegeria* sp; Rueger (22726448), from *Arthrobacter* sp.; Arthro (A0AW81), from *E. coli*; E.coli (37927532), from *Syntrophus aciditrophicus*; S.acid (85859492), from *Rhodobacter sphaeroides*; R.spha (77387013), from *Clostridium perfringens*; C.perf (86559649), from *Dechloromonas aromatica*; D.arom (71847775), from *Anabaena variabilis*; A.vari (75705484), from *Bacteroides thetaiotaomicron*; B.thet (29339960), from *Synechocystis* sp. PCC6803; Synech (38423903), from *Burkholderia cepacia*; B.cepa (A4JA91), and from *Rhizobium* sp. NGR234; Rhizob (2496664). Astetisk and octothorpes indicate catalytic cysteine residue and residues forming a conserved stabilizing contact in E2 from eukaryotes, respectively. Flap histidine and asparagine residues are shown by plus. Identical and similar amino acids are shaded in black and gray, respectively. (**D**) Sequence alignment of JAMM family proteins; proteins from human; Human COPS5 (12654695) and Human PSMD14 (5031981), from *A. thaliana*; A.thal CSN5A (15219970), from *S. cerevisiae*; S. cere RPN11 (14318526), from *T. brucei*; T.bruc RPN11 (18463065) and T.bruc SCN5 (72393165), from *G. lamblia*; G.lamb RPN11 (159114272), from *S. pombe*; S.pomb AMSHP (19115685), from *C. subterraneum*; CSUB_C1473, from *Archaeoglobus flugidus*; A.flugi JAB (11499780), from *Pyrococcus horikoshii*; P.hori JAB (3257912), from *Pseudomonas aeruginosa*; P.aeru JAB (15597298), from *Pyrobaculum aerophilum*; Py.aer JAB (18313041), from *E. coli*; E.coli RadC (15801143), from *B. subtilis*; B.subt RadC (16079856), from *M. acetivorans*; M.acet RadC (20090827), from *Thermotoga maritima*; T.mari RadC (15644305), from *Aquifex aeolicus*; A.aeol (2984019); from *Deinococcus radiodurans*; D.radi (15805429), from *Pseudomonas putida*; P.puti (84994017), from *Salinibacter rubber*; S.rubb (83814538), from *M. tuberculosis*; M.tube (13880984), from *Nocardia farcinica*; N.farc (54014564), from *Wolinella succinogenes*; W.succ, and from *Geobacter metallireducens*; G.meta. Asterisks indicate the JAMM motif residues. Identical and similar amino acids are shaded in black and gray, respectively.

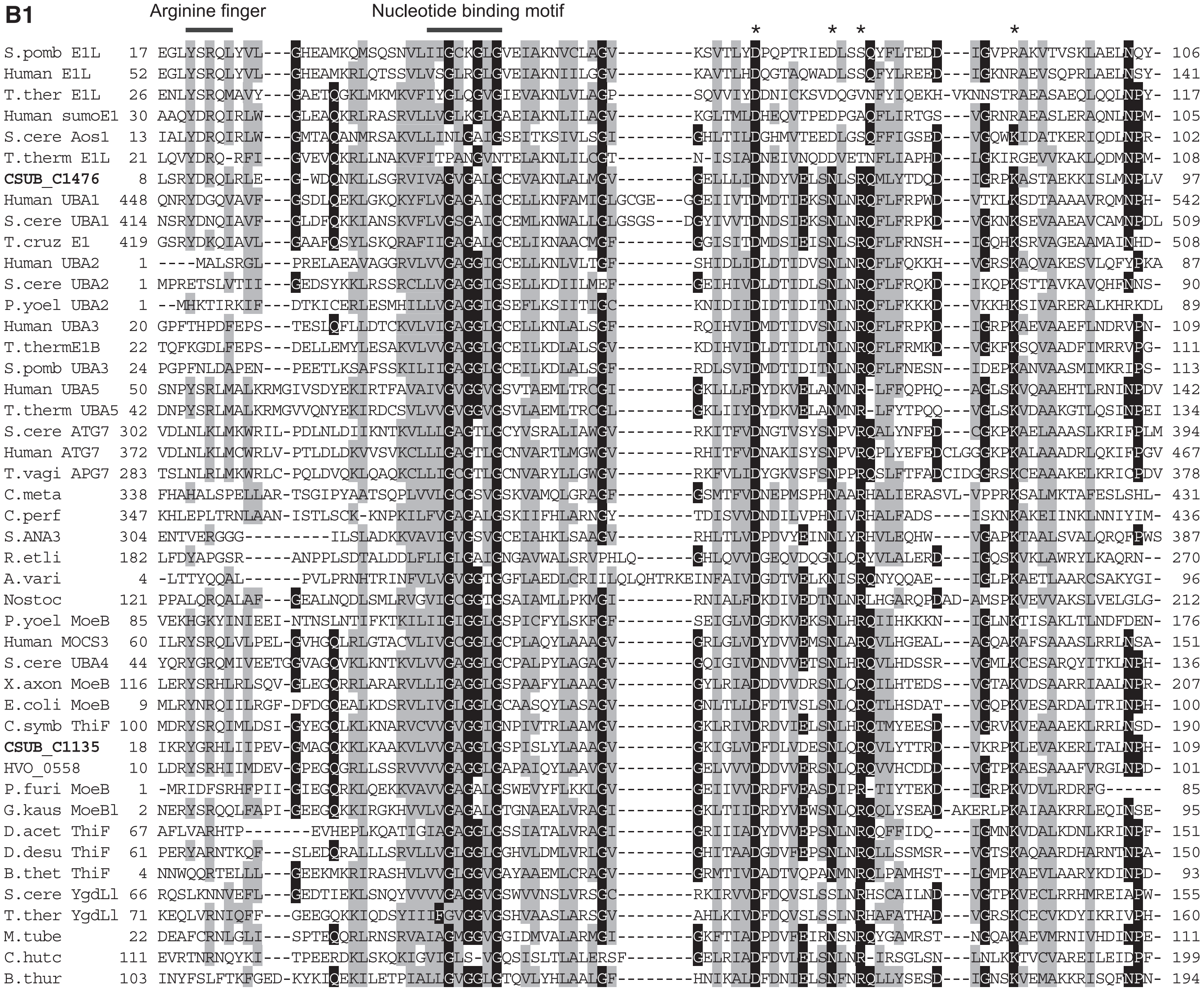


Figure 2. Continued.

eukaryotic RPN11, consisting of ∼55 residues and including one helix.

Phylogenetic analyses

In order to confirm the phylogenetic position of HWCGI, we used the genomic information of *C. subterraneum* along with those from other archaeal complete genome sequences and environmental genome fragments to perform phylogenetic analyses based on (i) concatenated SSU+LSU rRNA genes; (ii) concatenated ribosomal proteins and RNA polymerase subunits; and (iii) translation elongation factor 2 (EFII) (Figure 4). Taken together, all of these phylogenetic analyses demonstrate that *C. subterraneum* forms a robust cluster with the *Thaumarchaeota*, and is distinct from the hyperthermophilic *Crenarchaeota*. The *Korarchaeota* is placed in a deeply branching lineage with affinity to the crenarchaeal cluster in the trees of SSU+LSU rRNA genes and EFII, and occupies the deepest position of the *Archaea* in the tree based on concatenated r-proteins+RNAP subunits sequence. Most orders in the *Euryarchaeota* are sturdily recovered in all of these trees (Figure 4). The phylogenetic positions of *C. subterraneum* based on these multiple gene phylogenetic analyses are consistent with those suggested from previously reported phylogenetic trees including environmental SSU rRNA gene sequences (7,21,22; Supplementary Figure S1). The results appear to conflict with the deep branching of *Thaumarchaeota* as a sister group of all other *Archaea*, and the potential of a mesophilic last archaeal common ancestor (4,8,9).

Furthermore, in order to examine the origin of the ‘euryarchaeal genes’ in the novel creanarchaeal lineages, we performed phylogenetic analyses targeting DNAP, which is a signature of *Euryarchaeota* (47) (Table 1). The phylogenetic tree of concatenated SSU+LSU D-type DNAP presents a robust cluster of crenarchaeal lineages that can be considered as a sister group of the enzymes from *Euryarchaeota* (Figure 5). When the cluster of crenarchaeal sequences was placed as an outgroup of the euryarchaeal sequences, the tree topology does not

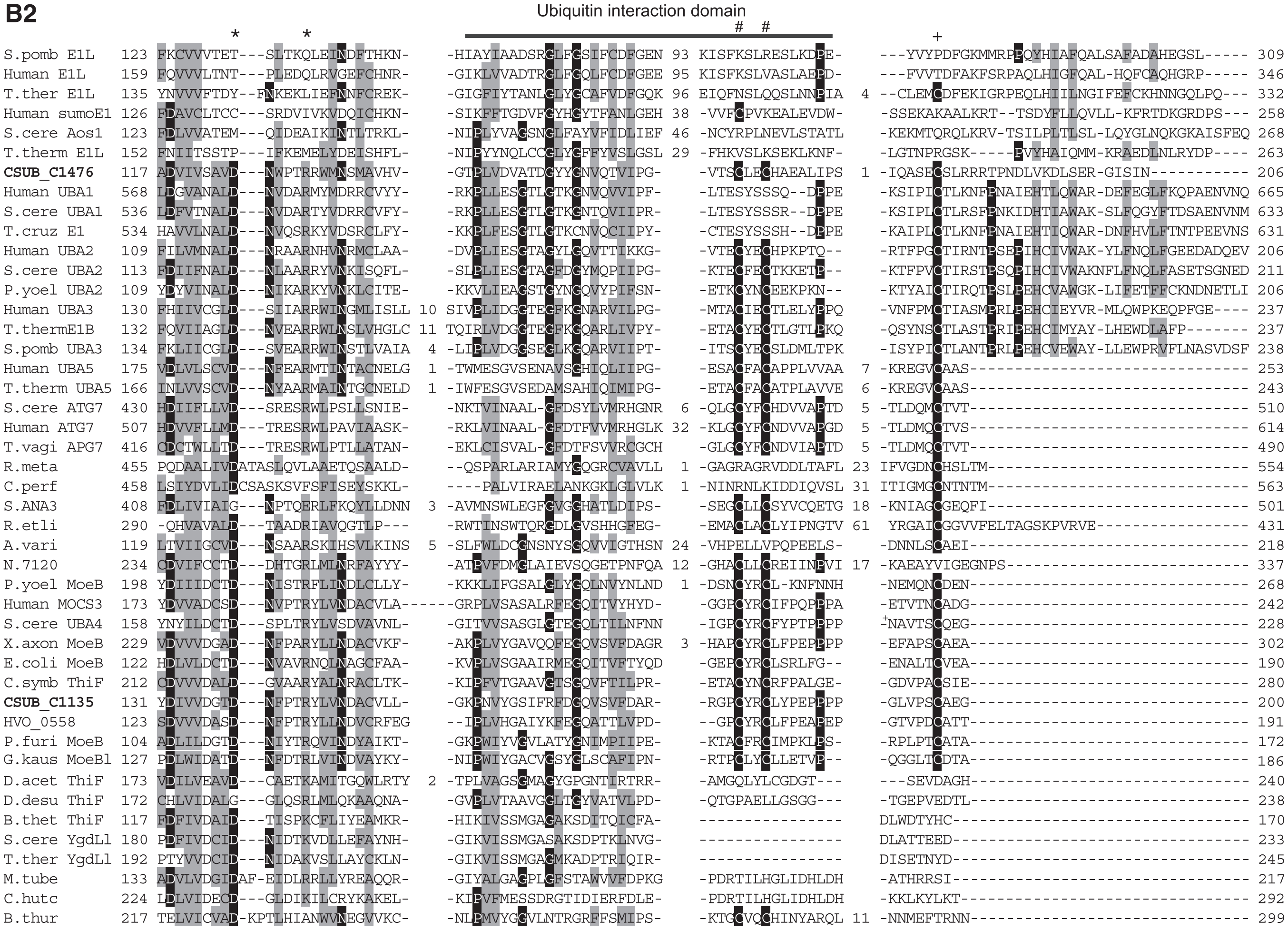


Figure 2. Continued.

contradict with the phylogenetic analyses for rRNA genes, r-proteins+RNAP subunits and EFII (Figures 4 and 5). It can thus be concluded that the D-type DNAPs in the novel crenarchaeal lineages were vertically inherited from the last common archaeal ancestor and did not originate in euryarchaeotes.

Genome core

In order to compare the gene complement among the novel crenarchaeal lineages, *C. subterraneum*, *Thaumarchaeota* and *Korarchaeota*, and to investigate the differences between *C. subterraneum* and hyperthermophilic *Crenarchaeota*, the numbers of arCOGs in these crenarchaeal lineages that are in common with the genome core genes of *Euryarchaeota* (E) and hyperthermophilic *Crenarchaeota* (HC) were examined (Figure 6, Supplementary Table S2). The CDSs in *C. subterraneum* were tentatively assigned to arCOGs based on BLASTP analysis ($<e^{-3}$) targeting the arCOG database (29). In this study, genome cores were defined as follows: (i) genes defined in an arCOG that are represented in all sequenced genomes of one division, but are missing in at least some organisms of the other division (5); (ii) genes present in more than two-thirds of the genomes from one division and absent in the other division (5); and (iii) genes that are present in at least one representative of each order of one division, but are absent from all genomes in the other division (4). When examining the presence of euryarchaeotic or crenarchaeotic genome core genes based on definition (I), *C. subterraneum* (HC:E = 80%:59%) and *Korarchaeota* (HC:E = 81%:22%) apparently show higher affinity with hyperthermophilic *Crenarchaeota* than *Euryarchaeota*, while *Thaumarchaeota* (HC:E = 58%:79%) had a more euryarchaeotic genomic feature (Figure 6, Supplementary Table S2). When the numbers of genome core genes defined by (II) and (III) were compared, we found that all three lineages shared similar euryarchaeotic features, but *Thaumarchaeota* exhibited fewer crenarchaeal features among the three. With definition (III), we found that *C. subterraneum* and *Korarchaeota* demonstrate significant euryarchaeotic features (Figure 6; Supplementary Table S2). Interestingly, only a small number of HC and E genome core genes defined by (II) and (III) are shared among the three novel crenarchaeal lineages (Figure 6; Supplementary Table S2). In addition, we summarized

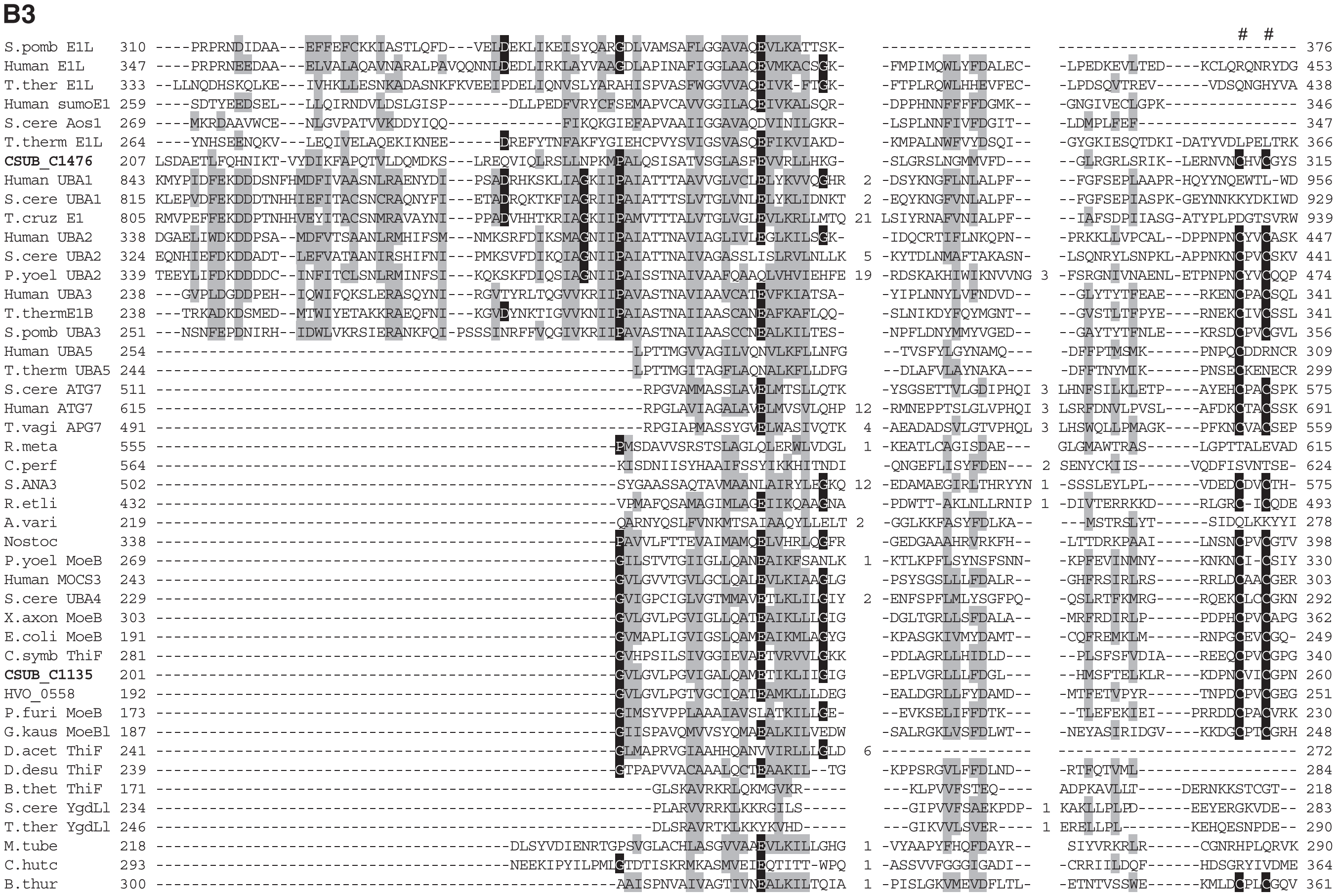


Figure 2. Continued.

the numbers of shared arCOGs among the novel crenarchaeal lineages in order to examine genomic affinity (Figure 6). The numbers of arCOGs shared in two lineages but lost in the other probably reflect the relative affinity among the three lineages. As a result, while a total of 446 arCOGs was shared among the three lineages, *K. cryptofilum* and *C. subterraneum* showed higher affinity (194 arCOGs shared) to one another compared to the affinity between *C. subterraneum* and *Thaumarchaeota* (134 arCOGs shared), and *K. cryptofilum* and *Thaumarchaeota* (78 arCOGs shared).

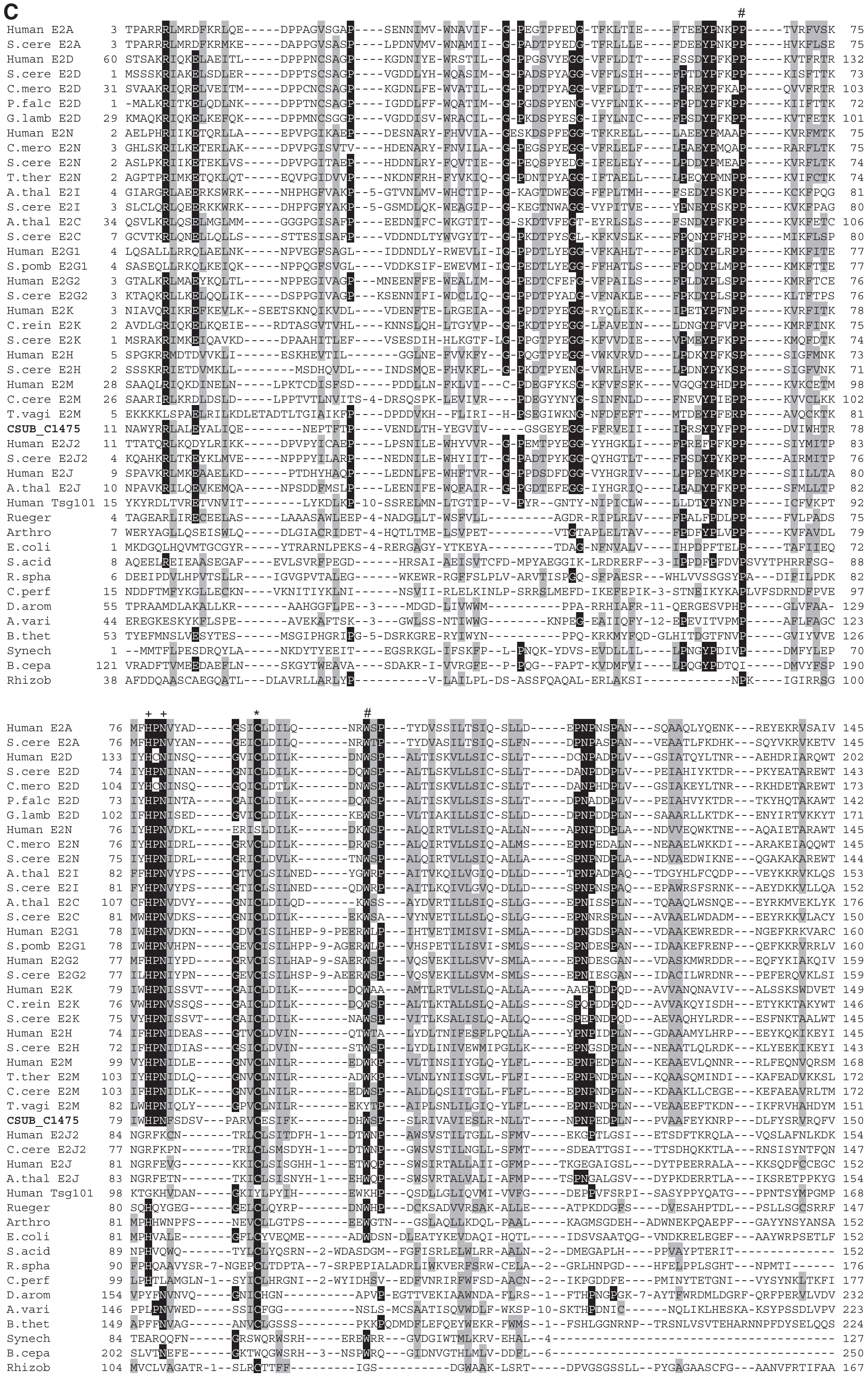
DISCUSSION

Genomic coherence and assembly

The high similarities of overlapping regions and the presence of potential insertion/deletion regions indicate that the composite genome sequence of *C. subterraneum* was successfully assembled from individual, closely related sympatric donor genotypes. The metagenomic library contains DNA from two uncultivated crenarchaeotic lineages; the HWCGI and HWCGIII (Ca. '*Nitrosocaldus*' sp.). The HWCGIII populations were thought to be more abundant than the HWCGI populations in the metagenomic library based on PCR dependent screening for archaeal SSU rRNA genes (21). However, the dot-blot screening in this study indicates that the number of genome fragments encoding SSU rRNA gene from the HWCGI is seven times as much as those from the HWCGIII, and the result is consistent with the successful genome assemblage of *C. subterraneum*. Among the 19 *C. subterraneum* SSU rRNA genes found in the metagenomic library, we observed the co-existence of intron coding or non-coding SSU rRNA genes within one ribotype population. The finding suggests the occurrence of intron transfer events among the *C. subterraneum* populations associated with double-strand break by intron-coding homing endonuclease and homologous recombination for double strand break repair (85).

Metabolism and ecology

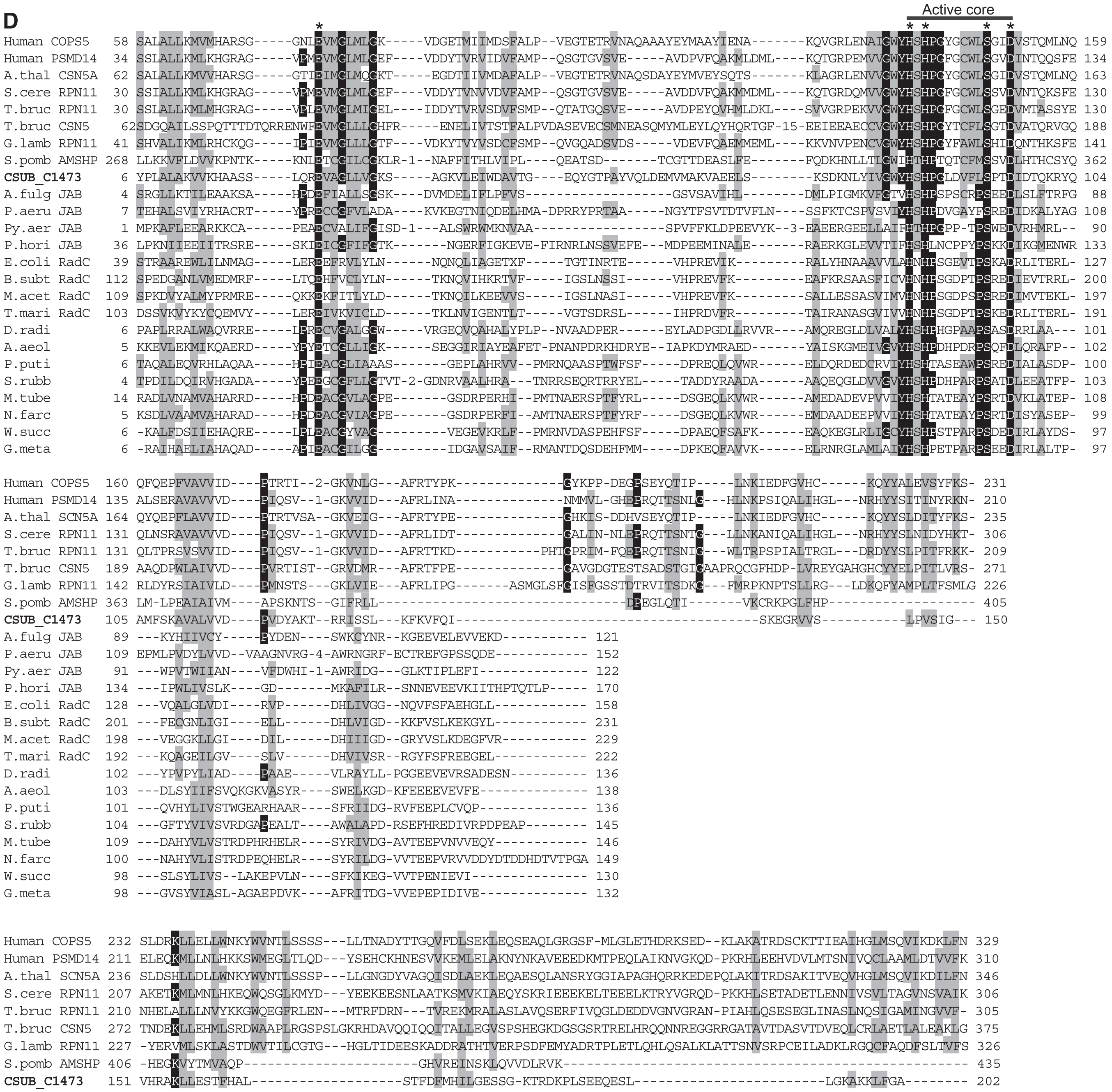
The bacterial communities of microbial mat formation in the geothermal water stream in the subsurface gold mine are dominated by hydrogen-, ammonia- or nitrite-oxidizing chemolithoautotrophs and methanotrophs while hetetrotrophs represent minor populations (86). Considering the high abundance of *C. subterraneum* and bacterial chemolithoautotrophs and methanotophs in the microbial ecosystem, the archaeon also likely displays chemolithoautotrophic metabolism. In fact, the presence of hydrogen up-take hydrogenase and aerobic carbon



```
C                                                                                                        #
Human E2A      3 TPARRRLMRDFKRLQE-----DPPAGVSGAP----SENNIMV-WNAVIF--G-PEGTPFEDG-TFKLTIE----FTEEYPNKPP-----TVRFVSK  75
S.cere E2A     3 TPARRRLMRDFKRMKE-----DAPPGVSASP----LPDNVMV-WNAMII--G-PADTPYEDG-TFRLLLE----FDEEYPNKPP-----HVKFLSE  75
Human E2D     60 STSAKRIQKELAEITL-----DPPPNCSAGP----KGDNIYE-WRSTIL--G-PPGSVYEGG-VFFLDIT----FSSDYPFKPP-----KVTFRTR 132
S.cere E2D     1 MSSSKRIAKELSDLER-----DPPTSCSAGP----VGDDLYH-WQASIM--G-PADSPYAGG-VFFLSIH----FPTDYPFKPP-----KISFTTK  73
C.mero E2D    31 SVAAKRIQRELVEITM-----DPPCNCSAGP----KGDNLFE-WVATIL--G-PPDTPYAGG-VFFLDIT----FPREYPFKAP-----QVVFRTR 103
P.falc E2D     1 -MALKRITKELQDLNK-----DPPTNCSAGP----IGDDLFF-WQATIM--G-PGDSPYENG-VYFLNIK----FPPDYPFKPP-----KIIFTTK  72
G.lamb E2D    29 KMAQKRIQKELKEFQK-----DPPMNCSGGP----VGDDISV-WRACIL--G-PKDSPYESG-IFYLNIC----FPSDYPFKPP-----KVTFETK 101
Human E2N      2 AELPHRIIKETQRLLA-----EPVPGIKAEP----DESNARY-FHVVIA--GESKDSPFEGG-TFKRELL----LAEEYPMAAP-----KVRFMTK  75
C.mero E2N     3 GHLSKRILKETERLAK-----DPVPGISVTV----HDENARY-FNVILA--G-PEGSPYEGG-VFRLELF----LPAEYPMQAP-----RARFLTK  75
S.cere E2N     2 ASLPKRIIKETEKLVS-----DPVPGITAEP----HDDNLRY-FQVTIE--G-PEQSPYEDG-IFELELY----LPDDYPMEAP-----KVRFLTK  74
T.ther E2N     2 AGPTPRIMKETQKLQT-----EQVPGIDVVP----NKDNFRH-FYVKIQ--G-PDNTPYAGG-IFTAELL----LPEQYPMNPP-----KVIFCTK  74
A.thal E2I     4 GIARGRLAEERKSWRK-----NHPHGFVAKP--5-GTVNLMV-WHCTIP--G-KAGTDWEGG-FFPLTMH----FSEDYPSKPP-----KCKFPQG  81
S.cere E2I     3 SLCLQRLQEERKKWRK-----DHPFGFYAKP--5-GSMDLQK-WEAGIP--G-KEGTNWAGG-VYPITVE----YPNEYPSKPP-----KVKFPAG  80
A.thal E2C    34 QSVLKRLQSELMGLMM-----GGGPGISAFP----EEDNIFC-WKGTIT--G-SKDTVFEGT-EYRLSLS----FSNDYPFKPP-----KVKFETC 106
S.cere E2C     7 GCVTKRLQNELLQLLS-----STTESISAFP----VDDNDLTYWVGYIT--G-PKDTPYSGL-KFKVSLK----FPQNYPFHPP-----MIKFLSP  80
Human E2G1     4 LQSALLLRRQLAELNK-----NPVEGFSAGL----IDDNDLY-RWEVLI-IG-PPDTLYEGG-VFKAHLT----FPKDYPLRPP-----KMKFITE  77
S.pomb E2G1    4 SASEQLLRKQLKEIQK-----NPPQGFSVGL----VDDKSIF-EWEVMI-IG-PEDTLYEGG-FFHATLS----FPQDYPLMPP-----KMKFTTE  77
Human E2G2     3 GTALKRLMAEYKQLTL-----NPPEGIVAGP---MNEENFFE-WEALIM--G-PEDTCFEFG-VFPAILS----FPLDYPLSPP-----KMRFTCE  76
S.cere E2G2    3 KTAQKRLLKELQQLIK-----DSPPGIVAGP---KSENNIFI-WDCLIQ--G-PPDTPYADG-VFNAKLE----FPKDYPLSPP-----KLTFTPS  76
Human E2K      3 NIAVQRIKREFKEVLK--SEETSKNQIKVDL----VDENFTE-LRGEIA--G-PPDTPYEGG-RYQLEIK----IPETYPFNPP-----KVRFITK  78
C.rein E2K     2 AVDLGRIQKELKQEIE----RDTASGVTVHL----KNNSLQH-LTGYVP--G-PKDTPYEGG-LFAVEIN----LDNGYPFVPP-----KMRFINK  75
S.cere E2K     1 MSRAKRIMKEIQAVKD-----DPAAHITLEF----VSESDIH-HLKGTF-LG-PPGTPYEGG-KFVVDIE----VPMEYPFKPP-----KMQFDTK  74
Human E2H      5 SPGKRRMDTDVVKLI------ESKHEVTIL-------GGLNE-FVVKFY--G-PQGTPYEGG-VWKVRVD----LPDKYPFKSP-----SIGFMNK  73
S.cere E2H     2 SSSKRRIETDVMKLL--------MSDHQVDL----INDSMQE-FHVKFL--G-PKDTPYENG-VWRLHVE----LPDNYPYKSP-----SIGFVNK  71
Human E2M     28 SAAQLRIQKDINELN-----LPKTCDISFSD-----PDDLLN-FKLVI---C-PDEGFYKSG-KFVFSFK----VGQGYPHDPP-----KVKCETM  98
C.cere E2M    26 SAARIRLKRDLDSLD-----LPPTVTLNVITS-4-DRSQSPK-LEVIVR----PDEGYYNYG-SINFNLD----FNEVYPIEPP-----KVVCLKK 102
T.vagi E2M     5 EKKKKLSPAELRILKDLETADTLTGIAIKFP----DPDDVKH--FLIRI--H-PSEGIWKNG-NFDFEFT----MTDEYPFERP-----AVQCKTK  81
CSUB_C1475    11 NAWYRRLALEYALIQE--------NEPTFTP----VENDLTH-YEGVIV-----GSGEYEGG-FFRVEII----IPRSYPYFPP-----DVIWHTR  78
Human E2J2    11 TTATQRLKQDYLRIKK-----DPVPYICAEP----LPSNILE-WHYVVR--G-PEMTPYEGG-YYHGKLI----FPREFPFKPP-----SIYMITP  83
S.cere E2J2    4 KQAHKRLTKEYKLMVE-----NPPPYILARP----NEDNILE-WHYIIT--G-PADTPYKGG-QYHGTLT----FPSDYPYKPP-----AIRMITP  76
Human E2J      9 SPAVKRLMKEAAELKD------PTDHYHAQP----LEDNLFE-WHFTVR--G-PPDSDFDGG-VYHGRIV----LPPEYPMKPP-----SIILLTA  80
A.thal E2J    10 NPAVKRILQEVKEMQA-----NPSDDFMSLP----LEENIFE-WQFAIR--G-PGDTEFEGG-IYHGRIQ----LPADYPFKPP-----SFMLLTP  82
Human Tsg101  15 YKYRDLTVRETVNVIT--------LYKDLKP-10-SSRELMN-LTGTIP--V-PYR--GNTY-NIPICLW----LLDTYPYNPP-----ICFVKPT  92
Rueger         4 TAGEARLIRECEELAS-----LAAASAWLEEP-4-NADGLLT-WSFVLL---------AGDR-RIPLRLV----FPALFPDLPP-----FVLPADS  75
Arthro         7 WERYAGLLQSEISWLQ-----DLGIACRIDET-4-HQTLTME-LSVPET---------VTGTAPLELTAV----FPDFYPLVPP-----KVFAVDL  79
E.coli         1 MKDGQLHQVMTGCGYR-----YTRARNLPEKS-4-RERGAGY-YTKEYA---------TDAG-NFNVALV----IHPDPFTELP-----TAFIIEQ  72
S.acid         8 AQEELREIEAASEGAF----EVLSVRFPEGD------HRSAI-AEISVTCFD-MPYAEGGIK-LRDRERF--3-IPPDFPFDVPSVYTPHRRFSG-  88
R.spha         6 DEEIPDVLHPVTSLLR----IGVGPVTALEG------WKEWR-RGFFSLPLV-ARVTISPGQ-SFPAESR---WHLVVSSGSYPA---DIFILPDK  82
C.perf        15 NDDFTMFYKGLLECKN----VKNITIYKLNI------NSVII-RLELKINLP-SRRSLMEFD-IKEFEPIK-3-STNEIKYKAPLVFSDRNDFPVE  97
D.arom        55 TPRAAMDLAKALLKR-------AAHGGFLPE--3----MDGD-LIVWWM---------PPA-RRHIAFR-11-QERGESVPHP-----GLVFAA- 129
A.vari        44 EREGKESKYKFLSPE-------AVEKAFTSK--3---GWLSS-NTIWWG--------KNPEG-EAIIQFY-12-EPEVITVPMP-----AFLFAGC 123
B.thet        53 TYEFMNSLVESYTES------MSGIPHGRIPG-5-DSRKGRE-RYIWYN---------PPQ-KRKMYFQD-GLHITDGTFNVP-----GVIYVVE 126
Synech         1 --MMTFLPESDRQYLA-----NKDYTYEEIT----EGSRKGL-IFSKFP--L-PNQK-YDVS-EVDLLIL----LPNGYPDIVP-----DMFYLEP  70
B.cepa       121 VRADFTVMEEDAEFLN-----SKGYTWEAVA----SDAKR-I-VVRGFE--P-PQG--FAPT-KVDMFVI----LPQGYPDTQI-----DMVYFSP 190
Rhizob        38 AFDDQAASCAEGQATL---DLAVRLLARLYP----------V-LAILPL-DS-ASSFQAQAL-ERLAKSI-----------NPK----IGIRRSG 100

                   + +          *             #
Human E2A     76 MFHPNVYAD-----GSICLDILQ-------NRWSP---TYDVSSILTSIQ-SLLD-----EPNPNSPAN---SQAAQLYQENK---REYEKRVSAIV 145
S.cere E2A    76 MFHPNVYAN-----GEICLDILQ-------NRWTP---TYDVASILTSIQ-SLFN-----DPNPASPAN---VEAATLFKDHK---SQYVKRVKETV 145
Human E2D    133 IYHCNINSQ-----GVICLDILK-------DNWSP---ALTISKVLLSIC-SLLT-----DCNPADPLV---GSIATQYLTNR---AEHDRIARQWT 202
S.cere E2D    74 IYHPNINAN-----GNICLDILK-------DQWSP---ALTLSKVLLSIC-SLLT-----DANPDDPLV---PEIAHIYKTDR---PKYEATAREWT 143
C.mero E2D   104 IYHCNINSQ-----GQICLDTLK-------DNWSP---ALTISKVLLSIC-SLLT-----DANPHDPLV---GSIAKEYLTNR---RKHDETAREWT 173
P.falc E2D    73 IYHPNINTA-----GAICLDILK-------DQWSP---ALTISKVLLSIS-SLLT-----DPNADDPLV---PEIAHVYKTDR---TKYHQTAKAWT 142
G.lamb E2D   102 IFHPNISED-----GVICLDILK-------KEWSP---VLTIAKILLSIC-SLLD-----DPNPDDPLN---SAAARLLKTDK---ENYIRTVKKYT 171
Human E2N     76 IYHPNVDKL-----ERISLDILK-------DKWSP---ALQIRTVLLSIQ-ALLN-----APNPDDPLA---NDVVEQWKTNE---AQAIETARAWT 145
C.mero E2N    76 IYHPNIDRL-----GRVCLDILK-------DRWSP---ALQIRTVLLSIQ-ALMS-----EPNPEDALN---NEAAELWKKDI---ARAKEIAQQWT 145
S.cere E2N    75 IYHPNIDRL-----GRICLDVLK-------TNWSP---ALQIRTVLLSIQ-ALLA-----SPNPNDPLA---NDVAEDWIKNE---QGAKAKAREWT 144
A.thal E2I    82 FFHPNVYPS-----GTVCLSILNED-----YGWRP---AITVKQILVGIQ-DLLD-----TPNPADPAQ---TDGYHLFCQDP---VEYKKRVKLQS 153
S.cere E2I    81 FYHPNVYPS-----GTICLSILNED-----QDWRP---AITLKQIVLGVQ-DLLD-----SPNPNSPAQ---EPAWRSFSRNK---AEYDKKVLLQA 152
A.thal E2C   107 CFHPNVDVY-----GNICLDILQD-------KWSS---AYDVRTILLSIQ-SLLG-----EPNISSPLN---TQAAQLWSNQE---EYRKMVEKLYK 176
S.cere E2C    81 MWHPNVDKS-----GNICLDILK-------EKWSA---VYNVETILLSLQ-SLLG-----EPNNRSPLN---AVAAELWDADM---EEYRKKVLACY 150
Human E2G1    78 IWHPNVDKN-----GDVCISILHEP-9-PEERWLP---IHTVETIMISVI-SMLA-----DPNGDSPAN---VDAAKEWREDR---NGEFKRKVARC 160
S.pomb E2G1   78 IWHPNVHPN-----GEVCISILHPP-9-AGERWLP---VHSPETILISVI-SMLS-----SPNDESPAN---IDAAKEFRENP---QEFKKRVRRLV 160
Human E2G2    77 MFHPNIYPD-----GRVCISILHAP-9-SAERWSP---VQSVEKILLSVV-SMLA-----EPNDESGAN---VDASKMWRDDR---EQFYKIAKQIV 159
S.cere E2G2   77 ILHPNIYPN-----GEVCISILHSP-9-AEERWSP---VQSVEKILLSVM-SMLS-----EPNIESGAN---IDACILWRDNR---PEFERQVKLSI 159
Human E2K     79 IWHPNISSVT----GAICLDILK-------DQWAA---AMTLRTVLLSLQ-ALLA-----AAEPDDPQD---AVVANQNAVIV---ALSSKSWDVET 149
C.rein E2K    76 VWHPNVSSQS----GAICLDILK-------DQWSP---ALTLKTALLSLQ-ALLS-----SPQPDDPQD---AVVAKQYISDH---ETYKKTAKYWT 146
S.cere E2K    75 VYHPNISSVT----GAICLDILK-------NAWSP---VITLKSALISLQ-ALLQ-----SPEPNDPQD---AEVAQHYLRDR---ESFNKTAALWT 145
Human E2H     74 IFHPNIDEAS----GTVCLDVIN-------QTWTA---LYDLTNIFESFLPQLLA-----YPNPIDPLN---GDAAAMYLHRP---EEYKQKIKEYI 145
S.cere E2H    72 IFHPNIDIAS----GSICLDVIN-------STWSP---LYDLINIVEWMIPGLLK-----EPNGSDPLN---NEAATLQLRDK---KLYEEKIKEYI 143
Human E2M     99 VYHPNIDLE-----GNVCLNILR-------EDWKP---VLTINSIIYGLQ-YLFL-----EPNPEDPLN---KEAAEVLQNNR---RLFEQNVQRSM 168
T.ther E2M   103 IYHPNIDLQ-----GNVCLNILR-------ADWKP---VLNLYNIISGVL-FLFI-----EPNPNDPLN---KQAASQMINDI---KAFEADVKKSL 172
C.cere E2M   103 IFHPNIDLK-----GNVCLNILR-------EDWSP---ALDLQSIITGLL-FLFL-----EPNPNDPLN---KDAAKLLCEGE---KEFAEAVRLTM 172
T.vagi E2M    82 LWHPNIQLY-----GPVCLNILR-------EKYTP---AIPLSNLILGIQ-YLFL-----EPNPNDPLN---VEAAEEFTKDN---IKFRVHAHDYM 151
CSUB_C1475    79 IWHPNFSDSV---PARVCESIFK-------DHWSP---SLRIVAVIESLR-NLLT-----NPNPEDPLN---PVAAFEYKNRP---DLFYSRVRQFV 150
Human E2J2    84 NGRFKCN-------TRLCLSITDFH-1---DTWNP---AWSVSTILTGLL-SFMV-----EKGPTLGSI---ETSDFTKRQLA---VQSLAFNLKDK 154
C.cere E2J2   77 NGRFKPN-------TRLCLSMSDYH-1---DTWNP---GWSVSTILNGLL-SFMT-----SDEATTGSI---TTSDHQKKTLA---RNSISYNTFQN 147
Human E2J     81 NGRFEVG-------KKICLSISGHH-1---ETWQP---SWSIRTALLAII-GFMP----TKGEGAIGSL---DYTPEERRALA---KKSQDFCCEGC 152
A.thal E2J    83 NGRFETN-------TKICLSISNYH-1---EHWQP---SWSVRTALVALI-AFMP----TSPNGALGSV---DYPKDERRTLA---IKSRETPPKYG 154
Human Tsg101  98 KTGKHVDAN-----GKIYLPYIH-------EWKHP---QSDLLGLIQVMI-VVFG-----DEPPVFSRPI--SASYPPYQATG---PPNTSYMPGMP 168
Rueger        80 SQHQYGEG------GELCLQYRP-------DNWHP---DCKSADVVRSAK-ALLE-----ATPKDDGFS----DVESAHPTDL---PSLLSGCSRRF 147
Arthro        81 MPHHWNPFS-----NEVCLLGTPS------EEWGTN---GSLAQLLKDQL-PAAL-----KAGMSGDEH--ADWNEKPQAEPF---GAYYNSYANSA 152
E.coli        81 MPHVALE-------GFLCYVEQME------ADWDSN-DLEATYKEVDAQI-HQTL-----IDSVSAATQG--VNDKRELEGEF---AAYWRPSETLF 152
S.acid        89 NPHVQWQ-------TYLCLYQSRN--2-WDASDGM--FGFISRLELWLRR-AALN--2--DMEGAPLH----PPVAYPTERIT------------- 152
R.spha        90 FPHQAAVYSR-7-NGEPCLTDPTA--7-SRPEPIALADRLIWKVERFSRW-CELA--2--GRLHNPGD----HFELPPLSGHT---NPMTI------ 176
C.perf        99 LPHTLAMGLN-1--SYICLHRGNI--2-WYIDHSV--EDFVNRIRFWFSD-AACN--2--IKPGDDFE----PMINYTETGNI---VYSYNKLTKFI 177
D.arom       154 VPYFNVNVQ-----GNICHGN---------APVP-EGTTVEKIAAWNDA-FLRS--1--FTHPNGPGK-7-AYTFWRDMLDGRF-QRFPERVLVDV 232
A.vari       146 PPLPNVWED-----SSICFGG---------NSLS-MCSAATISQVWDLF-WKSP-10-SKTHPDNIC------NQLIKLHESKA-KSYPSSDLVPV 223
B.thet       149 APFFNVAG------ANVCLGSSS--------PKKPQDMDFLEFQEYWEKR-FWMS--1--FSHLGGNRNP--TRSNLVSVTEHARNNPFDYSELQQS 224
Synech        84 TEARQQFN------GRSWQRWSRH-----EREWRR---GVDGIWTMLKRV-EHAL--4---------------------------------------- 127
B.cepa       202 SLVTNEFE------GKTWQGWSRH--3--NSPWRQ---GIDNVGTHLMLV-DDFL--6---------------------------------------- 250
Rhizob       104 MVCLVAGATR-1--SLRCTTFF---------IGS---------DGWAAK-LSRT-----DPVGSGSSLL--PYGAGAASCFG---AANVFRTIFAA 167
```

Figure 2. Continued.



```
Human COPS5   58 SALALLKMVMHARSG-----GNLEVMGLMLGK-----VDGETMIIMDSFALP--VEGTETRVNAQAAAYEYMAAYIENA-------KQVGRLENAIGWYHSHPGYGCWLSGIDVSTQMLNQ 159
Human PSMD14  34 SSLALLKMLKHGRAG-----VPMEVMGLMLGEF----VDDYTVRVIDVFAMP--QSGTGVSVE-----AVDPVFQAKMLDML----KQTGRPEMVVGWYHSHPGFGCWLSGVDINTQQSFE 134
A.thal CSN5A  62 SALALLKMVVHARSG-----GTIEIMGLMQGKT-----EGDTIIVMDAFALP--VEGTETRVNAQSDAYEYMVEYSQTS-------KLAGRLENVVGWYHSHPGYGCWLSGIDVSTQMLNQ 163
S.cere RPN11  30 SSIALLKMLKHGRAG-----VPMEVMGLMLGEF----VDDYTVNVVDVFAMP--QSGTGVSVE-----AVDDVFQAKMMDML----KQTGRDQMVVGWYHSHPGFGCWLSSVDVNTQKSFE 130
T.bruc RPN11  30 SSLALLKMLMHGRAG-----VPLEVMGLMIGEL----IDDYTVRVSDVFSMP--QTATGQSVE-----AVDPEYQVHMLDKL----SVVGRPEKVVGWYHSHPGFGCWLSGEDVMTASSYE 130
T.bruc CSN5     62SDGQAILSSPQTTTDTQRRENWFEVMGLLLGHFR-----ENELIVTSTFALPVDASEVECSMNEASQMYMLEYLQYHQRTGF-15-EEIEEAECCVGWYHSHPGYTCFLSGTDVATQRVGQ 188
G.lamb RPN11  41 SHVALIKMLRHCKQG-----IPIEVMGLLLGTF----VDKYTVYVSDCFSMP--QVGQADSVDS-----VDEVFQAEMMEML----KKVNVPENCVGWYHSHPGYFAWLSHIDQNTHKSFE 141
S.pomb AMSHP 268 LLKKVFLDVVKPNTK-----KNLETCGILCGKLR-1--NAFFITHLVIPL----QEATSD-------TCGTTDEASLFE-------FQDKHNLLTLGWIHTHPTQTCFMSSVDLHTHCSYQ 362
CSUB_C1473     6 YPLALAKVVKHAASS-----LQREVAGLLVGKS-----AGKVLEIWDAVTG----EQYGTPAYVQLDEMVMAKVAEELS--------KSDKNLYIVGWYHSHPGLDVFLSPTDIDTQKRYQ 104
A.fulg JAB      4 SRGLLKTILEAAKSA-----HPDEFIALLSGSK----DVMDELIFLPFVS------------------GSVSAVIHL---------DMLPIGMKVFGTVHSHPSPSCRPSEEDLSLFTRFG  88
P.aeru JAB      7 TEHALSVIYRHACRT-----YPRECCGFVLADA----KVKEGTNIQDELHMA-DPRRYPRTAA-----NGYTFSVTDTVFLN----SSFKTCSPVSVIYHSHPDVGAYFSREDIDKALYAG 108
Py.aer JAB      1 MPKAFLEEARKKCA------PEAECVALIFGISD-1---ALSWRWMKNVAA-----------------SPVFFKLDPEEVYK--3-EAEERGEELLAIFHTHPGPP-TPSWEDVRHMRL--  90
P.hori JAB     36 LPKNIIEEIITRSRE-----SKIEICGFIFGTK-------NGERFIGKEVE-FIRNRLNSSVEFE---MDPEEMINALE-------RAERKGLEVVTIFHSHLNCPPYPSKKDIKGMENWR 133
E.coli RadC    39 STRAAREWLILNMAG-----LEREEFRVLYLN------NQNQLIAGETXF-----TGTINRTE------VHPREVIK---------RALYHNAAAVVLAHNHPSGEVTPSKADRLITERL- 127
B.subt RadC   112 SPEDGANLVMEDMRF-----LTQEHFVCLYLN------TKNQVIHKRTVF-----IGSLNSSI------VHPREVFK---------EAFKRSAASFICVHNHPSGDPTPSREDIEVTRRL- 200
M.acet RadC   109 SPKDVYALMYPRMRE-----QKKEKFITLYLD------TKNQILKEEVVS-----IGSLNASI------VHPREVFK---------SALLESSASVIMVHNHPSGDPSPSREDIMVTEKL- 197
T.mari RadC   103 DSSVKVYKYCQEMVY-----LEREIVKVICLD------TKLNVIGENTLT-----VGTSDRSL------IHPRDVFR---------TAIRANASGVIVVHNHPSGDPTPSKEDRLITERL- 191
D.radi         6 PAPLRRALWAQVRRE-----LPRECVGALGGW-----VRGEQVQAHALYPLP--NVAADPER------EYLADPGDLLRVVR----AMQREGLDLVALYHSHPHGPAAPSASDRRLAA--- 101
A.aeol         5 KKEVLEKMIKQAERD-----YPYETCGLLIGK------SEGGIRIAYEAFET-PNANPDRKHDRYE--IAPKDYMRAED-------YAISKGMEIVGVYHSHPDHPDRPSQFDLQRAFP-- 102
P.puti         6 TAQALEQVRHLAQAA-----HPIEACGLIAAAS-------GEPLAHRVV--PMRNQAASPTWFSF-----DPREQLQVWR------ELDQRDEDCRVIYHSHTASEAWPSREDIALASDP- 100
S.rubb         4 TPDILDQIRVHGADA-----YPEEGCGFLLGTVT-2-GDNRVAALHRA---TNRRSEQRTRRYEL-----TADDYRAADA------AAQEQGLDVVGVYHSHPDHPARPSATDLEEATFP- 103
M.tube        14 RADLVNAMVAHARRD-----HPDEACGVLAGPE-------GSDRPERHI--PMTNAERSPTFYRL-----DSGEQLKVWR------AMEDADEVPVVIYHSHTATEAYPSRTDVKLATEP- 108
N.farc         5 KSDLVAAMVAHARAD-----HPDEACGVIAGPE-------GSDRPERFI--AMTNAERSPTFYRF-----DSGEQLKVWR------EMDAADEEPVVIYHSHTATEAYPSRTDISYASEP-  99
W.succ         6 -KALFDSIIEHAQRE-----LPLEACGYVAG---------VEGEVKRLF--PMRNVDASPEHFSF-----DPAEQFSAFK------EAQKEGLRLIGCYHSHPSTPARPSDEDIRLAYDS-  97
G.meta         6 -RAIHAELIAHAQAD-----APIEACGILGG---------IDGAVSAIF--RMANTDQSDEHFMM-----DPKEQFAVVK------ELRNRGLAMLAIYHSHPETPARPSEEDIRLALTP-  97

Human COPS5  160 QFQEPFVAVVID----PTRTI--2-GKVNLG---AFRTYPK--------------GYKPP-DEGPSEYQTIP-----LNKIEDFGVHC------KQYYALEVSYFKS- 231
Human PSMD14 135 ALSERAVAVVVD----PIQSV--1-GKVVID---AFRLINA--------------NMMVL-GHEPRQTTSNLG---HLNKPSIQALIHGL---NRHYYSITINYRKN- 210
A.thal SCN5A 164 QYQEPFLAVVID----PTRTVSA--GKVEIG---AFRTYPE--------------GHKIS-DDHVSEYQTIP-----LNKIEDFGVHC------KQYYSLDITYFKS- 235
S.cere RPN11 131 QLNSRAVAVVVD----PIQSV--1-GKVVID---AFRLIDT--------------GALIN-NLEPRQTTSNTG---LLNKANIQALIHGL---NRHYYSLNIDYHKT- 306
T.bruc RPN11 131 QLTPRSVSVVID----PIQSV--1-GKVVID---AFRTTKD-----------PHTGPRIM-FQEPRQTTSNIG---WLTRPSPIALTRGL---DRDYYSLPITFRKK- 209
T.bruc CSN5  189 AAQDPWLAIVVD----PVRTIST--GRVDMR---AFRTFPE--------------GAVGDGTESTSADSTGIGAAPRQCGFHDP-LVREYGAHGHCYYELPITLVRS- 271
G.lamb RPN11 142 RLDYRSIAIVLD----PMNSTS---GKLVIE---AFRLIPG-------ASMGLSFGISFGSSTDTRVITSDKG---FMRPKNPTSLLRG---LDKQFYAMPLTFSMLG 226
S.pomb AMSHP 363 LM-LPEAIAIVM----APSKNTS--GIFRLL-------------------------------DPEGLQTI-------VKCRKPGLFHP------------------ 405
CSUB_C1473   105 AMFSKAVALVVD----PVDYAKT--RRISSL---KFKVFQI--------------------------------------SKEGRVVS-----------LPVSIG--- 150
A.fulg JAB    89 ---KYHIIVCY-----PYDEN----SWKCYNR--KGEEVELEVVEKD----------- 121
P.aeru JAB   109 EPMLPVDYLVVD--VAAGNVRG-4-AWRNGRF-ECTREFGPSSQDE------------ 152
Py.aer JAB    91 ---WPVTWIIAN----VFDWHI-1-AWRIDG---GLKTIPLEFI-------------- 122
P.hori JAB   134 ---IPWLIVSLK----GD-------MKAFILR--SNNEVEEVKIITHPTQTLP----- 170
E.coli RadC  128 ---VQALGLVDI----RVP------DHLIVGG--NQVFSFAEHGLL------------ 158
B.subt RadC  201 ---FECGNLIGI----ELL------DHLVIGD--KKFVSLKEKGYL------------ 231
M.acet RadC  198 ---VEGGKLLGI----DIL------DHIIIGD--GRYVSLKDEGFVR----------- 229
T.mari RadC  192 ---KQAGEILGV----SLV------DHVIVSR--RGYFSFREEGEL------------ 222
D.radi       102 ---YPVPYLIAD----PAAE-----VLRAYLL-PGGEEVEVRSADESN---------- 136
A.aeol       103 ---DLSYIIFSVQKGKVASYR----SWELKGD-KFEEEEVEVFE-------------- 138
P.puti       101 ---QVHYLIVSTWGEARHAAR----SFRIIDG-RVFEEPLCVQP-------------- 136
S.rubb       104 ---GFTYVIVSVRDGAPEALT----AWALAPD-RSEFHREDIVRPDPEAP-------- 145
M.tube       109 ---DAHYVLVSTRDPHRHELR----SYRIVDG-AVTEEPVNVVEQY------------ 146
N.farc       100 ---NAHYVLISTRDPEQHELR----SYRILDG-VVTEEPVRVVDDYDTDDHDTVTPGA 149
W.succ        98 ---SLSYLIVS--LAKEPVLN----SFKIKEG-VVTPENIEVI--------------- 130
G.meta        98 ---GVSYVIASL-AGAEPDVK----AFRITDG-VVEPEPIDIVE-------------- 132

Human COPS5  232 SLDRKLLELLWNKYWVNTLSSSS---LLTNADYTTGQVFDLSEKLEQSEAQLGRGSF-MLGLETHDRKSED--KLAKATRDSCKTTIEAIHGLMSQVIKDKLFN 329
Human PSMD14 211 ELEQKMLLNLHKKSWMEGLTLQD---YSEHCKHNESVVKEMLELAKNYNKAVEEEDKMTPEQLAIKNVGKQD-PKRHLEEHVDVLMTSNIVQCLAAMLDTVVFK 310
A.thal SCN5A 236 SLDSHLLDLLWNKYWVNTLSSSP---LLGNGDYVAGQISDLAEKLEQAESQLANSRYGGIAPAGHQRRKEDEPQLAKITRDSAKITVEQVHGLMSQVIKDILFN 346
S.cere RPN11 207 AKETKMLMNLHKEQWQSGLKMYD---YEEKEESNLAATKSMVKIAEQYSKRIEEEKELTEEELKTRYVGRQD-PKKHLSETADETLENNIVSVLTAGVNSVAIK 306
T.bruc RPN11 210 NHELALLLNVYKKGWQEGFRLEN---MTRFDRN---TVREKMRALASLAVQSERFIVQGLDEDDVGNVGRAN-PIAHLQSESEGLINASLNQSIGAMINGVVF- 305
T.bruc CSN5  272 TNDEKLLEHMLSRDWAAPLRGSPSLGKRHDAVQQIQQITALLEGVSPSHEGKDGSGSRTRELHRQQNNREGGRRGATAVTDASVTDVEQLCRLAETLALEAKLG 375
G.lamb RPN11 227 -YERVMLSKLASTDWVTILCGTG---HGLTIDEESKADDRATHTVERPSDFEMYADRTPLETLQHLQSALKLATTSNVSRPCEILADKLRGQCFAQDFSLTVFS 326
S.pomb AMSHP 406 -HEGKVYTMVAQP---------------------GHVREINSKLQVVDLRVK------------------------------------------------ 435
CSUB_C1473   151 VHRAKLLESTFHAL-------------------STFDFMHILGESSG-KTRDKPLSEEQESL-----------------------LGKAKKLFGA-------- 202
```

Figure 2. Continued.

monoxide dehydrogenase implies the capability of chemolithotrophic metabolism in *C. subterraneum*. However, we cannot assert the metabolism because of several uncertainties in the function of these enzymes as described above. On the other hand, dicarboxylate/4-hydroxybutyrate cycle is the most likely carbon assimilation pathway though one key enzyme, 4-hydroxybutyryl-CoA dehydratase, is missing. This resembles the situation of *Pyrobaculum arsenaticum*, which is known to exhibit autotrophic growth with the dicarboxylate/4-hydroxybutyrate cycle but does not harbor a 4-hydroxybutyryl-CoA dehydratase gene on its

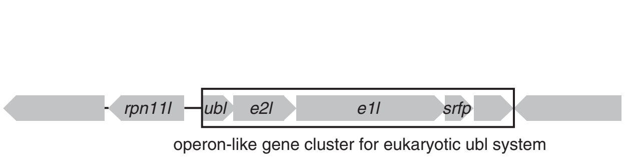


Figure 3. The gene cluster of the Ub-like protein modifier system in *C. subterraneum*. CDSs without gene annotation encode hypothetical proteins. CDSs; *rpn11l* (CSUB_C1473), *ubl* (CSUB_C1474), *e2l* (CSUB_C1475), *e1l* (CSUB_C1476) and *srfp* (CSUB_C1477) encode eukaryotic RPN11, Ubl, E2l and E1l and small RING finger protein, respectively.

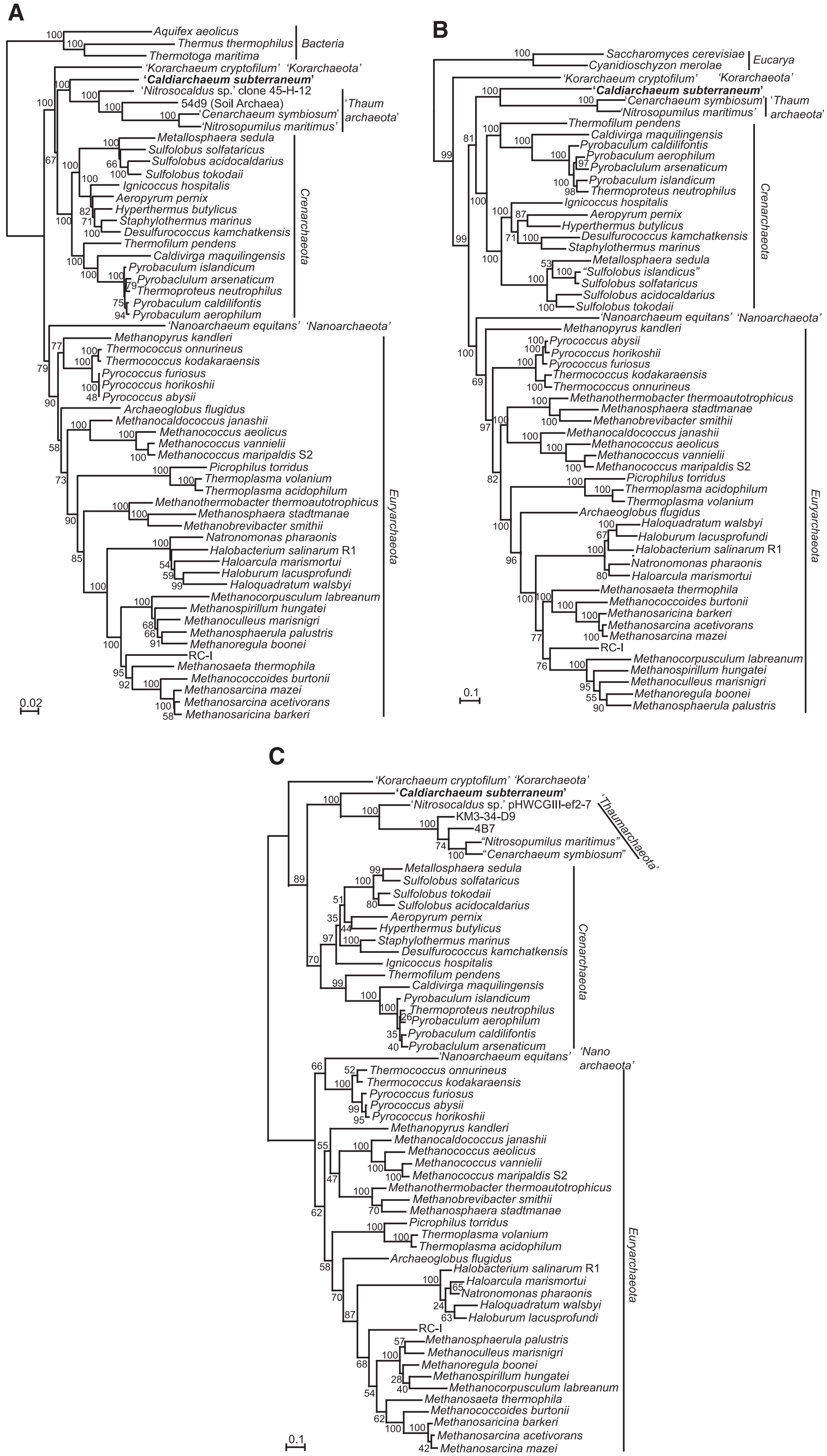


Figure 4. Phylogenetic analyses of *Archaea* including *C. subterraneum*. (**A**) Maximum likelihood phylogenetic tree of concatenated (SSU+LSU) rRNA genes using 3063 identical nucleotide positions. Bacterial sequences were used as out-group. Numbers indicate bootstrap values from 100 replications. (**B**) Maximum likelihood phylogenetic tree of concatenated universally conserved 45 ribosomal proteins and nine RNA polymerase subunits using aligned identical 5993 amino acid residues. Eukaryotic sequences were used as out-group. Numbers indicate bootstrap values (%) from 200 replications. (**C**) Maximum likelihood phylogenetic tree made from archaeal translation EF2 proteins based on 590 identical residues. Numbers indicate bootstrap values (%) from 200 replications.

genome (68). Non-homologous enzymes, such as members of other dehydratase groups, which are present on the composite genome, may be used as an alternative to support the function of the dicarboxylate/4-hydroxybutyrate cycle.

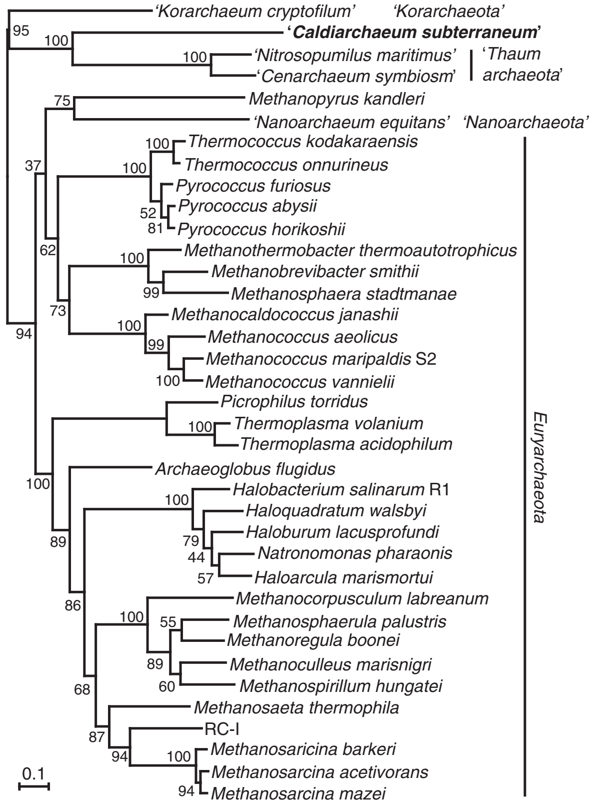


Figure 5. Maximum likelihood phylogenetic tree of concatenated (SSU+LSU) DNAP. Number of identical amino acid residues used were 829. Numbers indicate bootstrap values (%) from 200 replications.

The HWCGI has been detected from terrestrial and subsurface hot springs, and, recently, dominance of the group in anaerobic hot hydrothermal sediments was reported (7, 19–22). In such hot anaerobic environments, the most probable metabolism is anaerobic hydrogen oxidation dependent chemolithoautotrophy coupled with sulfur or sulfate reduction (22). Judging from the genome sequence, this does not seem to be the case in *C. subterraneum*. Consequently, the HWCGI is expected to be driven by a versatile energy metabolism as in the case of hyperthermophilic crenarchaeotes (87), and the composite genome of *C. subterraneum* probably does not represent all of the diverse energy metabolisms of the HWCGI.

In the unique archaeal genome, we found genomic signatures of potential hyperthermphilic life such as the presence of reverse gyrase and the relatively high G+C content of the SSU rRNA gene (21,51). On the other hand, we also observed the presence of DnaJ, DnaK and GrpE genes, reported only in the mesophilic and thermophilic, but not hyperthermophilic, archaea. The microbial mat formation studied here derives from a geothermal water stream with a temperature of 70°C, and other HWCGI SSU rRNA gene sequences have been detected from hot water (70°C, 72°C and 92°C) (7,20), hot spring sediments (74°C) (19,88) and hydrothermal sediments (from 35°C to 60°C) (22). Genes for reverse gyrase have recently been found from genomes of thermophilic bacteria (52,53), and it has been clarified that the gene is not necessarily a prerequisite for hyperthermophilic life (89). Taking all of these factors into account, the HWCGI including *C. subterraneum* can be considered to be thermophilic, but their optimum growth temperatures are most likely lower than those of hyperthermophilic crenarchaeotes. Considering the potential growth temperatures of *C. subterraneum*, the *Nitrosocaldales*, the most deeply branching thaumarchaeal group (74°C) (7) and mesophilic thaumarchaeotes, and the branch lengths of *C. subterraneum* and thaumarchaeal sequences in the

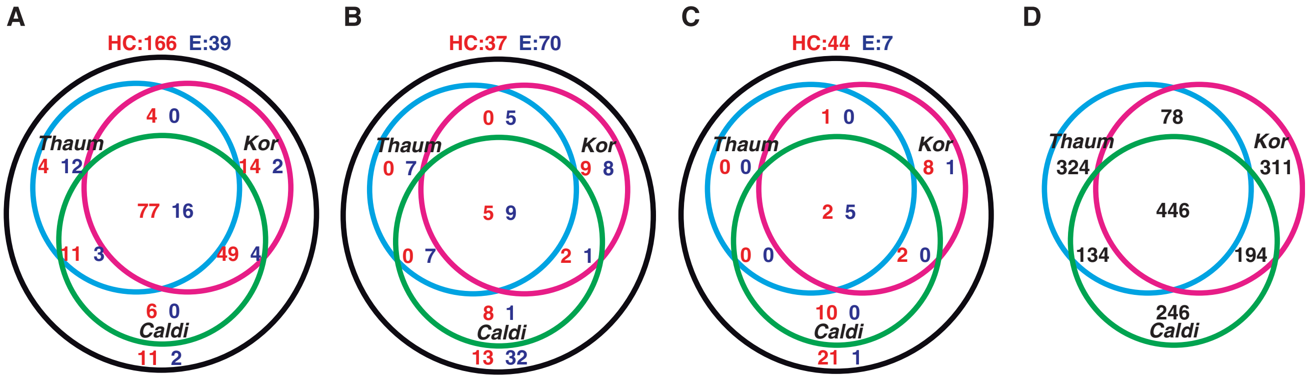


Figure 6. Venn diagrams presenting number of arCOGs among crenarchaeotic lineages; *Caldiarchaeum*, *Korarchaeum* and *Thaumarchaeota*. (**A**, **B** and **C**) Venn diagrams presenting number of arCOGs represents genome core genes of hyperthermophilic *Crenarchaeota* (HC: red) and *Euryarchaeota* (E: blue) in the genomes of the novel crenarchaeal lineages; *Caliarchaeum subterraneum* (*Caldi*), *Thaumarchaeota* (*Thaum*) and *K. cryptofilum* (*Kor*). A total of 11 hyperthermophilic-crenarchaeal and 27 euryarchaeal genomes in arCOG database were used in this analysis. (A) Genes that are represented in all sequenced genome used in arCOG from the represented division, but that are missing in at least some organisms of the other division. (B) Genes present in more than two-thirds of the genomes from one division and absent in the other division. (C) Genes that are present in at least one representative of each order of one division, but are absent from all genomes in the other division. (**D**) A Venn diagram presenting number of arCOGs shared among three crenarchaeotic lineages; *Caldiarchaeum*, *Korarchaeum* and *Thaumarchaeota*.

phylogenetic tree (Figure 4 and Supplementary Figure S1), the HWCGI and *Thaumarchaeota* most likely evolved from a (hyper-)thermophilic common ancestor in the course of adapting to lower temperature environments.

Evolutionary considerations

Differences in replicative functions, transcription and translation are one of the major criteria for phylum level characterization in the domain *Archaea* (90). The overall mechanisms of DNA replication/repair and cell division in *C. subterraneum* are more typical of the *Euryarchaeota* whereas the ribosomal proteins of this archaeon are shared more with crenarchaeotic lineages than with euryarchaeotes (Table 1). We also examined the number of arCOGs present on the genomes of *C. subterraneum*, *Thaumarchaeota* and *K. cryptofilum* that correspond to genome core genes of the *Euryarchaeota* and hyperthermophilic *Crenarchaeota*. These comparisons, along with the number of shared arCOGs among the novel crenarchaeal lineages, were used to clarify the affinity between *C. subterraneum* and other archeal phyla/divisions (Figure 6; Supplementary Table S2). The results indicate that (i) *C. subterraneum* is distinct from hyperthermophilic *Crenarchaeota*; (ii) *Thaumarchaeota* differs from *C. subterraneum* and *K. cryptofilum* with its significant euryarchaeotic features; and (iii) *C. subterraneum* shares more genes with *K. cryptofilum* than *Thaumarchaeota*. Moreover, judging from phylogenetic topology, indications of horizontal gene transfer (HGT) were not observed in most of the other euryarchaeal proteins in the crenarchaeal lineages that we examined, a typical case represented in the phylogenetic tree of D-type DNAPs (Figure 5). Taking all of these observations into consideration, we conclude that the complexity in the genomic core structures of the archaeal domain is mostly attributed to a combination of inheritance from an archaeal common ancestor and gene loss events, and that HGT events are not a major factor.

Considering the unique genomic features of *C. subterraneum* among the crenarchaeal lineages described above (*C. subterraneum*, the hyperthermophilic *Crenarchaeota*, *Thaumarchaeota* and *Korarchaeota*), the HWCGI occupies a position that can be considered an independent candidatus division among these lineages. On the other hand, phylogenetic trees suggest a close relationship between the *Thaumarchaeota* and *C. subterraneum* with high-bootstrap values (Figure 4), also raising the possibility that HWCGI represented by *C. subterraneum* is a deeply branching group in the *Thaumarchaeota*. Although conclusions will have to await further data accumulation, we would like to note several points that seem difficult to explain with the latter interpretation. At least two uncultivated crenarchaeal groups; Miscellaneous Crenarchaeotic Group (MCG) and Deep Sea Archaeal Group (DSAG) [also known as the Marine Benthic Group B (MBGB), whose phylogenetic position is still under debate] have been recognized (91) (Supplementary Figure S1). Although the HWCGI and *Thaumarchaeota* appear to be closely related in the phylogenetic trees shown in Figure 4, the inclusion of the MCG and DSAG sequences in the phylogenetic analysis based on SSU rRNA genes may influence the topology between the HWCGI and *Thaumarchaeota*. In addition, the genomes of *Thaumarchaeota* present more euryarchaeotic and less hyperthermophilic crenarchaeotic features than that of *C. subterraneum* as described above. It is difficult to explain, without considering the occurrence of HGT, that a deeply branching group conserves more crenarchaeotic features while a related group with longer branches within the same phylum/division shares more euryarchaeotic features. Therefore, there is the possibility that the HWCGI can be proposed as a novel division among the crenarchaeal lineages as '*Aigarchaeota*' (from the Greek αυγη '*aigi*', meaning dawn and aurora for the intermediate features of hyperthermophilic and mesophilic life during the evolution of the crenarchaeal lineage). However, the current analyses are based on the comparison of one HWCGI genome, one korarchaeal genome and two complete and two partial thaumarchaeal genomes. Thus, we cannot rule out the possibilities of the HWCGI as members of the *Crenarchaeota* or *Thaumarchaeota*. The classification of *Archaea* described in this study may have to be reconsidered in the light of future genomic analyses.

The genome of *C. subterraneum* also represents several eukaryotic features that have not observed in most of the previously known archaeal lineages. One such feature could be the presence of a type I DNA topoisomerase IB (TopoIB) family that has been found only in the *Thaumarchaeota* in the domain *Archaea* (8,92). The gene in *C. subterraneum* forms a clade with the *Thaumarchaeota* as a sister group of the eukaryotic cluster, and the phylogenetic topology supports the hypothesis presented by Brochier-Armanet *et al.* (92) that TopoIB was present in the last common ancestor of the *Archaea* and *Eucarya*, and lost in the *Euryarchaeota* and hyperthermophilic *Crenarchaeota*.

A striking eukaryotic feature of *C. subterraneum* is the presence of a potential protein degradation pathway that utilizes an Ub conjugation system. Although the possibility of the *C. subterraneum* Ubl gene cluster originating in eukaryotes was of concern, the structure of the gene cluster rules out the potential of HGT from eukaryotes. Most importantly, the gene cluster consists of five genes, which are partially overlapped (Figure 3), strongly indicating that this cluster is transcribed as an operon, a signature of prokaryotes. In addition, genes encoding prokaryote-type Ubl, E1l, E2l and JAMM proteins usually constitute fusion genes and/or form operon-like structures. The gene order of prokaryote-type Ubl, E2l and E1l genes in these operon-like gene clusters is highly conserved in the bacterial and archaeal genomes, and is also maintained in the eukaryote-type Ubl, E2l and E1l genes in *C. subterraneum*. No eukaryotic genome has ever been found to encode the protein modifier system in the form of a gene cluster, and it is highly unlikely that individual components derived by HGT from eukaryotes afterwards reorganize to form operon-like gene clusters. Furthermore, the gene for RPN11l is located adjacent to this gene cluster (Figure 3). The operon-like structure, the conserved prokaryotic gene organization, and the high similarity of the individual components to their eukaryotic counterparts strongly indicate that the eukaryote-type Ubl, E1l, E2l and adjacent RPN11l found in *C. subterraneum* had already evolved before the divergence between *Eucarya* and *Archaea*. The presence of the gene encoding the small Zn RING finger protein in this operon-like gene cluster raises the possibilities that a progenitor of RING-type E3, previously unidentified in prokaryotes, also occurred in the last common ancestor of *Eucarya* and *Archaea*. The only other possibility is that HGT occurred from an ancestral eukaryote still retaining prokaryotic gene organization. Such unexpected distributions of eukaryote-specific genes in particular archaeal groups have also been recently identified in cell division and vesicle-formation mechanisms, and these findings suggest a more complex gene composition in the genome of the last common ancestor of *Eucarya* and *Archaea* than those found in the genomes of individual modern *Archaea* (93).

As genes encoding the components of the *Haloferax* SAMPylation system, such as MoeB (prokaryote-type E1l) and MoaE, are present as single copies on various archaeal genomes (18,94), these genes might exhibit dual roles in both protein degradation and molybdenum/tungstate cofactor biosynthesis. *C. subterraneum* harbors both the molybdenum/tungstate cofactor biosynthesis systems in addition to the eukaryote-type Ub-like protein modifier system. The unique presence of the eukaryote-type Ub-like system in *C. subterraneum* and its absence in other *Archaea* are intriguing. As the *Haloferax* SAMPylation has been suggested to function in proteasome-dependent protein degradation, the eukaryote-type Ubl, E1l, E2l and RPN11l found in *C. subterraneum* might have been functionally replaced by the proteins for molybdenum cofactor/tungstate cofactor biosynthesis, allowing the gene loss of the eukaryotic system in most of the presently known archaeal lineages.

The composite genome of *C. subterraneum* provides further strong evidence that variations of the genome core in the domain *Archaea* are the result of a combination of vertically inherited ancient features and gene loss events rather than HGT. Furthermore, the genome provides novel insight into the evolutional relationship between *Archaea* and *Eucarya*, especially in the Ub–proteasome system. It is well recognized that many lineages of uncultivated *Archaea* exist on our planet that have yet to be examined. Future multidisciplinary studies combining cultivation, metagenomic or single-cell genomic analyses targeting these unexplored archaeal lineages will surely provide new perspective toward the understanding of the early evolution of life, especially in the *Archaea* and *Eucarya*.

ACCESSION NUMBERS

Sequences obtained or used in this study have been deposited in the DDBJ/EMBL/GenBank database under the accession numbers described below. A composite circular genome of *C. subterraneum*; BA000048. Complete or partial fosmid sequences from *of C. subterraneum*; AP011633, AP011650, AP011675, AP011689, AP011708, AP011723, AP011724, AP011727, AP011745, AP011751, AP011796 and AP011826-AP011902. Sequences and quality scores from pyrosequencing runs; DRP000160. Fosmid-end sequences of the metagenomic library; AG993735–AG999698. Fosmid sequences encoding representative intron-coding SSU rRNA genes from *Caldiarchaeum* type I (*C. subterraneum*); AP011786 and AP011878. *Caldiarchaeum* type II SSU rRNA gene sequence identified from the metagenomic library; AB566230. Partial *ef2* sequence from the *Nitrosocaldus* sp. (HWCGIII), pHWCGIII-ef2-7; AB543518. Sequences from *C. subterraneum* are also publically accessible from our ExtremoBase web site (http://www.jamstec.go.jp/gbrowser/cgi-bin/top.cgi).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank S. Matsui and M. Hirai (JAMSTEC) for technical supports in bioinformatics and gene cloning, respectively, and K. Furuya, C. Yoshino (University of Tokyo), A. Yamashita, A. Nakazawa and N. Ito (Kitasato University) for technical assistance in fosmid end-sequencing. We also thank Y. Watanabe (University of Tokyo) and Y. Ishino (Kyushu University) for discussion.

FUNDING

Ministry of Education, Culture, Sports, Science & Technology of Japan (Grant-in-Aid No. 18681030 to T.N., in part); Ministry of Education, Culture, Sports, Science & Technology of Japan (Grant-in-Aid No. 20310124 to H.T., in part); Ministry of Education, Culture, Sports, Science & Technology of Japan (Grant-in-Aid No. 22370066 to A.K., in part); Scientific Research on Priority Areas 'Comprehensive Genomics' (Grant-in-Aid to M.H., in part); Research fund at the Institute for Fermentation, Osaka, Japan (Grant-in-Aid to A.K., in part); Research funds at the Yamagata Prefectural Government and Tsuruoka City in Japan (Grant-in-Aid to A.K., in part). Funding for open access charge: Japan Agency for Marine-Earth Science & Technology (JAMSTEC).

Conflict of interest statement. None declared.

REFERENCES

1. Schleper,C. (2007) Diversity of uncultivated Archaea: perspectives from microbial ecology and metagenomics. In Garrett,R.A. and Klenk,H.-P. (eds), *Archaea; Evolution, Physiology, and Molecular Biology*. Blackwell Publishing, Oxford, pp. 39–50.

2. Hallam,S.J., Konstantinidis,K.T., Putnam,N., Schleper,C., Watanabe,Y., Sugahara,J., Preston,C., de la Torre,J., Richardson,P.M. and DeLong,E.F. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc. Natl Acad. Sci. USA*, **103**, 18296–18301.
3. Hallam,S.J., Mincer,T.J., Schleper,C., Preston,C.M., Roberts,K., Richardson,P.M. and DeLong,E.F. (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine *Crenarchaeota*. *PLoS Biol.*, **4**, e95.
4. Brochier-Armanet,C., Boussau,B., Gribaldo,S. and Forterre,P. (2008) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.*, **6**, 245–252.
5. Elkins,J.G., Podar,M., Graham,D.E., Makarova,K.S., Wolf,Y., Randau,L., Hedlund,B.P., Brochier-Armanet,C., Kunin,V., Anderson,I. *et al.* (2008) A korarchaeal genome reveals insights into the evolution of the *Archaea*. *Proc. Natl Acad. Sci. USA*, **105**, 8102–8107.
6. Könneke,M., Bernhard,A.E., de la Torre,J.R., Walker,C.B., Waterbury,J.B. and Stahl,D.A. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, **437**, 543–546.
7. de la Torre,J.R., Walker,C.B., Ingalls,A.E., Könneke,M. and Stahl,D.A. (2008) Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ. Microbiol.*, **10**, 810–818.
8. Spang,A., Hatzenpichler,R., Brochier-Armanet,C., Rattei,T., Tischler,P., Spieck,E., Streit,W., Stahl,D.A., Wagner,M. and Schleper,C. (2010) Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends Microbiol.*, **18**, 331–340.
9. Walker,C.B., de la Torre,J.R., Klotz,M.G., Urakawa,H., Pinel,N., Arp,D.J., Brochier-Armanet,C., Chain,P.S., Chan,P.P., Gollabgir,A. *et al.* (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc. Natl Acad. Sci. USA*, **107**, 8818–8823.
10. Huber,H., Hohn,M.J., Rachel,R., Fuchs,T., Wimmer,V.C. and Stetter,K.O. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, **417**, 63–67.
11. Brochier,C., Gribaldo,S., Zivanovic,Y., Confalonieri,F. and Forterre,P. (2005) Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol.*, **6**, R42.
12. Hochstrasser,M. (2009) Origin and function of ubiquitin-like proteins. *Nature*, **458**, 422–429.
13. Iyer,L.M., Burroughs,A.M. and Aravind,L. (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biol.*, **7**, R60.
14. Burroughs,A.M., Jaffee,M., Iyer,L.M. and Aravind,L. (2008) Anatomy of the E2 ligase fold: implications for enzymology and evolution of ubiquitin/Ub-like protein conjugation. *J. Struct. Biol.*, **162**, 205–218.
15. Burroughs,A.M., Iyer,L.M. and Aravind,L. (2009) Natural history of the E1-like superfamily: implication for adenylation, sulfur transfer, and ubiquitin conjugation. *Proteins*, **75**, 895–910.
16. Pearce,M.J., Mintseris,J., Ferreyra,J., Gygi,S.P. and Darwin,K.H. (2008) Ubiquitin-like protein involved in the proteasome pathway of *Mycobacterium tuberculosis*. *Science*, **322**, 1104–1107.
17. Darwin,K.H. and Hofmann,K. (2010) SAMPyling proteins in archaea. *Trends Biochem. Sci.*, **35**, 348–351.
18. Humbard,M.A., Miranda,H.V., Lim,J.-M., Krause,D.J., Pritz,J.R., Zhou,G., Chen,S., Wells,L. and Maupin-Furlow,J.A. (2010) Ubiquitin-like small archaeal modifier proteins (SAMPs) in Haloferax volcanii. *Nature*, **463**, 54–62.
19. Barns,S.M., Delwiche,C.F., Palmer,J.D. and Pace,N.R. (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl Acad. Sci. USA*, **93**, 9188–9193.
20. Marteinsson,V.T., Hauksdóttir,S., Hobel,C.F.V., Kristmannsdóttir,H., Hreggvidsson,G.O. and Kristjánsson,J.K. (2001) Phylogenetic diversity analysis of subterranean hot springs in Iceland. *Appl. Environ. Microbiol.*, **67**, 4242–4248.
21. Nunoura,T., Hirayama,H., Takami,H., Oida,H., Nishi,S., Shimamura,S., Suzuki,Y., Inagaki,F., Takai,K. and Nealson,K.H. (2005) Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments. *Environ. Microbiol.*, **7**, 1967–1984.
22. Nunoura,T., Oida,H., Nakaseama,M., Kosaka,A., Ohkubo,S., Kikuchi,T., Kazama,H., Tanabe,S.H., Nakamura,K., Kinoshita,M. *et al.* (2010) Archaeal diversity and distribution along thermal and geochemical gradients in hydrothermal sediments at the Yonaguni Knoll IV hydrothermal field in the Southern Okinawa Trough. *Appl. Environ. Microbiol.*, **76**, 1198–1211.
23. Lane,D.J. (1985) 16S-23S rRNA sequencing. In Stackebrandt,E. and Goodfellow,M. (eds), *Nucleic Acid Techniques in Bacterial Systematics*. John Wiley and Sons, New York, pp. 115–175.
24. DeLong,E.F. (1992) *Archaea* in coastal marine environments. *Proc. Natl Acad. Sci. USA*, **89**, 5685–5689.
25. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A. and Merrick,J.M. (1995) Wholegenome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
26. Takami,H., Nakasone,K., Takaki,Y., Maeno,G., Sasaki,R., Masui,N., Fuji,F., Hirama,C., Nakamura,Y., Ogasawara,N. *et al.* (2000) Complete genome of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilils*. *Nucleic Acids Res.*, **28**, 4317–4331.
27. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
29. Makarova,K.S., Sorokin,A.V., Novichkov,P.S., Wolf,Y.I. and Koonin,E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct*, **2**, 33.
30. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
31. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
32. Sugahara,J., Yachie,N., Arakawa,K. and Tomita,M. (2007) In silico screening of archaeal tRNA-encoding genes having multiple introns with bulge-helix-bulge splicing motifs. *RNA*, **13**, 671–681.
33. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
34. Ludwig,W., Strunk,O., Westram,R., Richter,L., Meier,H., Yadhukumar,H., Buchner,A., Lai,T., Steppi,S., Jobb,G. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.
35. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
36. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
37. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
38. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
39. Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

40. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
41. Hochstrasser,M. (2000) Evolution and function of ubiquitin-like protein-conjugation systems. *Nat. Cell Biol.*, **2**, E153–157.
42. Lake,M.W., Wuebbens,M.M., Rajagopalan,K.V. and Schindelin,H. (2001) Mechanism of ubiquitin activation revealed by the structure of a bacterial MoeB-MoaD complex. *Nature*, **414**, 325–329.
43. Wang,J., Hu,W., Cai,S., Lee,B., Song,J. and Chen,Y. (2007) The intrinsic affinity between E2 and the Cys domain of E1 in ubiquitin-like modifications. *Mol. Cell*, **27**, 228–237.
44. Lee,I. and Schindelin,H. (2008) Structural insights into E1-catalyzed ubiquitin activation and transfer to conjugating enzymes. *Cell*, **134**, 268–278.
45. Filée,J., Siguier,P. and Chandler,M. (2007) Insertion sequence diversity in archaea. *Microbiol. Mol. Biol. Rev.*, **71**, 121–157.
46. Sorek,R., Kunin,V. and Hugenholtz,P. (2008) CRISPR- a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.
47. Cann,I.K. and Ishino,Y. (1999) Archaeal DNA replication: identifying the pieces to solve a puzzle. *Genetics*, **152**, 1249–1267.
48. Rogozin,I.B., Makarova,K.S., Pavlov,Y.I. and Koonin,E.V. (2008) A highly conserved family of inactivated archaeal B family DNA polymerases. *Biol. Direct*, **3**, 32.
49. Komori,K., Fujikane,R., Shinagawa,H. and Ishino,Y. (2002) Novel endonuclease in Archaea cleaving DNA with various branched structure. *Genes Genet. Syst.*, **77**, 227–241.
50. Roberts,J.A., Bell,S.D. and White,M.F. (2003) An archaeal XPF repair endonuclease dependent on a heterotrimeric PCNA. *Mol. Microbiol.*, **48**, 361–371.
51. Forterre,P. (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet.*, **18**, 236–237.
52. Brochier-Armanet,C. and Forterre,P. (2007) Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea*, **2**, 83–93.
53. Campbell,B.J., Smith,J.L., Hanson,T.E., Klotz,M.G., Stein,L.Y., Lee,C.K., Wu,D., Robinson,J.M., Khouri,H.M. and Eisen,J.A. (2009) Adaptations to submarine hydrothermal environments exemplified by the genome of *Nautilia profundicola. PLoS Genet.*, **5**, e1000362.
54. Lindås,A.C., Karlsson,E.A., Lindgren,M.T., Ettema,T.J. and Bernander,R. (2008) A unique cell division machinery in the Archaea. *Proc. Natl Acad. Sci. USA*, **105**, 18942–18946.
55. White,M.F. (2006) DNA repair. In Garrett,R.A. and Klenk,H.-P. (eds), *Archaea; Evolution, Physiology, and Molecular Biology*. Blackwell Publishing, Oxford, pp. 171–184.
56. Nohmi,T., Yamada,M. and Gruz,P. (2008) DNA repair and DNA damage tolerance in archaeal bacteria: extreme environments and genome integrity. In Blum,P. (ed.), New models for prokaryotic biology, Casier Academic Press, Norfolk, pp. 147–169.
57. Sugahara,J., Kikuta,K., Fujishima,K., Yachie,N., Tomita,M. and Kanai,A. (2008) Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order Thermoproteales. *Mol. Biol. Evol.*, **25**, 2709–2716.
58. Sugahara,J., Fujishima,K., Morita,K., Tomita,M. and Kanai,A. (2009) Disrupted tRNA gene diversity and possible evolutionary scenarios. *J. Mol. Evol.*, **69**, 497–504.
59. Stock,T. and Rother,M. (2009) Selenoproteins in Archaea and Gram-positive bacteria. *Biochim. Biophys. Acta*, **1790**, 1520–1532.
60. Koonin,E.V., Makarova,K.S. and Elkins,J.G. (2007) Orthologs of the small RPB8 subunit of the eukaryotic RNA polymerases are conserved in hyperthermophilic Crenarchaeota and "Korarchaeota". *Biol. Direct.*, **2**, 38.
61. Blombach,F., Makarova,K.S., Marrero,J., Siebers,B., Koonin,E.V. and van der Oost,J. (2009) Identification of an ortholog of the eukaryotic RNA polymerase III subunit RPC34 in Crenarchaeota and Thaumarchaeota suggests specialization of RNA polymerases for coding and non-coding RNAs in Archaea. *Biol. Direct*, **4**, 39.
62. Lecompte,O., Ripp,R., Thierry,J.-C., Moras,D. and Poch,O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.*, **30**, 5382–5390.
63. Kelly,M., Lappalainen,P., Talbo,G., Haltia,T., Van der Oost,J. and Saraste,M. (1993) Two cysteines, two histidines, and one methionine are ligands of a binuclear purple copper center. *J. Biol. Chem.*, **268**, 16781–16787.
64. Hügler,M., Huber,H., Molyneaux,S.J., Vetriani,C. and Sievert,S.M. (2007) Autotrophic CO_2 fixation via the reductive tricarboxylic acid cycle in different lineages within the phylum *Aquificae*: evidence for two ways of citrate cleavage. *Environ. Microbiol.*, **9**, 81–92.
65. Berg,I.A., Kockelkorn,D., Buckel,W. and Fuchs,G. (2007) A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. *Science*, **318**, 1782–1786.
66. Huber,H., Gallenberger,M., Jahn,U., Eylert,E., Berg,I.A., Kockelkorn,D., Eisenreich,W. and Fuchs,G. (2008) A dicarboxylate/4-hydroxybutyrate autotrophic carbon assimilation cycle in the hyperthermophilic archaeum *Ignicoccus hospitalis*. *Proc. Natl Acad. Sci. USA*, **105**, 7851–7856.
67. Kockelkorn,D. and Fuchs,G. (2009) Malonic semialdehyde reductase, succinic semialdehyde reductase, and succinyl-coenzyme A reductase from *Metallosphaera sedula*: enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in *Sulfolobales*. *J. Bacteriol.*, **191**, 6352–6562.
68. Ramos-Vera,W.H., Berg,I.A. and Fuchs,G. (2009) Autotrophic carbon dioxide assimilation in *Thermoproteales* revisited. *J. Bacteriol.*, **191**, 4286–4297.
69. Teufel,R., Kung,J.W., Kockelkorn,D., Alber,B.E. and Fuchs,G. (2009) 3-hydroxypropionyl-coenzyme A dehydratase and acryloyl-coenzyme A reductase, enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in the *Sulfolobales*. *J. Bacteriol.*, **191**, 4572–4581.
70. Grochowski,L.L., Xu,H. and White,R.H. (2005) Ribose-5-phosphate biosynthesis in *Methanocaldococcus jannaschii* occurs in the absence of a pentose-phosphate pathway. *J. Bacteriol.*, **187**, 7382–7389.
71. Orita,I., Sato,T., Yurimoto,H., Kato,N., Atomi,H., Imanaka,T. and Sakai,Y. (2006) The ribulose monophosphate pathway substitutes for the missing pentose phosphate pathway in the archaeon *Thermococcus kodakaraensis*. *J. Bacteriol.*, **188**, 4698–4704.
72. Rashid,N., Imanaka,H., Fukui,T., Atomi,H. and Imanaka,T. (2004) Presence of a novel phosphopentomutase and a 2-deoxyribose 5-phosphate aldolase reveals a metabolic link between pentoses and central carbon metabolism in the hyperthermophilic archaeon *Thermococcus kodakaraensis*. *J. Bacteriol.*, **186**, 4185–4191.
73. Tumbula,D.L., Teng,Q., Bartlett,M.G. and Whitman,W.B. (1997) Ribose biosynthesis and evidence for an alternative first step in the common aromatic amino acid pathway in *Methanococcus maripaludis*. *J. Bacteriol.*, **179**, 6010–6013.
74. White,R.H. (2004) L-Aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannaschii*. *Biochemistry*, **43**, 7618–7627.
75. Meereis,F. and Kaufmann,M. (2008) Extension of the COG and arCOG databases by amino acid and nucleotide sequences. *BMC Bioinformatics*, **9**, 479.
76. Joazeiro,C.A. and Weissman,A.M. (2000) RING finger proteins: mediators of ubiquitin ligase activity. *Cell*, **102**, 549–552.
77. Verma,R., Aravind,L., Oania,R., McDonald,W.H., YatesIII,J.R., Koonin,E.V. and Deshaies,R.J. (2002) Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome. *Science*, **298**, 611–615.
78. Murata,S., Yashiroda,H. and Tanaka,K. (2009) Molecular mechanisms of proteasome assembly. *Nat. Rev. Mol. Cell. Biol.*, **10**, 104–115.
79. Amerik,A.Y. and Hochstrasser,M. (2004) Mechanism and function of deubiquitinating enzymes. *Biochim. Biophys. Acta*, **1695**, 189–207.
80. Walden,H., Podgorski,M.S. and Schulman,B.A. (2003) Insights into the ubiquitin transfer cascade from the structure of the activating enzyme for NEDD8. *Nature*, **422**, 330–334.

81. Lois,L.M. and Lima,C.D. (2005) Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1. *EMBO J.*, **24**, 439–451.
82. Cope,G.A., Suh,G.S., Aravind,L., Schwarz,S.E., Zipursky,S.L., Koonin,E.V. and Deshaies,R.J. (2002) Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science*, **298**, 608–611.
83. Maytal-Kivity,V., Reis,N., Hofmann,K. and Glickman,M.H. (2002) MPN?, a putative catalytic motif found in a subset of MPN domain proteins from eukaryotes and prokaryotes, is critical for Rpn11 function. *BMC Biochem.*, **3**, 28–39.
84. Ambroggio,X.I., Rees,D.C. and Deshaies,R.J. (2004) JAMM: a metalloprotease-like zinc site in the proteasome and signalosome. *PLoS Biol.*, **2**, E2.
85. Jurica,M.S. and Stoddard,B.L. (1999) Homing endonucleases: structure, function and evolution. *Cell. Mol. Life Sci.*, **55**, 1304–1326.
86. Hirayama,H., Takai,K., Inagaki,F., Yamato,Y., Suzuki,M., Nealson,K.H. and Horikoshi,K. (2005) Bacterial community shift along a subsurface geothermal water stream in a Japanese gold mine. *Extremophiles*, **9**, 169–184.
87. Huber,R., Huber,H. and Stetter,K.O. (2000) Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. *FEMS Microbiol. Rev.*, **24**, 615–623.
88. Barns,S.M., Fundyga,R.E., Jeffries,M.W. and Pace,N.R. (1994) Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc. Natl Acad. Sci. USA*, **91**, 1609–1613.
89. Atomi,H., Matsumi,R. and Imanaka,T. (2004) Reverse gyrase is not a prerequisite for hyperthermophilic life. *J. Bacteriol.*, **186**, 4829–4833.
90. Bernander,R. (2000) Chromosome replication, nucleoid segregation and cell division in archaea. *Trends Microbiol.*, **8**, 278–283.
91. Teske,A. and Sørensen,K.B. (2008) Uncultured archaea in deep marine subsurface sediments: have we caught them all? *ISME J.*, **2**, 3–18.
92. Brochier-Armanet,C., Gribaldo,S. and Forterre,P. (2008) A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol. Direct.*, **3**, 5.
93. Makarova,K.S., Yutin,N., Bell,S.D. and Koonin,E.V. (2010) Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat. Rev. Microbiol.*, **8**, 731–741.
94. Hartman,A.L., Norais,C., Badger,J.H., Delmas,S., Haldenby,S., Madupu,R., Robinson,J., Khouri,H., Ren,Q., Lowe,T.M. *et al.* (2010) The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. *PLoS One*, **5**, e9605.