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

Takuro Nunoura^{1,*}, Yoshihiro Takaki², Jungo Kakuta¹, Shinro Nishi², Junichi Sugahara^{3,4}, Hiromi Kazama¹, Gab-Joo Chee², Masahira Hattori⁵, Akio Kanai^{3,4}, Haruyuki Atomi⁶, Ken Takai¹ and Hideto Takami²

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

[©] The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/ by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 (E11) and E2-like (E21) proteins, respectively. Although only distantly related to their eukaryotic counterparts, Ubl, Ell and E21 proteins are present in prokaryotes (13–15). For simplicity, based on primary structure, we will refer to these proteins as the 'prokaryote-type' Ubl, Ell and E2l proteins. In prokarvotes, some of the prokarvote-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; prokaryotic Ub-like protein (Pup)-proteasome the 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 hydro-thermal 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 which include the the lineages 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 JFF030 F06 (AP011724), (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 JFF009 A08 (AP011867), JFF009 F01 (AP011864). JFF009 F10 (AP011875), (AP011844), JFF011 A11 (AP011858, AP011859), JFF012_C01 (AP011870), JFF015 C06 JFF013 A09 (AP011845), (AP011842), JFF015 E11 JFF015 C07 (AP011830), (AP011831), JFF017_C01 JFF021 E09 (AP011851), (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) ar | nd JFF055 E04 | 4 (AP011835) |
| analyzed by | | n this study. | Numbers in |
| parentheses f | | fosmid clone | are accession |
| | | enBank databa | |
| | | initially defined | |

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 genes of HWCGI and HWCGIII (Ca. rRNA 'Nitrosocaldus' sp.; SSU rRNA gene similarity between ammonia oxidizing thaumarchaeon Ca. 'Nitrosocaldus vellowstonii' (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 genome circular 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

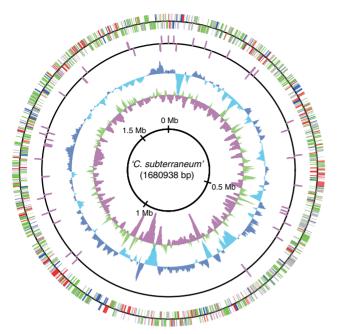


Figure 1. Circular representation of the *C. subterraneum* composite genome. From the inside, the first and second circles show the GC skew (values >0 or <0 are indicated in green and pink, respectively) and the G+C percent content (values greater or smaller than the average percentage in the overall 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 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 thermophiles, is observed (51–53). Genes for in 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

| | 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 | _ | + | + , | + | _ |

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

+, present; -, absent.

^aCharacterization of DNA polymerase is based on Ref. (47).

^bOnly C-terminal domain is found in C. symbiosum and N. maritimus.

^cOnly found in *Thermofilum pendens* and *Caldivirga maquilingensis*.

^dFusion form is observed in *Thermococcales* and *Thermoplasmatales*.

^eOnly 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 amido-transferase (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 citrvl-CoA synthetase and citrvl-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 succinvl-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 а 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 SAMPsproteasome 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 E11 (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 molvbdenum cofactor/tungstate cofactor biosynthesis pathway including a single prokaryote-type E11 (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, E11 and E21 (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 (RPN111) (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, E11 and E21 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 Ell 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). significant eukaryote-type feature in The the C. subterraneum Ell is the presence of two insertion helices (Asp₁₉₇-Ser₂₀₈ and Ile₂₂₄-Leu₂₃₉) between the Ub-E1 interaction domain and second Mg²⁺-chelating domain, which are found only in eukaryote-type E1s such as UBA1, UBA2, UBA3 and Aos1 (15) The JAMM (JAB1/MPN/Mov34 2B). (Figure 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 RPN111 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

| Α | | | ** |
|--|--|--|---|
| S.cere smt3 19KPETHINLKVSD S- | SEIFFKIKKTTPLRRLMEAFAKRQ <mark>C</mark> KEMDSLRFL | Y-D <mark>G</mark> IRIQADQTPEDLDMEDN- | -DIIEAHREQI <mark>GG</mark> ATY 101 |
| Human sumo1 17KEGEYIKLKVIGQDS | SEIHFKVKMTTHLKKLKESYCQRQ <mark>C</mark> VPMNSLRFL | F-E <mark>C</mark> QRIADNHTPKELGMEEE· | -DVIEVYQEQT <mark>GC</mark> HSTV 101 |
| Human sumo2 16DHINLKVAGQD <mark>G</mark> | SVVQFKIKRHTPLSKLMKAYCERKAYCER- | | -DTIDVFQQQT <mark>GG</mark> VY 71 |
| C.mero smt3 18SGGDQINLRVRDADG | vevqfrikkhtplrklmdayctrk <mark>e</mark> vdlhsyrfl | F-D <mark>G</mark> NRINEDDTPEKLGMED <u>M</u> - | |
| T.ther ubl1 8 ANANSEYLNLKVKSQEC | EEIFFKIKKTTQFKKLMDAYCQRAQVNAHNVRFL | F-D <mark>G</mark> DRILESHTPADLKMES <mark>G</mark> - | -DEIDVVVEQV <mark>GG</mark> SF 90 |
| G.lamb sumo 19 KPEQAQKIMIKVSDEHE | | | DIIEVMRNQI <mark>GC</mark> H 102 |
| | KEIEIDIEPTDKVERIKERVEEKE <mark>C</mark> IPPQQQRLI | | -SVLHLVLALR <mark>GG</mark> GGLRQ 81 |
| C.mero ubl 23 RSEPSETMLVKVKTLTG | | | -SVLHLVLALR <mark>GC</mark> HVC 108 |
| | KTITLEVESSDTIDNVKSKIQDKE <mark>G</mark> IPPDQQRLI | | -STLHLVLRLR <mark>GG</mark> IIEPSLKALA 86 |
| | -TVALTASPADSLTSIRQRLLAVYSGHV-VDSQRFV | | -SVLDLVPRLF <mark>GG</mark> VMEPTLINLA 86 |
| G.lamb ub1 1MGGFYMQIFVKTLTG | | F-SCKQLEDNRTLQDYSIQKD· | |
| · · · · · · · · · · · · · · · · · · · | | Y-S <mark>C</mark> KQMSDDLRLLDYKVTA <mark>G</mark> - | |
| | KVIQITSLTDDNTIAELKGKLEESE <mark>G</mark> IPGNMIRLV | | -ATFHMVVALRAGC 78 |
| | | Y-LAQQLQNTTTVEEANLKAG- | |
| | | y-k <mark>g</mark> ralkdtetleslgvad <mark>g</mark> - | |
| | PRLPYKVLS-VPESTPFTAVLKFAAEEFKVPAATSAII | ~ ~ | -SELRIIPRDRVGSC 85 |
| | PNLPFRTIS-VPEEAPFSACIKYVAEQFKVNHATSAII | | |
| ~ ~ ~ | RIVSTNVLATDSLAVVLSRVT <mark>G</mark> LDADAVYGT | | |
| | VTRSLEVDPTMSVKELRHIISEFS <mark>G</mark> ISIDSQCIS | | |
| | SMKSLILYVEENIIQYRKDHFI-ET <mark>G</mark> SKIKPGIIV | | -DLVTFIMTLH <mark>GG</mark> 98 |
| ~ ~ ~ ~ | NLNGLVQLLKTNYVKERPDLLVDQT <mark>G</mark> QTLRPGILV | | DTVEFISTLHGG 102 |
| ~ ~ ~ | DMVELDGSTVGEVLNKLVSRYTA-LQKHLFNENGAIRSFVNV | | DVVYIIPSIAGCLSIAAPAAVA 118 |
| | ASEEFELPQGSTVIDFLEKLRQVYGG-VLGDLFEGD <mark>G</mark> LRTGFAL | | DVVVVLPPIAGGYLKLGSLTPR 107 |
| | ETIRLEESPRTVRELLDLLAAKLGKSFEELVYDPRQK-TLKRAIVL | | |
| | -TVKINGRDMVCVGKTISQVLVSV <mark>G</mark> VDP-ARQGIAV | | |
| | RTVRVDVDGDATVGDALDALVGAHPALESRVF <mark>C</mark> DDGELYDHINV. | | -DELALFPPVS <mark>GG</mark> 87 |
| | ETSEVAVDDDGTYADLVRAVDLSPHEVTVL | | -DRVKVLRLIK <mark>GG</mark> 66 |
| | KDVKWKKDTGTIQDLLASYQLENKIVIV | | |
| | errriapgtaldtlvktltaappsgvaa | | |
| | QQQSYDGPMNVQQLVEKLSLQNKRFAI | | -DQLEIIVAVG <mark>GG</mark> 66 |
| | ~ ~ ~ | LED <mark>GKLLAAVNQTLVSFDHSLTD</mark> G- | |
| | DEEEIELPEGARVRDLIEEIKKRHEKFKEEVF <mark>C</mark> EGYDEDADVNI | | DVVGVFPPVSGG 99 |
| | PEIELEILPGDTVGTALQALQARYGPEFKEATTGTTAGG-IPKVRF | | DVMVFVPPVA <mark>GG</mark> 94 |
| | WLNLEVPDGTTVGAAIERSGILAQFPHIDLTVQKV | | DRVEIYRPITCDPKAVRKKADA 89 |
| P.syri RnfH 1 MADASIQIEVVYASVQR-Q | VLKTVDVPTGSSVRQALALSGIDKEFPELDLSQCAV | GIF C KVVTDPAARVLEA <mark>C</mark> - | -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 sumol (54792065), Human NEDD8 (5453760) and Human Ufm1 (7705300), from Cyanidioschyzon 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. subtertaneum; 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 (IVJK A), from Methanosarcina acetivorans; M.acet MoaB (AAM05120), from Aromatoleum aromaticum; A.arom NrfH (CA107579) 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 Schizosaccharomyces 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; E11 (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; SANA3 (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. pombe 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; *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.

..

....

. . . .

| B1 A | rginine finger | Nucleotide binding motif | |
|--------------------------------------|--|---|------------------|
| S.pomb E1L 17 E | GUYSROUYVUCHEAMKON | SQSNVLIICCKCLCVEIAKNVCLACVKSVTLYDPQPTRIEDLSSCYFLTEDDICVPRAKVTVSKLAELNQY | Y- 106 |
| | | x XTSSVLVSGLRCLGVEIAKNIILGGVKAVTLHDQGTAQWADLSSQFYLREEDICKNRAEVSQPRLAEINSY | |
| | | ĨĸſĸĸĸŦŦŸĠĹĊĔŶĊĬĔŸĸĸŇĿŶĹĂĠ₽SQŸŸŦŶĎŇĬĊĸŚŸĎQĠŸŇŦŸĬQĔĸŀ-VĸŀŇSTRĂĔĂŠĂĔQĹQQĹŇĔŸ | |
| | | RASRVLLVELKELGAEIAKNLILAGVKGLTMIDHEQVTPEDPGAOFLIRTGSVERNRAEASLERAQNINE | |
| S.cere Aos1 13 I | LALYDROIRLWGMTACANN | RSAKVLLINLEALCSEITKSIVLSCICHLTILDGHWVTEEDLGSOFFIGSEEVCOWKIDATKERIQDINE | |
| T.therm E1L 21 L | LQVYDRQ-RFIGVEVQKRI | .nakvfitpan <mark>e</mark> vntelaknlilC <mark>e</mark> tn-isia d neivnoddvetnfliaph <mark>e</mark> lekirgevvkaklodm <mark>ne</mark> m | M- 108 |
| CSUB_C1476 8 I | LSRYDRQLRLE <mark>G</mark> -WD <mark>Q</mark> NKI | LSGRVIVAGVCALCCEVAKNLALMCVCELLLIDNDYVELSNLSROMLYTDQDICRPKASTAEKKISLMNEI | LV 97 |
| Human UBA1 448 Q | | 3KQKYFLV <mark>CAC</mark> AICCELLKNFAMICIGCGECCEIIVTDMDTI <mark>B</mark> KS <mark>NLNRO</mark> FLFRPW <mark>DVT</mark> KL <mark>K</mark> SDTAAAAVROM <mark>NE</mark> H | |
| S.cere UBA1 414 N | | anskvflv <mark>escale</mark> cemlknwallelgsgsdevivvtdndsieksninreflfrpkdveknksevaaeavcamne | |
| T.cruz E1 419 G | GSRYDKQIAVL <mark>C</mark> AAF <mark>O</mark> SYI | SKQRAFIICACALCCELIKNAACMCFCGISITDMDSIEISNLSRQFLFRNSHICQHKSRVAGEAAMAINHI | |
| | | AGGRVLVVCACCIC CELLKNLVLTCFSHIDLIDLDTIDVSNLNRCFLFQKKHVCRSKAQVAKESVLQFYEK | |
| | 1PRETSLVTII <mark></mark> EDSYKKI | RSSRCLLVGAGCIGSELLKDIILMEFGEIHIVDLDTIDLSNLNROFLFRQKDIKQPKSTTAVKAVQHFNNS | |
| | | ESMHILLVGAGGICSEFLKSIITICCKNIDIIDIDTIDITNLNRQFLFKKKDVKKHKSIVARERALKHRKI | |
| | | DTCKVLVICAGGICCELLKNLALSGFRQIHVIDMDTIDVSNINRQFLFRPKDICRPKAEVAAEFINDRVEN | |
| | | ESAKVLVVCAGGICCEILKDLALSGVKDIHVIDLDTIDLTNLNRQFLFRMKDVGKFKSQVAADFIMRRVFG FSSKILIIGAGGICCEILKDLALSGFRDLSVIDMDTIDITNLNRQFLFNESNIDEPKANVAASMIMKRIS | |
| | | | |
| Human UBA5 50 S T.therm UBA5 42 D | SNPYSRLMALKRMGIVSDYEK. | RTFAVAIV <mark>EVGEVE</mark> SVTAEMLTRCEIEKLLLFDYDKV <mark>ELANMNR</mark> -LFFQPHQAELS <mark>K</mark> VQAAEHTLRNINE RDCSVLVVEVGEVESVLAEMLTRCELEKLIIYDYDKVELA <u>NMNR</u> -LFYTPQQVELS <mark>K</mark> VDAAKGTLQSINE | JV 142 |
| | | CDCSVLVVEVGEVESVLAEMLIRCELEKLIIIIDIKVELAAMMR-LFIIPDOVELSAVDAARGILOSIME KNTKVLLLEACTIECYVSRALIAWEVRKITFVDNGTVSYSNPVROALYNFEDCEKPKAELAAASLKRIFEI | |
| | | NTRULLIGAGTLECTUSRALLAWGURATTFULNGTUSTSNPWRALTNEDGEREWALAAASLIKTFU JSVKCLLLIGAGTLECNVARTLMGWGURHITFUDNAKISYSNPVROPLYEFEDCLGGEREWALAAADRLQKIFEG | |
| | | 99 VRCHINGASTIC CIVARTINGWS VRHTTFVLVRRLSTSKP NOPHTEF BCLIGG RAWIAADAL KITAG DAQKCLLI <mark>C</mark> CGTLCCNVARYLLGWGVRKFVLID YGKVSFSNP ROSLFTFAD CIDGCRSKCEAAAKELKRICT | 37 407 10 378 |
| | | rsopleblige freeventillewever internet in the first of an independent of the second second second second second | |
| C.perf 347 K | HLEPLTRNLAAN-ISTLSCK | MPKILFVEACALCSKIIFHLARNCYTDISVVDNDILVPHNLVRHALFADSISKNKAKEIINKLNNIYIM | |
| | | adkkvavigvesveceiahklsaaevrhltlvdpdvyeinnlyrhvleqhwvearktaalsvalqrqfe | |
| | | ALDDLFLIGIGAIGNGAVWALSRVPHLOCHLOVVDGEOVDOGNLORYVLALERDIGOSKVKLAWRYLKAORN- | |
| A.vari 4 - | -LTTYQQALPVLPRNH | TRINFVLV <mark>evGGTG</mark> GFLAEDLCRTILQLQHTRKEINFAIVDGDTVELKNISRQNYQQAEIGLFKAETLAARCSAKYGI | I - 96 |
| Nostoc 121 F | PPALOROALAF CEALNODI | SMLRVGVICCCCTCSAIAMLLPKWCIRNIALFDKDIVEDTNLNRLHGAROPDAD-AMSPKVEVVAKSLVELGLG | |
| P.yoel MoeB 85 V | /EKHGKYINIEEI-NTNSLNT | FKTKILIIGIGGLGSPICFYLSKFGFSEIGLVDGDKVEKSNLHRQIIHKKKNIGLNKTISAKLTLNDFDEN | N - 176 |
| Human MOCS3 60 I | ILRYSRQLVLPEL <mark>-</mark> GVHG <mark>Q</mark> LRI | FTACVLIVCCGGLCCPLAQYLAAAGVCRLGLVDYDVVEMSNLARQVLHGEALAGQAKAFSAAASLRRINSA QNTKVLVVCAGGLCCPALPYLAGAGVCQIGIVDNDVVETSNLHRQVLHDSSRVGMLKCESARQYITKLNEF | |
| S.cere UBA4 44 Y | /QRYGRQMIVEETG <mark>G</mark> VAG <mark>O</mark> VKI | (NTKVLVVCAGGLCCPALPYLAGAGVCQIGIVDNDVVETSNLHRQVLHDSSRVCMLKCESARQYITKLNEH | |
| | | ARARVLLI <mark>C</mark> AGGL <mark>C</mark> SPAAFYLAAACVCYLRIADDDVVDRSNLQ <mark>RC</mark> ILHTEDSVCTAKVDSAARRIAALNER | |
| | | (DSRVLIVGLGGLCCAASQYLASAGVGNLTLLDFDTVSLSNLCRQTLHSDATVGQPKVESARDALTRINE) | |
| | | QVARVCVVCVGGGICNPIVTRLAAMGVCKLRIVDRDVI <mark>E</mark> LSNLHRQTMYEESDVCRVKVEAAAEKLRRLNSI | |
| CSUB_C1135 18 I | LKRYGRHLIIPEV-GMAGOKKI | KAAKVLVVCAGGLCSPISLYLAAAGVCKIGLVDFDLVDESNLCRQVLYTTRDVKRPKLEVAKERLTALNE LSSRVVVVCAGGLCAPAIQYLAAVCVCELVVVDDDVVERSNLCRQVVHCDDDVCTPKAESAAAFVRGLNEL | |
| | | | |
| | | 2EKKVAVVCASALCSWEVYFLKKLCVCEIIVVDRDFVBASDIPR-TIYTEKDICRPKVDVLRDRFG RSKHVVLVCACALCTGNAEALVRACICKLTIIDRDYVEWSNLOROQLYSEAD-AKERLPKAIAAKRRLEQINSE | |
| | | KQATIGIAÇAÇÇIÇSSIATALVRAÇIÇRIIIADYDVEPS <u>NINRQ</u> QFFIDQICMNKVDALKDNIKRINEF | |
| | | LSRVLLVCLGGLCGHVLDMLVRLCVGHITAADGDVFEPSNLNRCLLSSMSRVCTSKAQAARDHARNTNEA | |
| B.thet ThiF 4 N | INWOORTELLL GEEKMKRI | RASHVLIVV <mark>CLGGVC</mark> AYAAEMLCRAGVCRMTIVDADTVQPA <mark>MNRQLPAMHSTICMPKAEVLAARYKDINE</mark> | |
| | | SNQYVVVV <mark>C</mark> AGGVCSWVVNSLVRSGCRKIRVVDFDQVSLSSLNRHSCAILNDVCTPKVECLRRHMREIA | |
| T.ther YqdLl 71 K | EQLVRNIQFFGEEGOKK | DDSYIII <mark>B</mark> GVGGVCSHVAASLARSCVAHLKIVDFDQVSLSSLNRHAFATHADVCRSKCECVKDYIKRIVE | |
| | | RNSRVAIACMCGVCGIDMVALARMCICKFTIADPDVFEIRNSNRQYGAMRSTNGQAKAEVMRNIVHDINE | |
| C.hutc 111 E | SVRTNRNQYKI TPEERDKI | SKOKIGVI <mark>G</mark> LS-V <mark>G</mark> OSISLTLALERSF <mark>G</mark> ELRIA D FDVI <mark>E</mark> LS <mark>N</mark> LN R -IRSGLSNLNLK <mark>K</mark> TVCVAREILEID <mark>E</mark> F | F - 199 |
| B.thur 103 I | INYFSLFTKFGED-KYKI <mark>Q</mark> EKI | jetpiall <mark>gvægle</mark> tqvlyhlaal <mark>e</mark> fhnikaldfdni <mark>e</mark> ls <mark>nfnre</mark> llysesdiens <mark>k</mark> vemakkrisqf <mark>ne</mark> n | N- 194 |

Figure 2. Continued.

eukaryotic RPN11, consisting of \sim 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. distinct from and is 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

| B2 | | Ubiquitin interaction domain # # | |
|--|---|--|---|
| B2 S.pomb E1L Human E1L T.ther E1L Human sumoE: S.cere Aos1 T.therm E1L CSUB_C1476 Human UBA1 S.cere UBA1 T.cruz E1 Human UBA2 S.cere UBA2 Human UBA3 T.thermE1B S.pomb UBA3 Human UBA5 T.therm UBA4 S.cere ATG7 Human ATG7 | 123 PDLVVATEMOTDEAIKINTLTRKL- 152 FNITTSSTPIFKEMELYDEISHFL- 117 ADVIVSAVDNWPTRRWINSMAVHV- 556 IDFVTNAIDNVDARYYMDRRCVFY- 536 IDFVTNAIDNVDARYYMDRRCVFY- 534 HAVVINAIDNVARYYMDRRCVFY- 135 FIITNAIDNIAARYINKISOFL- 136 PIITVAIDNIAARYINKISOFL- 139 YDVINAIDNIAARYINKISOFL- 132 FOVIIAGDNIAARYINKISOFL- 134 FKLIICGIDSVEARWINSTVAIA 135 VDLVLSCVDNFEARMINTACNELG | <pre># # + -HIAYIAADSRCIFCS FODFGEN 93 KISFKSLRESLKDEGIKLVVADTRCIFCS FODFGEN 93 KISFKSLRESLKDEGIKLVVADTRCIFCS FODFGEN 95 KISFKSLVASLABEDGIGTYTANLGIFCS FODFGEN 95 KISFKSLVASLABEDGIGTYTANLGIFCS FODFGEN 95 KISFKSLVASLABEDSIGFTGANGVFGYFGALGEH 38 -VVFCPVKEALEVDWSSEKAKAALKRT-TSDVFLIQVLL-KFRTDKGRDES2 -NIELYVASSNGJFAYVFIDLIEF 46 -NCYRPINEVLSTATL -KEKMTORQLKRV-TSILPITISL-LOYGUNKGKAISFEQ 26 -NIEYVATTGYYGNOTVIPGVTSLEHABALIPS 1 -IQABESLRRTPHDLVKDLSER-GISIN20 -RKELLESCTIGTKGNQVTIFFLTESYSSSSDEPE -KSIPICTLKNFENALEHTIC/AR-DEFEGJFKQPAENVNR 63 -KGLEESCTIGTKGNQVTIFFLTESYSSSR-DEPE -KSIPICTLKNFENALEHTIC/AR-DEFEGJFKQPAENVNR 63 -KGLESSCTIGTKGNQVTIFFLTESYSSSR-DEPE -KSIPICTLKNFENALEHTIC/AR-DEFEGJFKQPAENVNR 63 -KGLESSCTIGTKGNQVTIFFLTESYSSSR-DEPE -KSIPICTLKNFENALEHTIC/AR-DEFEGJFKQPAENVNR 64 -KKVLIEASTGYNQVTYFSNETKGYNGEKEPNKTYPICTIRSTESCIHCTVWAKFLENDJFASTGKNE 21 -QSYNSTLASTERIEHCUEYVR-MLOHFASTGRND2112 CSIVGLIDCGTEGTKGARATIIFGKTALEDTKKGFNETKGYNGEKEPNKTYPICTLANTERIEHCUEYVR-MLOHFAASCSP2 1 TWFESGVSENAWSHOLIFGETAGFAGAPLVVA 7 -KREGVGAAS</pre> | 46 32 58 66 63 65 33 10 65 33 10 65 33 10 6 37 7 8 3 37 38 3 53 31 0 11 6 11 6 7 7 18 8 31 10 10 10 10 10 10 10 10 10 10 10 10 10 |
| T.vagi APG7 R.meta C.perf S.ANA3 R.etli A.vari N.7120 P.yoel MoeB Human MOCS3 S.cere UBA4 X.axon MoeB E.coli MoeB C.symb ThiF CSUB_C1135 HVO_0558 P.furi MoeB G.kaus MoeB D.acet ThiF D.desu ThiF B.thet ThiF | 416 ODCTWLITD TRESRWLPTLLATAN- 455 PODAALIVDATASLOVLAAETOSAALD- 458 ISYTVLID CSASKSVFSFISEYSKL- 408 FDLTVIAIG NPTOERLFKQYLLDNN 290 -OHVAVALD TAADRIAVOGTLP 119 ITVIIGCUD NSAARSKIHSVLKINS 234 ODVIFCTD NITGRIMINRFAYYY- 135 YDVVADCSD NITGRIMINRFAYYY- 136 YDIILDCTD NISTRFLINDLCLLY- 173 YDVVADCSD NIPTRYLNDACVLA 158 YNYLDCTD SPLITKYLNSDVAVLL- 122 HDVVDGSD NIPPRYLINDACVLL- 123 YDVVDGSD NIPPRYLINDACVLL- 124 HDVVDGSD NIPPRYLINDACVLL- 125 YDVVDGSD NIPPRYLINDACVLL- 126 HDVVDGTD SVARYOLNACGFAA 121 YDVVDGSD NIPPTRYLINDACVLL- 123 SOVVDASD NIPPTRYLINDACVLL- 124 HDVDGTD NIPPTRYLINDACVLL- 125 YDIIDCTD NIPPTRYLINDACVLL- 126 HDVDDTD CATKAMITGONARTY 127 FOLWIDATE TISPKCFILYEAMTA- 127 HUNDATETISPKCFILYEAMTA- | -EKLCISVAL-CTDTFSVVRGCH -GLC YFENDVIATID 5 -TLDMCTVT | 90 54 63 01 31 83 76 80 90 80 91 72 86 90 91 72 84 0 38 70 |
| S.cere YgdL T.ther YgdL M.tube C.hutc B.thur | 180 FIVDCIDNIDIKVDLLEFAYNH- 192 FTYVUDCIDNIDAKVSLLAYCKLM- 133 ADVLVDGIDAF-EIDLRRLLYREAQQR- 224 IDLVIDECDGIDIKILCRYKKEL- 217 TELVICVAD-KPTHIANWYNEGVVKC- | -GIKVISSMGASAKSDPTKINVG- DLATTEED | 45 17 92 |

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

| - |
|-----|
| 00 |
| ה.ח |
| |

| DU | | | | л л |
|----------------------------|-----|---|--------------------------------------|--|
| S.pomb ElL | 310 | PRPRNDIDAAEFFEFCKKIASTLOFDVELDEKLIKEISYOARCDLVAMSAFLGGAVACDVLKATTSK- | | # # 376 |
| Human E1L | 347 | PRPRNEEDAAELVALAQAVNARALPAVQQNNLDEDLIRKLAYVAACDLAPINAFIGGLAAQDVMKACSCK- | -FMPIMQWLYFDALEC- | -LPEDKEVLTEDKCLQRQNRYDG 453 |
| T.ther E1L | | LLNQDHSKQLKEIVHKLLESNKADASNKFKVEEIPDELIQNVSLYARAHISPVASFWGGVVAQBIVK-FTCK- | -FTPLRQWLHHEVFEC- | -LPDSQVTREVVDSQNGHYVA 438 |
| Human sumoEl | 259 | SDTYEEDSELLLQIRNDVLDSLGISPDLLPEDFVRYCFSEMAPVCAVVGGILAQEIVKALSQR- | -DPPHNNFFFFDGMK | -GNGIVECLGPK 346 |
| S.cere Aosl | | MKRDAAVWCENLGVPATVVKDDYIQQFIKQKGIEFAPVAAIIGGAVAQDVINILGKR- | -LSPLNNFIVFDGIT | -LDMPLFEF 347 |
| T.therm ElL | | YNHSEENQKVLEQIVELAQEKIKNEE <mark>D</mark> REFYTNFAKFYGIEHCPVYSVIGSVASQ <mark>E</mark> FIKVIAKD- | -KMPALNWFVYDSQI | -GYGKIESQTDKI-DATYVDLPELTRK 366 |
| CSUB_C1476 | | LSDAETLFQHNIKT-VYDIKFAPQTVLDQMDKSLREQVIQLRSLLNPKWEALQSISATVSGLASFEVVRLLHKG- | -SLGRSLNGMMVFD | GLRGRLSRIKLERNVNCHVCGYS 315 |
| Human UBA1 | 843 | KMYPIDFEKDDDSNFHMDFIVAASNLRAENYDIPSADRHKSKLIACKIIEAIATTTAAVVGLVCLELYKVVCCHR 2 | -DSYKNGFLNLALPF | FGFSEPLAAPR-HQYYNQEWTL-WD 956 |
| S.cere UBA1 | | KLEPVDFEKDDDTNHHIEFITACSNCRAQNYFIETADRQKTKFIAGRIIEAIATTTSLVTGLVNLELYKLIDNKT 2 | | FGFSEPIASPK-GEYNNKKYDKIWD 929 |
| T.cruz El | | RMVPEFFEKDDPTNHHVEYITACSNMRAVAYNIPPADVHHTKRIACKIIPAMVTTTALVTGLVGIPVLKRLLMTQ 21 | LSIYRNAFVNIALPF | IAFSDPIIASG-ATYPLPDGTSVRW 939 |
| Human UBA2 | | DGAELIWDKDDPSAMDFVTSAANLRMHIFSMNMKSRFDIKSMACNIIEAIATTNAVIAGLIVLEGLKILSCK- | -IDQCRTIFLNKQPN | -PRKKLLVPCALDPPNPNCYVCASK 447 |
| S.cere UBA2 | | EQNHIEFDKDDADTLEFVATAANIRSHIFNIPMKSVFDIKQIAGNIIEAIATTNAIVAGASSLISLRVLNLLK 5 | | -LSQNRYLSNPKL-APPNKNCPVCSKV 441 |
| P.yoel UBA2 | | TEEYLIFDKDDDDCINFITCLSNLRMINFSIKQKSKFDIQSIACNIIBAISSTNAIVAAFQAAQLVHVIEHFE 19 | -RDSKAKHIWIKNVVNG 3 | -FSRGNIVNAENL-ETPNPNCYVCQQP 474 |
| Human UBA3 | | GVPLDGDDPEHIQWIFQKSLERASQYNIRGVTYRLTQGVVKRIIEAVASTNAVIAAVCATEVFKIATSA- | -YIPLNNYLVFNDVD | GLYTYTFEAERKENCPACSQL 341 |
| T.thermE1B | | TRKADKDSMEDMTWIYETAKKRAEQFNIKGVDYNKTIGVVKNIIPAIASTNAIIAASCANDAFKAFLQQ- | -SLNIKDYFQYMGNT | GVSTLTFPYERNEKCIVCSSL 341 |
| S.pomb UBA3 | | NSNFEPDNIRHIDWLVKRSIERANKFQI-PSSSINRFFVQGIVKRIIEAVASTNAIIAASCCNBALKILTES- | -NPFLDNYMMYVGED | GAYTYTFNLEKRSDCPVCGVL 356 |
| Human UBA5 | 254 | LPTIMGVVAGILVQNVLKFLLNFG | TVSFYLGYNAMQ | -DFFPTMSMKPNPQCDDRNCR 309 |
| T.therm UBA5 | 244 | LPTTMGITAGFLAQNALKFLLDFG | DLAFVLAYNAKA | -DFFTNYMIKPNSECKENECR 299 |
| S.cere ATG7 | 511 | RPGVAMMASSLAV <mark>P</mark> LMTSLLQTK | -YSGSETTVLGDIPHQI 3 | LHNFSILKLETPAYEHCPACSPK 575 |
| Human ATG7 | 615 | RPGLAVIAGALAVPLMVSVLQHP 12 | -RMNEPPTSLGLVPHQI 3 | LSRFDNVLPVSLAFDKCTACSSK 691 |
| T.vagi APG7 | 491 | RPGIAPMASSYGVPLWASIVQTK 4 | | LHSWQLLPMAGKPFKNCVACSEP 559 |
| R.meta | 555 | | -KEATLCAGISDAE | GLGMAWTRASLGPTTALEVAD 615 |
| C.perf | | KISDNIISYHAAIFSSYIKKHITNDI | | |
| S.ANA3 | 502 | SYGAASSAQTAVMAANLAIRYLEEKQ 12 VEMAFQSAMAGIMLACEIIKQAAENA | | -SSSLEYLPLVDEDCDVCTH- 575 |
| R.etli | 432 | VIMAFQSAMAGIMLACEIIKQAAENA QARNYQSLFVNKMTSAIAAQYLLELT 2 | -PDWTT-AKLNLLRNIP 1 | -DIVTERRKKDRLGRC-ICODE 493 MSTRSLYTSIDOLKKYYI 278 |
| A.vari | 219 | QARNYQSIFVNRMISAIAAQYLLEEF 2 | -GGLKKFASYFDLKA | MSTRSLYISIDQLKKIYI 278 -LTTDRKPAAILNSNOPVOGTV 398 |
| Nostoc | 338 | | -GEDGAAAHRVRKFH -KTLKPFLSYNSFSNN- | -KPFEVINMNYKNKNCI-CSIY 330 |
| P.yoel MoeB | 269 | | -RILKPFLSINSFSNN- | -GHFRSIRLRSRRLDQAACGER 303 |
| Human MOCS3 S.cere UBA4 | 243 | | -PSYSGSLLLFDALR -ENFSPFLMLYSGFPO- | -GHERSTRIKSRELDGAAGGER 303 -OSLRTFKMRGROEKOLCOGKN 292 |
| X.axon MoeB | 229 | CVLGVCPCVGLOATPAIKLLCIG | -DGLTGRLLSFDALA | -OSLRIFAMRGROEACLCOGAN 292 -MRFRDIRLPPDPHOPVOAPG 362 |
| E.coli MoeB | | CVMAPLIGVIGSLQAMDAIKMLACYG | -KPASGKIVMYDAMT | -COFREMKLMRNPGCEVCQQ- 249 |
| C.symb Thif | 291 | CVMAPLIGUIGAIGAINLACIG | -PDLAGRLLHIDLD | PLSFSFVDIAREECOPVCGPG 340 |
| CSUB C1135 | 201 | CVLGVLPGVIGALQAM <mark>B</mark> TIKLIIC | -EPLVGRLLLFDGL | HMSFTELKLRKDPNQVIQGPN 260 |
| HVO 0558 | 192 | CVLGVLPGTVGCIQATEAMKLLLDEG | -EALDGRLLFYDAMD | -MTFETVPYRTNPDCPVCGEG 251 |
| P.furi MoeB | 173 | CIMSYVPPLAAAIAVSLATKILLEE- | EVKSELIFFDTK | TLEFEKIEIPRRDDCPACVRK 230 |
| G kaus MoeBl | 187 | GIISPAVQMVVSYQMABALKILVEDW | -SALRGKLVSFDLWT | -NEYASIRIDGVKKDGOPTOGRH 248 |
| D.acet ThiF | 241 | CLMAPRVGIAAHHQANVVIRLLLCLD 6 | | 272 |
| D.desu Thif | | GTPAPVVACAAALQCTBAAKILTG | -KPPSRGVLFFDLND | -RTFOTVML 284 |
| B.thet ThiF | 171 | GLSKAVRKRLOKMGVKR | KLPVVFSTEO | ADPKAVLLTDERNKKSTCGT 218 |
| | | ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ | GIPVVFSAEKPDP- 1 | KAKLLPLEDEEYERGKVDE 283 |
| | | DLSRAVRTKLKKYKVHD | GIKVVLSVER 1 | ERELLPLKEHQESNPDE 290 |
| M.tube | | DLSYVDIENRTGPSVGLACHLASGVVAADVLLGHG 1 | -VYAAPYFHQFDAYR | SIYVRKRLRCGNRHPLQRVK 290 |
| C.hutc | 293 | NEEKIPYILPMICTDTISKRMKASMVEIPOTITT-WPO 1 | -ASSVVFGGGIGADI | -CRRIILDOFHDSGRYIVDME 364 |
| B.thur | 300 | AAISPNVAIVAGTIVNEALKILTQIA 1 | | -ETNTVSSWEKMLDCPLCGOV 361 |
| | | | | |

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 nitriteoxidizing 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

| June June Johnson Totals Totals <th>•</th> <th></th> | • | | | | | | | | | | | | | | | | | |
|---|---|---|--|--|---|--|---|--|---|---|--|---|---|--|--|---|---|--|
| Success Dial 1 Transaction of control of contro of control of control of control of contro of con | C Human E2A | ٦ | TPAPP | IMPDEKE | LOF | | WSGAR- | SENNT | MV-W | JAVTE- | - PRG | TPFFD | ידיי.דאידי | RTER | ≠ ZPNKPI | # = | VSK | 75 |
| C meto Elli 3 20042000 2012 11 TOMAA - BEYER ALL ALL STRUCTURE ALL STRUC | | | | | | DAPPG | VSASP- | LPDN | | | | | | | | | | |
| C meto Elli 3 20042000 2012 11 TOMAA - BEYER ALL ALL STRUCTURE ALL STRUC | | | | ~ | | DPPPN | ICSAG <mark>P</mark> - | KGDNI | | | | | | | | | | |
| p. f. al. 1 ALA: STILL C. DEPRESSION | | | | | | | CSAGP- | VGDDI | | | | | | IFPTD | YPFKPI | PKISF | | |
| | | | | | | | | | | | | | | | | | | |
| C. maro Elizi 4. Gala Lange Lin Tan Lan. — Devolve Tr. — STERNING - March Lan. — Stellar T. — St | G.lamb E2D | 29 | KMAQK | RIQKELKE | FQK | | | VGDDI | SV-WH | RACIL- | -G-PKD | | | CF <mark>P</mark> SD | Y P F K P I | | etk | 101 |
| | | | | | | | | | | | | | | | | | | |
| T. L. B. 221 2. ADJET TOWN CONCEPT. CONSIDER Constraint | | | | | | | | | | | | | | | | | | |
| A. Lob. 1 6 LAB. M. BUCKE (C | | | | | | EOVPG | IDVVP- | | | | | | | | | | | |
| A. thal B2C 3 0.0712/B. DOLLARS MARKET | | | | | | NHPHG | FVAK <mark>P</mark> - | -5-GTVNI | MV-WI | HCTIP- | | | | | | | | |
| S. cere E2C 1 CONTROLLED. STREETERS CONTROL TO BUILT AND IN CONTRAL TO BUILT AND IN CONTROL TO BUILT AND IN CONTROL TO BUILT A | | | | | | | | | ~ | _ | -G-KEG | TNWAGG- | -VYPITVI | | | | | |
| Juman E201 4. LGALLING, ALMON | | | | | | | | | | | -G-PKD | TPYSGL- | -EIRLSL: | | | | | |
| Human E222 3 OTAL BORNE TODICL. SPECOVYSE SUBJECT 100 Control 100 SPECOVYSE SUBJECT 100 SPECOVYSE SPECOVYSE SUBJECT 100 SPECOVYSE < | | | | | | | FSAGL- | IDDNI | DLY-RV | VEVLI- | IG-PPD | | | | | | | |
| B. cere H232 3 FIAORENE SOLLES BEFEORVESS | | | | | | | FSVGL- | VDDKS | SIF-EV | VEVMI- | | | | | | | | |
| Human EZX 3 HIAVGORGE PERGENE - EFFECTION OF UNDER - UNDER THE - DOUGLA - ENVIREMENT - ENVIREMEN | | | | | | | | | | | | | | | | | | |
| C . PE I EX 2 AVLC 2002 ELG . ENTROY DELLE . ENTROY DELLE . MANNEL . ENTROY DELLE . AVLC . ENTROY DELLE . LEXES . | | | | | | | | | | | | | | | | | | |
| Bunna E224 5 ENERGENETEVUEL - ESCREPT L COLUME FUEL - ELOCIPTE CONTROL - ELOCIPTE CONT | | | | | | | | | | | -G-PKD | TPYE <mark>GG</mark> - | | | | | | |
| S. cere H231 2 BEAKE STUDIES. LEPTICE SED. FOLDUE. FUNCT. C. DEFYLER STUDIES. CONTROL THE SCHULT 12 SCHULT SERVICE SERVICE SERVICE SERVICE SERVICE SERVICE SERVICE. SERVICE S | | | | | | | | | | | | | | | | | | |
| Buman E2M 28 EAADLEUXCHOUNDELD EPRCIENT CONTROL + CONTROL + CONTROL + CONTROL - NUMBER - THE CONTROL - NUMBER - NUMBER - THE CONTROL - NUMBER | | | | | | | | | | | | | | | | | | |
| C. cree EXM 5 8 SAARLEUKOLDORU. LEPYTYLENET = 4-DESCENE. BYLE DESCRIPT. STRUEL DESCRIPT THEOREM. THE - CREEKES AND | | | | | | | | | | | | | | | | | | |
| GEUE C1475 11 NATYERIAL 02 | C.cere E2M | 26 | SAARI | RLKRDLDS | LD | -LPPTVT | LNVI <u>T</u> S | | | EVIVR- | PDE | GYYNY <mark>G</mark> - | | | | | | |
| Human E2.32 11 TYAYO BUOYLE KK | | | | | | | | | | | | | | | | | | |
| S. cere E2.2 4 KOAHEURE YEAHVE - NPPYELBAL - NEUKELE-NYLET - LAURENCE - NYLET - LAURENCE - STUTY - LAURENCE | | 11 | TTATO | LALBYAL LKODYLR | ТŐR | NE | TCAEP- | | | | | | | | | | | |
| Human E2J 9 5PAYNELING AAELEON FUDDYING - LEUNING ON STUTUET FUDDYING ON STUTUET ON TANKET STUTUE ON STUTUE ON TANKET STUTUE ON STUTUET ON TANKET STUTUE ON STUTUE ON TANKET STUTUE ON TAN | | | | | | NPPPY | ILARP- | NEDNI | LE-WH | -TIIYH | | | | | | | | |
| Human E2A 16 HUMAN T20101 10 SERENG - TOTTE V ST - MIT - MI | | | | | | PTDH | IYHAQP- | LEDNI | FE-WH | HFTVR- | -G-PPD | | | /LPPE | y P <mark>mk</mark> pi | | | |
| Fueger 4 TACELARD TRESSERIAL | | | | | | | | | | | | | | | | | | |
| Arthro 7 MEXAGLOS ELSUAD | - | | | | | | | | | | -v- e ir | | | | | | | |
| B. Acid B. AGELSET ENABLESCH | | | | | | | | | | | | | | | | | | |
| B. B. Da 6 DEELTUPLALIPTELIA | | | | | | | | | | | | | | | | | | |
| C perf 15 NDPTMEYKGLIEBNSVURTITIKIT | | | | | | | | | | | | | | | | | | |
| D. arom 55 TPRAAMDLAXILLERARBGDTURE -3MOG-BITWOM | | | | | | | | | | | | | | | | | | |
| B. thet 5.7 TYPEPNISLAWSTESS MOST PERKERS | - | 55 | TPRAAM | IDLAKALL | | | | - 3 MI | GD-L | I VWWM- | | PPA- | RRHIAF | R-11-Q <u>E</u> RGI | | | | 129 |
| Synch 1 MITTERESDATIA MITTERESDATIA MITTERESDATIA MITTERESDATIA B1.copa 12 VRADETWEDADELA MITTERESDATIA MI | | | | | | | | | | | | | | | | | | |
| B. copa 12.1 VRADPTINE DARLATION SKATTURE VALUE UNDER TO | | 53 | | | | | | | | | T. DNO | | | | | | VVE VVE | |
| Human N2A 7 1 <th1< th=""> 1 <th1< th=""> <th1< th=""></th1<></th1<></th1<> | | 121 | | | | | | | | | | | | | | | FSP | |
| S. CERE E2A 76 MPEREVYAN SCIUDILG NEW STATUSIO-SILM DESIGNATIONAL SCIUDILG DESIGNATIONAL SCIUDILG DESIGNATIONAL DESIG | Rhizob | 38 | AFDDQA | ASCAEGQ | ATL | DLAVRLL | ARLYP- | | - V- L2 | AILPL-I | DS-ASS | FOAOAL- | -ERLAKS | [| N | KIGIR | RSG | 100 |
| S. CERE E2A 76 MPERYVAN SCUDDIQ NEW - NEW - TYD ASTLETSIO_SIZM CERTAS SAN - VERATLEFORK - SQYVERVEENT 145 S. CERE E2D 74 MPERINAN NEGDILK DOGS - ALTISKULISIC-SILT DOLGEDIK PETAHITYETOR - SQYVERVEENT G. Lance E2D 74 MPERINAN NEGDILK DOGS - ALTISKULISIC-SILT DARGDELV SQLETITIKE PETAHITYETOR | | | | | | | | | | | | 2.12.1 | and age to 1 | | | | | |
| Human E2D 131 TYPESTINAD WYPE CDUIK Desc ALT SYNULSIC SILT Desc ALT SYNULSIC SILT Desc Desc <thdesc< th=""> <thdes< td=""><td></td><td></td><td>+ +</td><td></td><td>*</td><td></td><td></td><td>#</td><td></td><td></td><td>_</td><td></td><td></td><td>_</td><td></td><td>-</td><td></td><td></td></thdes<></thdesc<> | | | + + | | * | | | # | | | _ | | | _ | | - | | |
| S. cere E2D 74 TYMENINAN | | | | | * GSIC | LDILQ | NR | | | ILTSIQ | SLLD | EPN | PNSPAN- | | | | | |
| P. Fale P. TYTENINTA CALCUDILK ALTLSYNLLSISS-SLIT TOTAL SPLATE G. Lamb P. FALSSON FUENDEDK FRESCONTINT ALTLSYNLLSISS-SLIT TOTAL SPLATE Human P. TONE FRESCONTINT ALTLSYNLLSISS-SLIT AND REALWART TATLSYNLLSISS-SLIT Score P. TONE FRESCONTINT FRESCONTINT ALTLSYNLLSISS-SLIT AND REALWART Score P. TONE FRESCONTINT FRESCONTINT FRESCONTINT ALTLSYNLLSISS-SLIT ALTLSYNLLSISS-SLIT Score P. TONE FRESCONTINT FRE | S.cere E2A | 76 | MFHPN | VYAN | GEIC | LDILQ | NR | WТРТУ | DVAS | ILTSIQ- ILTSIQ- | SLLD | EPN | PNSPAN- PASPAN- | VEAATLF | KDHK- | SQYVKRVK | ETV | 145 |
| G. Lamb E2D 102 IPERTSED | S.cere E2A Human E2D | 76 133 | MFHPN IYH <mark>C</mark> N | VYAN INSQ | GEIC | LDILQ | NR DN | WTPTY WSPAI | ZDVASI JTISKV | ILTSIQ- ILTSIQ- VLLSIC- | SLLD | EPN DPN DCN | PNSPAN- PASPAN- PADPLV- | VEAATLF GSIATQY | KDHK- LTNR- | SQYVKRVK AEHDRIAR | ETV QWT | 145 202 |
| Human E2N 76 TYTERYDKL | S.cere E2A Human E2D S.cere E2D C.mero E2D | 76 133 74 104 | MFHPN IYHCN IYHPN IYHCN | VYAN INSQ INAN INSQ | GEIC GVIC GNIC GQIC | LDILQ LDILK LDILK LDTLK | DNR DN | WTPTY WSPAI WSPAI WSPAI | ZDVASI JTISKA JTLSKA JTISKA | ILTSIQ ILTSIQ /LLSIC /LLSIC /LLSIC | SLLD | EPN DPN DCN DAN DAN | PNSPAN- PASPAN- PADPLV- PDDPLV- PHDPLV- | VEAATLF GSIATQY PEIAHIY GSIAKEY | KDHK- LTNR- KTDR- LTNR- | SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR | ETV QWT EWT EWT | 145 202 143 173 |
| S. cere EX. TYPERTORL CREADURY CREADURY <thcreadury< th=""> <thcreadury< th=""> <thcre< td=""><td>S.cere E2A Human E2D S.cere E2D C.mero E2D P.falc E2D</td><td>76 133 74 104 73</td><td>MFHPN IYH<mark>C</mark>N IYHPN IYH<mark>C</mark>N IYHPN</td><td>VYAN INSQ INAN INSQ</td><td>GEIC GVIC GNIC GQIC</td><td>LDILQ LDILK LDILK LDTLK</td><td>DN DQ DQ DN</td><td>WTPTY WSPAI WSPAI WSPAI WSPAI</td><td>DVASI JTISKV JTLSKV JTISKV JTISKV</td><td>ILTSIQ ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIS</td><td>SLLD SLFN SLLT SLLT SLLT SLLT</td><td><mark>E</mark>PN DPN DCN DAN DAN DAN</td><td>PNSPAN- PASPAN- PADPLV- PDDPLV- PHDPLV- ADDPLV-</td><td>VEAATLF GSIATQY PEIAHIY GSIAKEY PEIAHVY</td><td>KDHK- LTNR- KTDR- LTNR- KTDR-</td><td>SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR TKYHQTAK</td><td>ETV QWT EWT EWT</td><td>145 202 143 173 142</td></thcre<></thcreadury<></thcreadury<> | S.cere E2A Human E2D S.cere E2D C.mero E2D P.falc E2D | 76 133 74 104 73 | MFHPN IYH <mark>C</mark> N IYHPN IYH <mark>C</mark> N IYHPN | VYAN INSQ INAN INSQ | GEIC GVIC GNIC GQIC | LDILQ LDILK LDILK LDTLK | DN DQ DQ DN | WTPTY WSPAI WSPAI WSPAI WSPAI | DVASI JTISKV JTLSKV JTISKV JTISKV | ILTSIQ ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIS | SLLD SLFN SLLT SLLT SLLT SLLT | <mark>E</mark> PN DPN DCN DAN DAN DAN | PNSPAN- PASPAN- PADPLV- PDDPLV- PHDPLV- ADDPLV- | VEAATLF GSIATQY PEIAHIY GSIAKEY PEIAHVY | KDHK- LTNR- KTDR- LTNR- KTDR- | SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR TKYHQTAK | ETV QWT EWT EWT | 145 202 143 173 142 |
| A. thal E2I 82 FFHENVPSGYG_SILLEDYGREATTKKOLUSGO-DLDTDRADEACTDGYHLFCODPVEYKKRKUGS 153 S. cere E3C 81 MHENVDKSGYG_SILLEDYGREATTKKOLUSGO-DLDSDENSEGSDENSESPENKARVDKKULGA 152 A. thal E2C 107 GFHENVDY | S.cere E2A Human E2D S.cere E2D C.mero E2D P.falc E2D G.lamb E2D | 76 133 74 104 73 102 | MFHPN IYHCN IYHPN IYHCN IYHPN IFHPN | VYAN INSQ INAN INSQ INTA ISED | GEIC GVIC GNIC GQIC | LDILQ LDILK LDILK LDTLK | DN DQ DQ DN DQ DQ | WTPTY WSPAI WSPAI WSPAI WSPAI WSPVI | (DVAS) JTISK JTLSK JTISK JTISK JTISK | ILTSIQ ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIS ILLSIC | SLLD SLFN SLLT SLLT SLLT SLLT SLLD | <mark>E</mark> PN DPN DCN DAN DAN DPN DPN | PNSPAN- PASPAN- PADPLV- PDDPLV- PHDPLV- ADDPLV- PDDPLV- PDDPLN- | VEAATLF GSIATQY PEIAHIY GSIAKEY PEIAHVY SAAARLL | KDHK- LTNR- KTDR- LTNR- KTDR- KTDR- | SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR TKYHQTAK ENYIRTVK | ETV QWT EWT EWT AWT | 145 202 143 173 142 171 |
| S. cere E2I 81 FYHENVYPS | S.cere E2A Human E2D S.cere E2D C.mero E2D P.falc E2D G.lamb E2D Human E2N C.mero E2N | 76 133 74 104 73 102 76 76 | MFHPN IYH <mark>C</mark> N IYHPN IYHCN IYHPN IFHPN IYHPN IYHPN | VYAN INSQ INAN INSQ ISED VDKL IDRL | GEIC GVIC GNIC GQIC GXIC GVIC ERIS | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK | DR DQ DQ DQ DQ DQ DQ KE DK | WTPTY WSPAI WSPAI WSPAI WSPAI WSPVI WSPAI WSPAI | (DVAS) JTISK JTISK JTISK JTISK JTISK JUSK JUSK JUSK JUSK | ILTSIQ ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIC /LLSIQ /LLSIQ | SLLD SLFN SLLT SLLT SLLT SLLD ALLN | | PNSPAN- PASPAN- PDDPLV- PDDPLV- PHDPLV- ADDPLV- PDDPLN- PDDPLA- PEDALN- | VEAATLF GSIATQY PEIAHIY GSIAKEY PEIAHVY SAAARLL NDVVEQW NEAAELW | KDHK- LTNR- KTDR- LTNR- KTDR- KTDK- KTNE- KKDI- | SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR TKYHQTAK ENYIRTVK AQAIETAR ARAKEIAQ | ETV QWT EWT EWT AWT AWT QWT | 145 202 143 173 142 171 145 145 |
| S. cere E2C 81 MMTERVOKS | S.cere E2A Human E2D S.cere E2D C.mero E2D P.falc E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N | 76 133 74 104 73 102 76 76 75 | MFHPN IYHCN IYHPN IYHCN IYHPN IFHPN IYHPN IYHPN IYHPN | VYAN INSQ INSQ INTA ISED VDKL IDRL IDRL | GEIC GVIC GNIC GQIC GAIC GVIC ERIS GRVC GRIC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LDILK | DR DQ DQ KE DR DR DR | WTP TY WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI | /DVASI JTISK\ JTISK\ JTISK\ JTISK\ JTIAKI JQIRT\ JQIRT\ JQIRT\ | ILTSIQ- ILTSIQ- /LLSIC- /LLSIC- /LLSIC- /LLSIC- /LLSIQ- /LLSIQ- /LLSIQ- | SLLD SLFN SLLT SLLT SLLT SLLT SLLD ALLN ALLA | BPN DQN DAN DAN DAN DPN DPN | PNSPAN- PADPLV- PDDPLV- PDDPLV- ADDPLV- PDDPLN- PDDPLA- PEDALN- PNDPLA- | VEAATLF GSIATQY PEIAHIY GSIAKEY PEIAHVY SAAARLL NDVVEQW NEAAELW NDVAEDW | KDHK- LTNR- KTDR- KTDR- KTDR- KTDK- KTDE- KKDI- I KNE- | SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR TKYHQTAK ENYIRTVK AQAIETAR - ARAKEIAQ QGAKAKAR | ETV QWT EWT EWT AWT KYT AWT QWT EWT | 145 202 143 173 142 171 145 145 144 |
| Human E2G1 78 TMHENUDXH | S.cere E2A Human E2D S.cere E2D C.mero E2D P.falc E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I | 76 133 74 104 73 102 76 76 76 75 82 | MFHPN IYHCN IYHCN IYHCN IYHPN IFHPN IYHPN IYHPN FFHPN | VYAN INSQ INSQ INTA ISED VDKL IDRL IDRL VPS | GEIC GVIC GNIC GQIC GAIC GVIC ERIS GRVC GRIC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LDILK | DN DQ DN DQ DQ DQ DQ DQ DQ DQ | WTP TY WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI | (DVAS) JTISK JTISK JTISK JTISK JURT JURT JURT JURT JURT | ILTSIQ ILTSIQ VLLSIC VLLSIC VLLSIC VLLSIC VLLSIQ VLLSIQ VLLSIQ ILVGIQ | SLLD SLFN SLLT SLLT SLLT SLLT SLLD ALLN ALLN ALLA DLLD | | PNSPAN- PASPAN- PADPLV- PDDPLV- ADDPLV- PDDPLN- PDDPLA- PEDALN- PADPLA- PADPAQ- | VEAATLF GSIATQY PEIAHIY GSIAKEY PEIAHVY SAAARLL NDVVEQW NEAAELW NDVAEDW TDGYHLF | KDHK- LTNR- KTDR- KTDR- KTDR- KTDK- KTNE- KKDI- IKNE- CQDP- | SQYVKRVK - AEHDRIAR - PKYEATAR - RKHDETAR - TKYHQTAK - ENYIRTVK - AQAIETAR - AQAIETAR - QGAKAKAR - VEYKKRVK | ETV QWT EWT AWT XYT AWT QWT EWT LQS | 145 202 143 173 142 171 145 145 144 153 |
| S pomb E201 78 IMHENVHPN GEV GISTLHAP-9-AGERAGE WISPETLISVI-SKLSSENDEG ANDAAKEFRENPOEFKKEVRRUV 160 Human E22 77 MFHENTYPD GEV GISTLHAP-9-AGERAGE WOSVEKILLSVV-SKLAENTESGANDAAKEFRENPDEFKKEVRRUV 160 S. cere E22 77 ILHENTYPD GEV GISTLHAP-9-AGERAGE WOSVEKILLSVV-SKLA | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C | 76 133 74 104 73 102 76 76 76 75 82 81 107 | MFHPN IYHCN IYHPN IYHCN IYHPN IYHPN IYHPN IYHPN FFHPN FFHPN CFHPN | VYAN INSQ INSQ ISED VDKL IDRL VYPS VYPS VYPS | GEIC GVIC GNIC GAIC GVIC ERIS GRVC GRIC GTVC GNIC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LDILQD | DN DQ DN DQ DQ DQ DQ DQ DQ DQ | WTP TY WSP AI WSP AI WSP AI WSP VI WSP AI WSP AI WSP AI WSP AI WRP AI WRP AI | IDVAS TISK TISK TISK TISK JISK JISK JISK JISK JISK JISK JISK ISK ISK ISK ISK JISK J | ILTSIQ- ILTSIQ- /LLSIC- /LLSIC- /LLSIC- /LLSIQ- /LLSIQ- /LLSIQ- ILVGIQ- ILLSIQ- | SLLD SLFN SLLT SLLT SLLT SLLT SLLD ALLN ALLA DLLD | EPN | PNSPAN- PASPAN- PDDPLV- PDDPLV- PHDPLV- ADDPLV- PDDPLN- PDDPLA- PNDPLA- PNDPLA- PADPAQ- PNSPAQ- ISSPLN- | VEAATLF -GSIATQY -DEIAHIY -GSIAKEY -PEIAHY -SAAARLL -NDVVEQW -NEAAELW -NDAELW -NDVAEDW -TDGYLF -TDGYLF -TQAAQLW | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KTDK- KKDI- IKNE- CQDP- SRNK- SNQE- | SQYVKRVK AEHDRIAR PKYEATAR RKHDETAR TXYHQTAK ENYIRTVK AQAIETAR QGAKAKAR VEYKKVK AEYDKKVL EYRKMVEK | ETV QWT EWT AWT AWT QWT EWT LQS LQA LYK | 145 202 143 173 142 171 145 145 145 145 152 152 |
| Human E2G2 77 MEHENITYPD | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2C | 76 133 74 104 73 102 76 76 75 82 81 107 81 | MFHPN IYHCN IYHPN IYHPN IYHPN IYHPN IYHPN IYHPN FFHPN FFHPN CFHPN MWHPN | VYAN INSQ INSQ ISED VDKL IDRL IDRL VYPS VYPS VDVY VDKS | GEIC GVIC GNIC GAIC GVIC ERIS GRVC GRIC GTVC GNIC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LDILQD LDILQ | NR DN DQ DN DQ KE DK DR VG K K K | WTP TY WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI WSP AI WRP - AI WRP - AI WRP - AI WRP - AI | IDVAS TISK TISK TISK TISK JIRT JURT JURT JURT TIKQI ITLKQI IDVRT INVET | ILTSIQ ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIC /LLSIQ /LLSIQ ILLSIQ ILLSIQ ILLSIQ | - SLLD - SLLT - SLLT - SLLT - SLLT - ALLA - ALLA - DLLD - SLLG - SLLG | EPN | PNSPAN- PASPAN- PDDPLV- PHDPLV- ADDPLV- PDDPLN- PDDPLA- PEDALN- PADPAQ- ISSPLN- NRSPLN- | VEAATLF -GSIATQY -PEIAHIY -GSIAKEY -PEIAHVY -SAARLL -NDVVEQW -NEAAELW -NDVAEDW -TDGYHLF -EPAWRSF -TCAAQLW -AVAAELW | KDHK- LTNR- KTDR- LTNR- KTDK- KTDK- KTDE- KKDI- IKNE- CQDP- SRNK- SNQE- DADM- | SQYVKRVK AEHDRIAR PKYBATAR RKHDETAR TKYHQTAK BNYIRTVK AQAIETAR QGAKAKAR QGAKAKAR VEYKKRVK BEYRKWEK EYRKWEK EYRKKVL | ETV QWT EWT AWT AWT QWT EWT LQS LQA LYK ACY | 145 202 143 173 142 171 145 145 145 153 152 176 150 |
| Human E2K 79 IMMEDISYTAITGLDILKDWAAAANTLIKTULSLQ-ALLS | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N A.thal E2I S.cere E2I A.thal E2I S.cere E2C Human E2G1 | 76 133 74 104 73 102 76 76 75 82 81 107 81 78 | MFHPN IYHCN IYHPN IYHPN IYHPN IYHPN IYHPN IYHPN FFHPN FFHPN CFHPN MWHPN IWHPN | VYAN INSQ INAN INTA INTA VDKL IDRL IDRL VYPS VYPS VDVY VDVY VDKS VDKN | GEIC GVIC GNIC GQIC GAIC GVIC GRIC GTIC GNIC GNIC GNIC GUVC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LDILQD LDILQD- ISILHEP- | DN DQ DQ DQ CR DR DR VG K K K K K EK | WTP TY WSP AI WSP AI WSP AI WSP VI WSP AI WSP AI WSP AI WRP AI WRP AI WSS AY WSS VY WLP IF | IDVAS TISK TISK TISK TISK JIR JURT JURT JURT ITLKQ ITLKQ ITLKQ INVET | ILTSIQ ILTSIQ ALLSIC ALLSIC ALLSIC ALLSIQ ALLSIQ ALLSIQ ILLSIQ ILLSIQ ILLSIQ ILLSIQ ILLSIQ | SLLD SLFN SLLT SLLT SLLT SLLT ALLN ALLN ALLN DLLD SLLG SLLG SLLG SLLG | | PNSPAN- PADPLV- PDDPLV- PDDPLV- PDDPLV- PDDPLV- PDDPLA- PDDPLA- PADPAQ- PNSPAQ- ISSPIN- NRSPLN- GDSPAN- | VEAATLF -GSIATQY -PEIAHIY -SIAKEY -PEIAHVY -SAAARLL -NDVVEQW -NDVAEDW -NDVAEDW -TDGYHLF -EPAWRSF -TQAAQLW -AVAAELW -VDAAKEW | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KTDE- KKDI- IKNE- CQDP- SRNK- SNQE- DADM- REDR- | SQYVKRVK AEHDRIAR PKHDETAR TKYHQTAK -ENYIRTVK AQAIETAR QGAKAKAR QGAKAKAR VEYKKRVK AEYDKKVL -EEYRKKVL -EEYRKKVL -NGEFKKVL | ETV QWT EWT AWT XYT QWT EWT LQS LQA LYK ACY ARC | 145 202 143 173 142 171 145 145 145 153 152 176 150 160 |
| C. crin E2K 76 WHENVSSQSGATGLDTLKDQSEGATTLKTALLSLQ-ALLSSECEDDEDCAUVAKQVISDHEYYKKTAKWT 146 S. cree E2K 75 VYHENISSVTGATGLDTLK | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 | 76 133 74 104 73 102 76 75 82 81 107 81 78 78 78 78 77 | MFHPN IYHCN IYHPN IYHCN IYHPN IYHPN IYHPN FFHPN FFHPN CFHPN IWHPN IWHPN MFHPN | VYAN INSQ INSQ INSQ ISED VDKL ISEL IDRL VYPS VYPS VYPS VDVY VDKN YPD | GEIC GVIC GNIC GQIC GAIC GVIC GRVC GTVC GTIC GNIC GNIC GDVC GUVC GEVC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LDILQD LDILQD- ISILHEP- ISILHEP- | | WIT P TY WS P AI WS P AI WS P AI WS P AI WS P VI WS P AI WS P AI WR P AI WR P AI WS S AY WS S YY WS A TH WI P TH WI P V WS P V | IDVAS TISK TISK TISK TISK TISK JISK JUSK JUSK JUSK JUSK TISK TISK TISK TVK TVK TVK TVK TVK TVK TVK TVK TVK TV | ILTSIQ- ILTSIQ- /LLSIC- /LLSIC- /LLSIS- /LLSIS- /LLSIQ- /LLSIQ- IVLGVQ- ILLSIQ- ILLSIQ- IMISVI- ILLSV- | - SLLD - SLLT - SLLT - SLLT - SLLT - SLLD - SLLD - ALLA - SLLG - SLLG - SMLA - SMLA | EPN | PNSPAN- PADPLV- PDDPLV- PDDPLV- PDDPLV- PDDPLA- PDDPLA- PNDPLA- ISSPLN- ISSPLN- RSPAQ- ISSPLN- GDSPAN- DESPAN- DESPAN- | VEAATLF GSIATQY BEIAHIY SIAKEY NEAAELW NDVAEDW NEAAELW NDYAEDW TDGYHLF FDAWRSF TCAAQLW TCAAQLW VDAKEW IDAAKEW IDAAKEW | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KKDI- IKNE- CQDP- SRNK- SNQE- DADM- REDR- RENP- RDDR- | SQYVKRVK AEHDRIAR -PKYEATAR -RKHDETAR -TKYHQTAK -NRYHQTAK -AQAIETAR -QAAKARR -VEYKKKVK -AEYDKKVL -EYYKKWEK -EYYKKVK -EYYKKVK -QEFKKVW -QEFKKRVR -QEFKKRVR -QEFKKRVR | ETV QWT EWT AWT KYT AWT QWT LQS LQA LYK ACY ARC RLV QIV | 145 202 143 173 142 171 145 145 145 153 152 176 150 160 159 |
| S. cere E2K 75 VMTENTSVTSATGLDTLKNAGGEVTTLKSLIGLO-ALDQSEENDDQDAEVAQHYLADRESFNKTAALWT 145 Human E2H 74 IFHENTDLASGYVGLDVINGYWTALYDLTNIFESFLPQLLAYDNIDGLNGDAAMYLHAPRESFNKTAALWT 145 S. cere E2H 72 IFHENTDLASGYVGLDVINSTWGELYDLTNIFESFLPQLLAYDNIDGLNGDAAMYLHAPREFYKQKIKEYI 143 Human E2M 99 VMTENTDLEGNVGLDVINSTWGELYDLTNIFESFLPQLLAYDNIDGLNKQAASQMINDIKLYEEKIKEYI 143 C. cere E2M 103 IFHENTDLKGNVGLDVIR | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2C Human E2G1 S.cere E2G2 | 76 133 74 104 73 102 76 75 82 81 107 81 78 78 78 77 77 | MFHPN IYHCN IYHPN IYHPN IYHPN IYHPN IYHPN FFHPN FFHPN CFHPN IWHPN IWHPN IWHPN IWHPN IWHPN | VYAN INSQ INTA INTA ISED VDKL IDRL VYPS VYPS VDVY VDKS VDKS VDKM IYPD IYPN | GEIC GVIC GNIC GQIC GQIC GVIC GRIC GRIC GNIC GNIC GEVC GEVC GEVC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LSILNED- LSILNED- LSILHEP- ISILHEP- ISILHSP- | | $\begin{split} & w_{11} \mathbf{F} = -\mathbf{T}_{11} \\ & w_{12} \mathbf{F} = -\mathbf{A}_{11} \\ & w_{12} \mathbf{F} = -\mathbf{A}_{12} \\ & w_{13} \mathbf{F} = -\mathbf{A}_{12} \\ & w_{13} F$ | VDVAS JISK JISK JISK JISK JISK JISK JUSK JUSK JUSK JUSK JUSK JUSK JUSK JU | ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIC /LLSIC /LLSIQ /LLSIQ ILLSQ ILLSQ ILLSVI ILLSVI ILLSVM | SLLD SLFN SLLT SLLT SLLT SLLD ALLN ALLN ALLN SLLG SLLG SLLG SLLG SLLG SSLLG SSLLG SSMLA SSMLA SSMLA SSMLA SSMLA | E PN | PNSPAN- PASPAN- PDPLV- PDPLV- PDPLV- PDPLA- PDPLA- PNDPLA- PNDPLA- PADPAQ- INSPLN- GDSPAN- GDSPAN- DESPAN- DESPAN- DESPAN- IESGAN- IESGAN- | VEAATLF - GSIATQY - PEIAHY - GSIAKEY - PEIAHVY - SAAARLL - NDVVEQW - NDVEQW - NDAELW - TDGYHLF - EPAWRSF - TCAAQLW - VDAAKEW - VDAAKEW - VDASKW | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KTNE- CQDP- SRNK- SNQE- DADM- REDR- REDR- REDR- REDR- RDRR- | SQYVKRVK BEHDRIAR PKYEATAR RKHDETAR RKHDETAR RKHDETAR RKHDETAR RALIETAR AQAIETAR QGAKAKAR VGYKKRVK BEYRKWZK BEYRKWZK DGEFKRVZ DEFKRVZ DEFKRVZ DEFKRVZ | ETV QWT EWT AWT KYT AWT QWT LQS LQA LYK ACY ARC RLV QIV LSI | 145 202 143 173 142 171 145 145 145 152 176 150 160 159 159 |
| Human E2H 74 IFFENTDEASFTVGLDVINOTWIALVDTINTEFSFLEQLIAYENGTDEANDEAAAWYLHRPEFYKQKIKEYI 145 S.cere E2H 72 IFFENTDIASGNTGLDVINSTWSELVDLINTVEWMIPGELKEDNSDEINNEAATLQLRCKKLYEEKIKEYI 145 S.cere E2H 72 IFFENTDIASGNTGLINTR | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 S.cere E2C2 | 76 133 74 104 73 102 76 76 75 82 81 107 81 78 78 78 77 77 79 | MFHPN IYHCN IYHPN IYHPN IYHPN IYHPN IYHPN IYHPN CFHPN WHPN IWHPN IWHPN IWHPN IWHPN IIHPN | VYAN INSQ INSQ INTA ISED VDKL IDRL IDRL VPFS VYFS VYFS VDVY VDKN VDKN IYPD ISSVT | GEIC GVIC GVIC GQIC GQIC GVIC GRIC GRIC GTIC GNIC GUIC GEVC GEVC GEVC GEVC GEVC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LSILNED- ISILHEP- ISILHEP- ISILHEP- LSILHAP- LSILSA | | WIT P TY WS P AI WS P AI WS P AI WS P VI WS P VI WS P VI WS P AI WR P AI WS AI WS AI WS VI WI P VI WS P VC WS P VC WS VC WS VC | IDVASI JISK JISK JISK JISK JISK JISK JISK JI | ILTSIQ ILTSIQ ALSIC ALSIC ALSIC ALSIC ALSIC ALSIQ ILLSIQ ILLSIQ ILLSQ ILLSV ILLSV ILLSV ILLSV ILLSV | SLLD SLLT SLLT SLLT SLLT SLLT ALLN ALLN DLLD SLLG SLLG SLLG SMLA SMLA SMLA SMLA SMLA | E PN | PNSPAN- PASPAN- PDDFLV- PDDFLV- PDDFLA- PDDFLA- PDDFLA- PADPAQ- ISSPIN- NRSFLN- GDSPAN- DESCAN- IESCAN- IESCAN- ESCAN- | - VEAATLF - GSIATCY - GSIATCY - GSIATCY - PEIAHIY - SAARLL - NDVYEOW - NEAAELW - NDVAEW - TOGYHLF - TOGYHLF - TOGAQLW - VDAAEW - VDAAEW - VDAAKW - VDAAKW - IDAALW - IDAACLW - VDAAKW | KDHK- LTNR- KTDR- KTDR- KTDK- KTNE- KKDI- IKNE- CQDP- SRNK- SNQE- DADM- REDR- REDR- REDR- RDDR- AVIV- | - SQYVKRVK - AEHDRIAR - PKYEATAR - RKHDETAR - TKYHQTAK - AQAIETAR - AQAIETAR - QAKAKAR - VEYKKRVK - AEYKKRVK - BYKKRVK - EYRKKVK - SEYRKKVL - OZFKKRVR - OZFKKRVR - DQFYKIAK - PQFYKIAK - PLSXSSWD | ETV QWT EWT AWT AWT QWT EWT LQS LQA LYK ACY ACY ACY QIV LSI VET | 145 202 143 173 142 171 145 145 145 153 152 176 150 160 160 159 159 149 |
| Human E2M 99 VYHENIDLEGNVCLNILRFDRKEVITINSIIYGLQ-YLFLEDRED LN | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N A.thal E2I S.cere E2N A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 S.cere E2G2 Luman E2K C.rein E2K | 766 1333 744 1044 733 1022 766 755 822 811 1077 811 788 788 778 777 779 766 | MFHPN IYHCN IYHPN IYHPN IYHPN IYHPN IYHPN FFHPN CFHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN | VYAN INSQ INTA INTA IDRL UDRL UDRL VYFS VYFS VJYS VDYS VDKN IYPD IYPD IYPD USSVT VSSQS | GEIC GVIC GVIC GQIC GQIC GVIC GRVC GTIC GTIC GNIC GDIC GEVC GEVC GEVC GAIC GAIC GAIC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- LSILNED- ISILHEP- ISILHEP- ISILHEP- ISILHEP- LDILK LDILK | | WIIP | IDVAS ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE ITISE | IIITSIQ JILSIC JLSIC JLSIC JLSIC JLSIC JLSIC JLSIQ JLSIQ IILSIQ IILSIQ IILSIQ IILSIQ ILLSV IILSV ILLSV JLLSLQ ALLSQ | SLLD SLFN SLLT SLLT SLLT SLLT SLLD ALLN ALLA SLLG SLLG SLLG SLLG SMLA SMLA SMLA ALLA ALLA ALLQ | | PNS PAN- PADPLV- PDDPLV- PDDPLV- PDDPLN- PDDPLN- PDDPLA- PADPAQ- ISSPIN- NRSPIN- GDSPAN- GDSPAN- DESGAN- IESGAN- IESGAN- IESGAN- | VEAATLE | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KKDI- IKNE- CQDP- SRNK- SRNK- SRNK- DADM- REDR- REDR- RDDR- RDDR- RDNR- AVIV- ISDH- LRDR- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NALHDETAR - ARAKEIAQ - QGAKAKAR - VEYKKRVK - AEYDKKVL - EYRKMVEK - DEYRKNVK - QEFKKRVR - QEFKKRVR - DEFFROVK - DEFFROVK - LSSKSMD - ESFNKTAA | ETV QWT EWT EWT AWT KYT AWT LQX LQA LQA LQA ACY QUV LSI VET YWT LWT | 145202202143317314221711451445153144515315221766150160015991599149914661455 |
| T.ther E2M 103 TYHENIDLQGNVGINTLRDBWGVINNIVNILGGL-FLFIENENDGINKQAASQMINDIKAFEADVKKSL 122 C.cere E2M 103 TFHENIDLKGNVGINTLRBAGESALDLQSITGLL-FLFIBNNDGINKQAASQMINDIKAFEADVKKSL 122 T.vagi E2M 82 LWHENIGLYGPVGINTLR | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 S.cere E2C2 Human E2K C.rein E2K S.cere E2K | 766 1333 744 733 1022 766 755 822 811 1077 811 788 778 778 779 766 755 744 | MFHPN IYHCN IYHCN IYHCN IYHCN IYHPN IYHPN IYHPN IYHPN CFHPN WWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN | VYAN | GEIC GVIC GVIC GQIC GQIC GVIC GRVC GRVC GTIC GNIC GEVC GEVC GAIC GAIC GAIC GAIC GAIC GAIC GAIC | LDILQ LDILK LDILK LDILK LDILK LDILK LDILK LDILK LSILNED- LSILNED- ISILHEP- ISILHEP- LSILHAP- LDILK LDILK LDILK LDILK LDILK LDILK | | WILT P TY WS P - AI WS P - AI WS P AI WS P AI WS P VI WS P AI WS P AI WS P AI WS P AI WS P AI WS P VI WS P AI WS P AI WS P AI WS P AI WS P VI | IDVAS TISKU TISKU TISKU JISKU JISKU JUTSKU JUTSKU JUTSKU ITVKU ITVKU ITVKU ITVKU ISPET 2SVEKU 2SVEKU ISPET 2SVEKU ISPET 2SVEKU ISPET | IIITSIQ IIITSIQ VILSIC VILSIC VILSIC VILSIC IILSIQ VILSIQ IIVGIQ IIVGQ IIVGQ IILSIQ IILSVI IILSVI IILSV VILSIQ VILSIQ VILSIQ VILSIQ | SLLD SLFT SLLT SLLT SLLT SLLD ALLN ALLA SLLG SLG | | PNS PAN- PAD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUA- PDD PUA- PND PLA- PND PAQ- PNS PAQ- PNS PAQ- PNS PAQ- PNS PAQ- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- PDD PDD- PDD PDD- PND PDD- PID PDD- | | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KKDI- IKNE- CQDP- SRNK- SNQE- DADM- REDR- REDR- REDR- RDDR- RDDR- AVIV- ISDH- LRDR- LHRP- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NGALETAR - QAALETAR - QAALETAR - QAALETAR - QAFKKVK - BYRKKVK - BYRKKVK - BYRKKVK - NGEFKRVK - QEFKKKV - QEFKKKV - PEFERQVK - PEFERQVK - ESFNKTAR - ESFNKTAR - ESFNKTAR | ETV QWT EWT EWT AWT QWT LQX LQA LQA LQA ACY ACY LQX LQIV LSI VET LWT EYI | 1452 2022 1433 1422 1711 1455 1445 1532 1502 1500 1600 1599 1499 1491 1455 1455 1455 1455 1455 |
| C.cere E2M 103 TFHENIOLKGNVGLNILRCAUSEALDLOSITGLL-FLFLBINNDELN | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N A.thal E2I S.cere E2N A.thal E2C S.cere E2C Human E2G1 S.uere E2G2 Human E2G2 S.cere E2G2 Human E2K C.rein E2K S.cere E2K S.cere E2K | 766 1333 744 73 1022 766 765 822 811 1077 81 78 81 78 81 77 77 77 79 96 75 74 72 | MFHPN IYHCN IYHCN IYHCN IYHPN IYHPN IYHPN IYHPN CFHPN IYHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IIHPN IWHPN IFHPN | VYAN | GEIC GVIC GVIC GQIC GVIC GVIC GRIC GRIC GRIC GNIC GNIC GRIC GRIC GRIC GAIC GAIC GIVC GIVC GIVC GIVC GIVC GIVC GIVC GIVC GIVC | LDILQ | | W II P T X W I P AI W W S V W S V - V W S V W S V - V - V W S V - V - V W S V - V - V - V - V - V - V - V - | IDVAS ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITISKU ITVKU ITVKU ITVKU ITVKU ITVKU ITVKU ITVKU ISPET ISPET ISPET ISPET ISPET ISPET ISPET ISPET ISPET ITLKSI ITLKSI ITLKSI | ILTSIQ /LLSIC /LLSIC /LLSIC /LLSIC /LLSIC /LLSIQ /LLSIQ /LLSIQ ILLSIQ ILLSIQ ILLSV ILLSV /LLSQ ALLSQ ALLSQ ALLSQ ILLSQ IFESFL IVEWMII | SLLD SLLT SLLT SLLT SLLT SLLT ALLD ALLA DLLD SLLG SLLG SMLA SMLA SMLS ALLA ALLA SMLS ALLA QLLA QLLA QLLA | | PNS FAN PADFLV- PDD FLV- PDD FLV- PDD FLV- PDD FLN- PDD FLN- PDD FLA- PND FLA- PND FLA- PND FLA- PND FAA- ISS FLN- GDS FAN- DES FAN- DES FAN- DES FAN- DES FAN- DES FAN- DES FAN- DES FAN- DES FAN- PDD FOD- PDD FOD- PDD FOD- SD FLN- GSD FLN- | VEAATLE GSIATU GSIAKEY SIAKEY SIAKEY NDVAEV NDVAEV NDVAEV NDVAEV TDGYHLE TDGYHLE TDGYHLE TDAKEW VDAKEW VDAKEW VDAKEW VDAKEW VDAKEW VDAKEV VDAKEV VDAKEV AVVAVAN AVVAVAN AVVANON AVVANON AVVANON AVVANON AVVANON AVVANON AVVANON | KDHK- LTNR- KTDR- LTNR- KTDR- KTDK- KTDK- KTNE- KKDI- IKNE- CQDP- SRNK- SNQE- DADM- RENP- RDNR- RDNR- RDNR- AVIV- ISDH- LRDR- LRDK- | - SQVYKRVK - ABHDRIAR - PKYEATAR - PKYEATAR - PKYHQTAK - ENYIRTVK - AQAIETAR - AQAIETAR - AQAIETAR - AGYDKWL - AGYDKWL - BYYKRWVK - EYYKRWVK - EYYKKVK - PEFERVK - PEFERVK - PEFERVK - EYYKKTAK - EYYKKTAK | ETV QWT EWT EWT XWT QWT LQS LQA LQX LQX LQX LQX LQX LQX LUX LUX LSI LWT LWT EYI EYI | 1452 2022 1433 1422 1711 1455 1445 1532 1766 1500 1600 1599 1499 1491 1455 1455 1455 1455 1455 |
| CSUB_C1475 79 IMHENPSDSVPARVGES FKPHGSESLRTAVIEGLE-NLLTNEXTEDELNPCAAFEYKNRPDLFYSRUROFV 150 Human E2J2 84 NGRFKCNTRLGLSTDFH-1DTMRGKSVSTLINGLL-SFWKGGTLGSIETSDFTKCQLAVQSLAFILLKI I C.ccre E2J2 77 NGRFKCNTRLGLSTDFH-1DTMRGKSVSTLINGLL-SFWKGGTLGSIDTSDFTKCQLAVQSLAFILLKI I A.thal E2J 81 NGRFEKGKKIGLSSSDYH-1TTMRGKSVSTLINGLL-SFWTSBAGALGSLDYFDEERKALAKKSOPFCCGC IS2 A.thal E2J 81 NGRFETKKIGLSSSOH-1TWGCSSVSTALLAII-GPMPTSBAGALGSVDYFDEERKALAKKSOPFCCGC IS2 A.thal E2J 83 NGRFET | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2C S.cere E2I A.thal E2C S.cere E2C Human E2G1 S.cere E2G2 Human E2K S.cere E2C Human E2K S.cere E2K Human E2H | 766 1333 744 73 102 766 766 768 822 811 1077 811 788 788 777 779 766 755 744 7299 | MFHPN IYHCN IYHCN IYHCN IYHPN IYHPN IYHPN IYHPN CFHPN CFHPN MFHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN IWHPN | YYAN | GEIC GVIC GQIC GQIC GAIC GRIC GRIC GRIC GTIC GRIC GRIC GAIC GAIC GAIC GAIC GSIC GNIC GSIC GIC | LDILQ | | $ \begin{split} & \forall x \\ x \\$ | VDVAS TISKV TISKV TISKV TISKV TISKV QIRTV QIRTV QIRTV TVKQQ TVKQ VDVRT TVKQQ VDVRT TVKQ VDVT TVKQ VDVRT TVKQ TVKQ TVKQ TVKQ TVKQ TVKQ TVKQ TVK | ILTSIQ ILTSIQ ILLSIC ILLSIC ILLSIC ILLSIC ILLSIQ ILLSIQ ILLSIQ ILLSQ ILLSQ ILLSV ILS | SLLD SLLT SLLT SLLT SLLT SLLT ALLN ALLN DLLD DLLD DLLD SLLG SLLT SLT | EPN | PNSPAN- PADPIV- PDDFIV- PDDFIV- PDDFIV- PDDFIN- PDDFIN- PDDFIN- PDDFIN- PDDFIN- PDDFIN- PDDFIN- PNDFIA- PNSPAQ- PNSPAQ- PNSPAQ- PNSPAQ- PNSPAQ- PNSPAQ- PNSPAQ- PNSPAQ- PNSPAQ- PDSPAQ- PDDPQD- PDDPQD- PDDPQD- SIDPIN- SSPIN- PDDPQD- SIDFIN- SSPIN- | VEAATLE | KDHK- LTNR- LTNR- LTNR- LTNR- KTDR- KTDR- KTDK- KTDK- KTDK- KTDK- SNQE- DADM- SNQE- DADM- SNQE- DADM- REDR- RENP- RDDR- RDDR- AVIV- ISDH- LRDR- LHRP- LHRP- LHRP- LHRP- QNNR- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NALHDETAR - JRYHRTVK - AQALETAR - QGAKAKAR - VEYKRVK - QEYKRVK - DEYRKVVE - NGEFKRVR - DEFKRVR - DEFKRVK - DEFKRVK - DEFKRVK - DEFKKKVK - DEFKKKVK - ESFNKTAA - ESFNKTAA - ESFNKTAA | ETV QWT EWT AWT AWT LQS LQA LQK ACY QUT LQX LQX LQX LQX LSI VET LWT LEYI EYI RSM | 1452 2022 1433 1733 1422 1711 1455 1445 1522 1766 1509 1499 1499 1455 1455 1455 1455 1455 145 |
| Human E2J2 84 NGRFKCMFRLGISTDFH-1DTMRAMS/STILUGLL-SFMVEKGTLGSIFISDFTKQLAVQSLAFNLKDK 154 C.cere E2J2 77 NGRFKCM | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C Human E2G1 S.cere E2G2 Human E2G2 S.cere E2G2 Human E2K C.rein E2K S.cere E2K Human E2H S.cere E2H Human E2M C.cere E2H | 766 1333 744 73 102 766 755 822 81 107 81 78 78 78 77 77 79 76 75 74 75 74 29 99 103 103 | MFHPN IYHCNN IYHCNN IYHCNN IYHCNN IYHPN IYHPN IYHPN IYHPN CFHPPN IYHPN IWHPN IWHPN IWHPN IWHPN IFHPN IFHPN IFHPN IFHPN IFHPN IFHPN | VYAN | GEIC GVIC GQIC GQIC GAIC GRVC GRVC GTIC GNIC GNIC GPVC GAIC GAIC GAIC GAIC GIC | LDILQ | | WII | IN STATUSE TISKU TISKU TISKU TISKU TISKU TISKU TISKU QIRTU QIRTU QIRTU TUVET TIVET SPEK SVEKU TIVET TIVET TIVET TILSJ CDLINN TILSJ CDLINN TINS SULLINN TINSS | LLTSIQ. LLTSIQ. ALLSIC. ALLSIC. ALLSIC. ALLSIC. ALLSIC. ALLSIC. LLSIC. LLSIC. LLSIQ. | SLLD SLLT SLLT SLLT SLLT SLLT SLLT SLLT ALLA DLD SLLG SLLG SLLG SLLG SMLA ALLA SMLA | = = PN | PNSPAN PADPUV PDDFUV PDDFUV PDDFUV PDDFUV PDDFUA PDDFIA PDDFIA PNDFIA PNDFIA PNDFIA PNSPAO SSPAN SSPAN DESGAN IESGAN IESGAN IESGAN EDDPOD PNDFIN PDDFIN EDDPIN PDDFIN PNDFIN | | KDHK- LTNR- KTDR- KTDR- KTDR- KTNK- KTNK- KTNK- STRK- SRNK- SRNK- DADM- REDR- REDR- REDR- REDR- REDR- REDR- LARDR- | - SQVVKRVK - ABHDRIAR - PKYEATAR - PKYEATAR - RKHDETAR - TKYHQTAK - ENYIRTVK - AQALETAR - QGAKAKAR - VEYKRVKK - BEYKKVR - BEYRKWL - BEYRKWL - BEYRKWL - BEYRKWL - EFYRKVKA - ESFNKTAR - ESFNKTAR - ESFNKTAR - ESFNKTAR - ESFNKTAR - ESFNKTAR - ESFNKTAR - KIYEBKIK - KAFEADVK - KAFEADVK | ETV QWT EWT AWT AWT LQS LQX ACY LQA ACY QUV LVX LQX LQX LQX LXX VET EYI EYI KSL LTM | 1452 2022 1433 1733 1422 1711 1455 145 145 1522 1766 1599 1499 1499 1495 1455 1435 1435 1435 1435 1437 1682 1722 1722 |
| C.cere E2J2 77 NGRFKPNTRLCLS/SDVH-1DTAMEGRSVSTILGL-SPMTSDEATTGSITTSDHQKKTLARNSISYNTFQN 147 Human E2J 81 NGRFECMKKUGLSISGHH-1TTAMESWS RTALLAII-GFMPTKGEGAIGSLDYTPEBRALAKKSQDFCCEGC 152 A.thal E2J 83 NGRFETM | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2C S.cere E2I A.thal E2C S.cere E2G Human E2G1 S.cere E2G2 Human E2G1 S.cere E2G2 Human E2K S.cere E2C Human E2K S.cere E2K Human E2H Human E2M T.ther E2M T.ther E2M T.vagi E2M | 76 133 74 104 73 76 76 76 76 82 81 107 75 81 107 77 77 77 77 77 77 75 74 72 99 90 103 82 | MFHPN IYHCNN IYHCNN IYHPN IYHPN IYHPN IYHPN IYHPN IYHPN IWHPN IWHPN VYHPN IFHPN VYHPN IFHP | YYAN | | LDILQ | | WILTS TY WS P - AI WS P - V WS P - V WS P - V WS P - AI WS P - | ZUVAS TISK TISK TISK TISK TISK TISK QIRT QURT TVKQ QIRT TVKQ QURT TVKQ VDVX VDVX TVKQ VDVX | ILTSIQ. ILT | SLLD SLLT SLLT SLLT SLLT ALLN ALLN ALLN DLD DLD SLLG | | PNSPAN- PADPUV- PDDPUV- PDDPUV- PDDPUV- PDDPUV- PDDPUA- PDDPIA- PADPAQ- PNDPAQ- PNSPAQ- PSSPAQ- ISSPIN- DSSPIN- DSSPIN- DESGAN- IESGAN- IESGAN- DESGAN- DESGAN- PDDPQD- PIDPQD- PIDPUN- SSDPIN- PDDPIN- PNDPIN- | | KDHK- LTNR- KTDR- KTDR- KTDK- KTDK- KTDK- KTNK- CQDP- SRNK- SNQE- DADM- REDR- REDR- REDR- REDR- REDR- RDDR- RDDR- RDDR- RDR- LRDR- LRDR- LRDR- LRDR- LRDR- LRDR- LRDR- LRDR- TKDN- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NALHDETAR - ARAKEIAQ - QAVARAR - VEYKRVK - QEYKRVK - ZEYRKWUE - BEYRKVU - VEYRKVK - ZEYRKVK - ZEYRKVK - ESFNKTAA - RLFEONVQ - KAFEADVK - IKFFNAH | ETV QWT EWT AWT AWT LQX AWT LQA ACY QUV LQA ACY QIV LVX LVX LVX LVX LVX LXX LXX LXX LXX LX | 1452 2022 1433 173 142 171 145 145 152 176 150 160 159 1496 1455 1455 1455 1455 1455 1455 1455 145 |
| A.thal E2J 83 NGRFETNTKIGLSISNYH-1EHAGESKSVRTALVALI-AFMPTSENGALGSVDYPKDERRTLAIKSRETPPKYG 154 Human Tsg101 98 KTGKHVDANFKIGLSISNYH-1EHAGESKSVRTALVALI-AFMPTSENGALGSVDYPKDERRTLAIKSRETPPKYG 154 Rueger 80 SCHQYGEGGELGLQYRPDNMHEOCKSADVVRSAK-ALEATPKDDEFSDWSSAHPTDLPSLSCGSRRF 147 Arthro 81 MPHWNPFSNEVCLLGYPPDNMHEOCKSADVVRSAK-ALEKAGMSGDEH-ADWNEKPQAEPFGAYYNSYANSA 152 E.coli 81 MPHWNPFSFLCYVEQMEADWDSN-DLEATYKEVDAQI-HQTLDSWSAATQCVNDKRELGEEFAAWRPSETLF 152 S.acid 99 NPHVQWQTYLGIVGSRN-2-2-MDASDGMFGISKLEJWLER-ALN-2DWGGAPLHPPKAYPTERIT | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2C Human E2G2 S.cere E2G2 Human E2G2 Human E2G2 K.cere E2K Human E2H S.cere E2K Human E2M T.ther E2M C.cere E2M G.cere E2M T.ther E2M C.cere E2M S.cere E2M Human E2M S.cere E2M | 766 133 74 104 73 76 76 75 81 107 78 81 107 75 81 77 77 77 77 77 77 70 76 75 74 72 99 9103 103 82 79 | MFH2N IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IYHEN IFHEN YYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN | VYAN | | LDILQ | | WILE TY WS P AI WS P VI WS P AI WS P AI | INVAS ITISKV TISKV TISKV QIRT QIRT QIRT QURT QURT TVKQ COVRT TVKQ COVRT TVKQ SVEK SVEK SVEK SVEK SVEK SVEK SVEK SVEK | IIITSIQ. IIITSIQ. IIITSIQ. IIITSIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIISIQ. IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII | SLLD SLLT SLLT SLLT SLLT SLLT SLLT ALLN ALLN ALLN SLLG SLG | | PNS PAN- PAD PAD PU- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PLA- PDD PLA- PND PLA- PND PLA- PND PLA- PND PAQ- PND PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- PD PD PUN- PND PUN- | | KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- RDR- RDR- RDR- RDR- RDR- RDR- RDR- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - QALETAR - JRYIRTVK - AQAIETAR - QALETAR - QALETAR - QEYKKWVEK - EYRKKVK - EYRKKVK - DEFYKKVK - DEFYKKVK - PEFERQVK - ESFNKTAA - ESFNKTAA - ESFNKTAA - ESFNKTAA - ESFNKTAA - RLFEQNVQ - KAFEADVK - KAFEADVK - KAFEADVK - KAFEADVK - KAFEADVK | ETV QWT EWT KYT KYT AWT LQV LQX LQX LQX LQX LQX LQX LQX LQX LXX EYI EYI EYI EYI EYI LYM LDYM QQFV | 1452 2022 1433 173 142 171 145 1454 153 152 1766 150 1600 159 1496 1455 1453 1455 1455 1455 1455 1455 1455 |
| Human Tsg101 98 KTGKHVDANGKTYLPYTH | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2C S.cere E2I A.thal E2C S.cere E2C Human E2G1 S.cere E2G2 Human E2G2 S.cere E2G2 Human E2K S.cere E2C Human E2H S.cere E2C Human E2H S.cere E2C Human E2M T.ther E2M T.ther E2M T.vagi E2M C.cere E2J | 766 133 74 104 73 766 766 768 811 78 811 78 78 78 777 79 9 766 755 744 722 99 90 1033 822 799 | MTEIDN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IYHEN | VYAN | | LDILQ | | WIL | IDVAS ITISK TTISK TTISK TTISK TTISK QIRT TTAK QIRT TTVKQ DVRT TTVKQ DVRT DVRT DVRT DVRT DVRT DVRT DVRT DVRT | ILTSIQ. ILTSIQ. JLSIC. JLSIC. JLSIC. JLSIC. JLSIC. ILSI | SLLD SLLT SLLT SLLT SLLT SLLT SLLT SLLT ALLN ALLN DLD DLD SLLG SLLT SFMT SFMT | = PN | PNS PAN- PAD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUA- PND PLA- PND PLA- PND PLA- PND PLA- PND PLA- PND PLA- PDD PUA- PND PLA- PDD POA- PDD POA- PDD POA- PND PUA- PND PLA- PND | | KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- CQDP- SNQE- SNQE- DADM- RDR- RENP- RDNR- AVIV- ISDH- LRDR- LRDR- LRDR- LRDR- LRDK- CGEGE- TKDN- KNRP- RQLA- KTLA- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NGYIRTVK - AQAIETAR - ARAKEIAQ - QGAKAKAR - VEYKKVK - AEYDKKVI - BYRKWVEK - BEYRKVL - NGEFKRVR - DEFEROVK - AEYSKKVK - DEFEROVK - ALSSKSWD - ETYKKTAK - EEYKKIA - EEYKKIA - EEYKKIA - EEYKKIA - FLFEONVQ - KAFEADVK - KEFAEAVK - VQSLAFNL - UNSLAFNL | ETV QWT EWT EWT KYT AWT LQS LQA LYK LQA LLYK LLYK LLYK LLYK LLYK LSI LYK LSI LSI KSL LTM KSL LTM KSL LTM KSL | 1452 2022 1433 1733 1422 1711 1455 1444 1533 1522 1760 1600 1599 1495 1455 1455 1455 1455 1455 1455 |
| Rueger 80 SCHQYGEGSELCLQYRPDXHHDCKSADVVRSAK-ALLEATPKDDGFSDVESAHPTDLPSLLSGSRRF 147 Arthro 81 MPHWNPFSNEVGLLGTFSERGTNGSLAGLKOGL-PAALKAGMSGEHADWBEKPQAEPFGAYYNSYANSA 152 E.coli 81 MPHWNPFSSELCYVEQMEDXDSN-DLEATYKEVDAQI-HQTLISVSAATQGVNDKRELEGEFQAYYNSYANSA 152 S.acid 89 NPHVQMQ | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N A.thal E2I S.cere E2I A.thal E2I S.cere E2C Human E2G S.cere E2G Human E2G S.cere E2G Human E2G S.cere E2C Human E2H S.cere E2K Human E2H S.cere E2K Human E2M T.ther E2M C.cere E2M T.vagi E2M CSUB_C1475 Human E2J2 C.cere E2J2 Human E2J2 Human E2J2 | 766 133 74 104 73 76 76 75 81 107 78 81 107 78 81 77 79 75 74 75 74 72 903 103 82 79 84 82 79 84 87 77 81 | MFHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN MFHEREN MFHEREN MFHEREN WHEREN VYHEN IYHEN IYHEN IYHEN IYHEN IGRFEK | VYAN | | LDILQ | | WII | IDVAS ITISKV ITISKV ITISKU ITISKQ ITISKQ ITISKQ QIRTT QIRTT QIRTT IVVQ ITIKQ ITIKQ ITIKQ ITIKQ SVEK IVVT ITVT ISPET ITVXQ ITIKQ SVEK IVVT ISSVEK IVVT ISSVET ISSUE ISITI | LITSIQ. TIT | SILD SLID SLIT SLIT SLIT SLIT SLIT SLIT ALLN ALLN ALLN SLID - | | PNS PAN- PAD PAU- PDD PLU- PDD PLU- PDD PLU- PDD PLU- PDD PLU- PDD PLA- PEDALN- PND PLA- PAD PAQ- PND PLA- PND PLA- PSS PAN- DESSAN- DESSAN- DESSAN- DESSAN- DESSAN- DESSAN- DESSAN- DESSAN- PDD PDD- PDD PLA- SDD PDD- PDD PLN- PND PLN- PND PLN- PND PLN- PND PLN- PND PLN- PTGSI- ATTGSI = ATTGSI = | | KDHK- LTNR- KTDR- LTNR- KTDK- KTDK- KTDK- KKDI- SRNK- SRNK- SRNK- SRNK- SRNK- SRNK- SRNK- SRNK- LKPR- LHRP- LHRP- LHRP- LHRP- LHRP- LKPK- KNRP- | - SQVYKRVK - ABHDRIAR - PKYEATAR - PKYEATAR - PKYEATAR - PKYEATAR - NCHORT - NCHORT - NCHORT - NCHORT - ARAKEIAQ - QGAKAKAR - VEYKRWVK - DEYKKVK - DEYKKVK - DEYKKVK - DEFYKKVK - PEFERQVK - ALSSKSWD - EYYKKTAK - EEYKKTAK - EEYKKTAK - EEYKKTAK - EEYKKTAK - EEYKKTAK - EEYKKTAK - EEYKKTAK - EEYKKTAK - KEFAEAVR - KEFAEAVR - VQSLAFNL - KKSJSVT | ETV QWT EWT AWT AWT LQS LQA QWT LQS LQA ARCY VET LQS LQA VET LVS LVX VET LWT LWT LWT LWT EYI LWT LWT EYI LWT FRSM CONTRACTOR CONTRAC | $\begin{array}{c} 1452\\ 2022\\ 143\\ 173\\ 142\\ 171\\ 145\\ 153\\ 152\\ 176\\ 160\\ 159\\ 145\\ 145\\ 145\\ 145\\ 145\\ 145\\ 145\\ 145$ |
| Arthro 81 MFHHWNPFSNEVCLUGTPSEEGGTNGSLAQLLKDQL-PAALKAGMSGDEH-ADWNEKPQAEPFGAYYNSYANSA 152 E.coli 81 MFHWABEEFGTVPEQHEADWDSN-DLEATYKEVDAQI-HOTLIDSVSAATQGVNDKRELGEFAAWRPSETLF 152 S.acid 89 NFHVQWQTYLGUYSRN-2-UNASDGMFGTISRLEJWLRF-AADN-2DWEGAPLHPPWAYTPERIT152 R.spha 90 FFHQAAVYSR-7-NGEPCLTDPTA7-SRPEPIALADRLINKVERFSRW-CELA-2-GRLHNPGDHFEJPLSGHTNPWTI | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2I Human E2G1 S.cere E2G2 Human E2G1 S.cere E2G2 Human E2K S.cere E2C2 Human E2K S.cere E2Z4 Human E2H T.ther E2M C.cere E2H Human E2M T.vagi E2M C.cere E2J4 Human E2J2 C.cere E2J2 Human E2J2 C.cere E2J2 A.thal E2J | 766 133 744 104 73 766 765 82 81 10775 81 788 788 788 788 788 788 788 799 766 755 744 799 766 755 744 81 1032 828 844 778 844 778 838 838 838 838 838 838 838 838 838 | MELDN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN CFHDN CFHDN CFHDN CFHDN CFHDN CFHDN IYHEN IWHEN IWHEN IFHEN IFHEN IFHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IFHEN IYHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IYHEN | VYAN | | LDILQ | | WILTS | IDVAS ITISKV ITISKV ITISKU ITISKU ITISKU ITISKU ITISKU QIRTT QURTT IVVET ITVET ISPET ISSVEK IDLTN ITVET ITVET ITVET ITVET ITVET ITVET ITVET ITVET ISSVEK INLYN INLYN INLYN INLYN ISSVEK ISSVET ISSVET ISSVET ISSVET | ILTSIQ. ILTSIQ. ILTSIQ. ILTSIQ. ILLSIC. ILLSIC. ILLSIQ. ILLSIQ. ILLSIQ. ILLSIQ. ILLSIQ. ILLSIQ. ILLSIQ. ILLSVI. ILL | SLLD SLFN SLTT SLLT SLLT SLLT ALNS ALNS ALNS ALNS SLLG - | = = P | PNSPAN- PADPIV- PDDFIV- PDDFIV- PDDFIV- PDDFIN- PDDFIA- PDDFIA- PDDFIA- PADPAQ- PNSPAQ- ISSPIN- NRSFIN- GDSPAN- DESGAN- IESGAN- IESGAN- DESGAN- DDSQD- PDDPQD- PIDPQD- PIDPQD- PIDPQD- PIDPQD- PIDPQD- PIDPQD- PIDPIN- PIDFIN- PEDFIN- PIDFIN- PIDFIN- PIDFIN- PITIGSI- ATIGSI- GALGSV- | | KDHK- LTNR- KTDR- LTNR- KTDR- KTDR- KTDK- KTDK- CQDP- SRNK- CQDP- SRNK- CQDP- SRNK- RDR- RDR- RDR- RDR- RDR- RDR- RDR- RD | - SQYVKRVK - ABHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - LENYIRTVK - AQAIETAR - QGAKAKAR - VEYKRVK - QGAKAKAR - VEYKRVK - EYRKNVEK - EYRKNVEK - DEFKKRVR - DEFKKRVR - DEFKKRVR - DEFKKRVR - ESFNKTAA - ESFNKTAA - ESYNKTAA - ESYNKTAA - ELYKKTAK - KLYEEKIK - KLYEEKIK - KLYEEKIK - KLYEEKIK - VQSLAFNL - VQSLAFNL - VQSLAFNL - KKSQDFC - KKSQDFC - KKSQDFC - KKSQDFC | ETV QWT EWT EWT AWT LQS LQA WT LQS LQA LQV LQV VET LVY VET LVY LWT LWT LWT LWT EYI LWT EYI LWT FQC KCK KSL KSL KSL KSL KSL KSL KSL KSL KSL KS | 1452 2022 143 173 1422 145 145 145 152 150 150 150 160 159 149 145 145 145 145 145 145 145 145 145 145 |
| S.acid 89 NPHVQWQTYLC1YQSRN-2-WDASDGMFGFISRLELWLRR-AAIN-2-DMEGAPLHPPVAYPTERIT152 R.spha 90 FPHQAVYSR-7-NGEPGITDPTA-7-SRPEPIALADRIIWKVERFSRM-CELA-2GRLHNPGDHFELPELSGHTNPMTI | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I S.cere E2N A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 S.cere E2G2 Human E2K C.rein E2K S.cere E2K Human E2H S.cere E2H Human E2M T.ther E2M T.ther E2M T.vagi E2M C.cere E2J2 Human E2J2 C.cere E2J2 Human E2J2 A.thal E2J Human TSg101 | 76674 13374 104 7307675 8218778 810778 810778 778777 779 766775 74472 755744 729 766775 74472 729 709103 822 709 824 77788 82884 777 813 8398 | MFEIN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IWHEN IYHEN | VYAN | | LDILQ | | WIL | IDVAS TTISKU TTISKU TTISKU TTISKU TTISKU TTISKU QIRTT QIRTT QIRTT VQIRTT TTVKQ QIRTT TTVKQ QIRTT TTVKQ QIRTT TTVKQ QIRTT TTVKQ QIRTT TTVKQ DITTU SVEK TTVKQ SVEK TTVKQ SVEK TTVKQ SVEK TTVKQ SVEK TTVKQ SVEK TTVKQ SVEK TTVKQ SVEK SVEK SVEK SVEK SVET SVEK SVET SVEK SVET | ILTSIQ. ILTSIQ. ILTSIQ. ILTSIQ. ILLSIC. ILLSIC. ILLSIC. ILLSIQ. ILLGIQ. ILLSIC. ILLGIQ. ILLSIC. ILL | SLLD SLLT SLT - | = = PN = | PNS PAN- PAD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUA- PDD PUA- PND PLA- PAD PAQ- ISS PUN- PND PLA- PND PLA- PND PLA- PDD PUA- PDD PUA- PDD PUA- PDD PUA- PDD PUA- PDD PUA- PND PUA- PUA- PUA- PUA- PUA- PUA- PUA- PUA- | | KDHK- LTNR- LTNR- KTDR- LTNR- KTDR- KTDR- SRNK- SRNK- SRNK- SRNK- SRNK- REDR- REDR- REDR- REDR- REDR- REDR- REDR- LRDR- LRDR- LRDR- LRDR- LRDR- LRDR- LRDR- REDR- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NYIRTVK - AQAIETAR - QAKAKAR - VEYKKVK - AEYDKKVI - AEYDKKVI - BYRKWVEK - EEYRKKVI - DEFEROVK - ALSSKSWD - EFYRKVKA - EEYKKVK - SEFNKTAA - EEYKKIK - KAFEADVK - KAFEADVK | ETV QWT EWT AWT AWT LQWT EWT LQS AWT LQS ARCV QWT LSI VET EYI EYI LYK VET EYI LYK KSL LTM QFV KCK CQN CFQN CKC CGMP | $\begin{array}{c} 1452\\ 2022\\ 1433\\ 1732\\ 1451\\ 1451\\ 1452\\ 1522\\ 1502\\ 1600\\ 1599\\ 1465\\ 1453\\ 1622\\ 1512\\ 1453\\ 1622\\ 1512\\ 1554\\ 1472\\ 1556\\$ |
| R.spha 90 PFPOAAVYSR-7-NGEPCITDPTA7-SRPEPIALADR.INKVERFSRW-CELA2-GRLHNPGDHFE.PPLSGHTNPMTI176 C.perf 99 LPHTLANGLN-1SVIGLHRGNI-2-WYIDHBVEDFVNRIFKPSD-AACN2-IKPGDDFEPMINYTETONIVYSYNKLTKFI 177 D.arom 1164 VPTPNVNVQSNIGFGAPVE-BCTTVEKIANNDA-FLSS-1-FIRS-1PTHRAGGAC-7-AYTURKMLDGFF-ORPERVLDVV 223 A.vari 146 PPIENVWEDSNIGFGGNSLS-MCSAATISOVMDLF-WKSP-10-SKTHEDNICNQLIKLHESKA-KSYPSSDLVVV 223 B.tht 149 APFFNVAGBNVGLGSSSPKKGQDMDPLEFQEVWEKR-FWNS-1-FFFLAGGNRNP-TRSNLVSVTEHARNNPPTYSELQS 224 Synech 84 TEARQOFNGKTWQCMSRHS-MVGIWTMLKRV-EHAL-4 | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C S.cere E2I Human E2G1 S.cere E2G2 Human E2G3 S.cere E2G2 Human E2K S.cere E2C2 Human E2K S.cere E2Z Human E2H T.ther E2M C.cere E2H Human E2M T.ther E2M C.cere E2J Human E2J C.cere E2J A.thal E2J Human E2J A.thal E2J Human E3D A.thal E2J Human E3D A.thal E2J Human E3D A.thal E2J Human E3D A.thal E2J Human E3D A.thal E2J Human E3D Human E3D A.thal E2J Human E3D Human E3D A.thal E2J Human E3D Human E3D A.thal E2J Human E3D Human E3D Hu | 76133 74144 7331022 76675582 822 823 8110778 7177777777777777777777777777777 | MEIDM IYHEA IYHEA IYHEA IYHEA IYHEA IYHEA IYHEA CHUN IYHEA CHUN CHUN CHUN CHUN CHUN CHUN CHUN IYHEA IMHEA IWHEA IY | VYAN | | LDILQ | | WII | DVAS TTISK TTISK TTISK TTISK TTISK TTISK QIRT TTISK QIRT TTVKQ TTVKQ TTVKQ QIRT TTVKQ TTVKQ TTVKQ TTVKQ TTVKQ QIRT TTVKQ TTVX TTVKQ TTVX TTVKQ TTVKQ TTVKQ TTVKQ TTVKQ TTVKQ TTVX TTVKQ TTVX TTVKQ TTVC | ILTSIQ. ILTSIQ. ILTSIQ. ILTSIQ. ILTSIQ. ILLSIC. ILLSIC. ILLSIQ. ILL | SLLD SLTT SLTT SLLT SLLT SLLT SLLT ALNS ALNS ALNS SLLG - | | PNSPAN- PAD PIV- PDD PIV- PDD PIV- PDD PIV- PDD PIV- PDD PIN- PDD PIN- PDD PIN- PIND P | | KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- SNQE- SNQE- DADM- SNQE- SNQE- SNRC SNQE- SNRC SNQE- SNRC SNQE- LARP- LHRP- LHRP- LHRP- LHRP- LHRP- LHRP- KTLA- QATG- PAEF- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NKHDETAR - TKYHQTAK - QQAKATAR - QQAKATAR - VEYKRVK - QQFKKKV - QEFKKRVR - DEFRKKV - QEFKKRVR - DEFRKKV - DEFRKKV - DEFRKKV - LESFNKTAA - ESFNKTAA - ESFNKTAA - ELYKKTAK - KLYEKTK - KLFEQNVQ - KAFEADVK - KEFAEAVR - KLFRVHAH - VQSLAFNL - KSQDFCC - IKSRETPP - PSNISYMP - GAYYNSYA | ETV QWT EWT AWT AWT LQX AWT LQX LLYK ARCY QUV LLYK LQA LLYK QUV LSI EXSL LTM QFV KCK KSL LTM FEQC KCGMP FNSA | $\begin{array}{c} 1452\\ 2022\\ 1433\\ 1771\\ 1452\\ 144\\ 1532\\ 176\\ 150\\ 160\\ 159\\ 149\\ 145\\ 168\\ 1722\\ 151\\ 168\\ 1722\\ 151\\ 155\\ 168\\ 1472\\ 152\\ 154\\ 168\\ 1472\\ 155\\ 168\\ 1472\\ 155\\ 168\\ 1472\\ 155\\ 168\\ 1472\\ 155\\ 168\\ 1472\\ 155\\ 168\\ 152\\ 152\\ 155\\ 156\\ 156\\ 156\\ 156\\ 156\\ 156\\ 156$ |
| C.perf 99 DFMTLAWGLN-1SVIGLHEGNI-2-WYIDHBYEDGWNRIKWF5D-AACN-2IKPGDDFEPMINYTETONIVYSYNKLTKFI 177 D.arom 154 MPYFNVNVQSNIGHRGNAPVE-EGTVEKIAAWNDA-FLRS-1FHENGEGK-7-AYTEWRDHLGRF-QFPERVLVDV 232 A.vari 146 PLENVWEDSSIGFG3NUS_SWCSAATISOVMDLF-WKSP-10-SKTHEDNIGNQLKLHESKA-KSYPSSDLVPV 223 B.thet 149 APFFNVAGANVGIGSSPKKGONDFLEFQSVWEKR-FMVS-1FSHLGGNRNP-TRSNLVSVTEHARNNPFDYSELQQS 224 Synech 84 TEARQGFNGRSWQRWSRHEREWRRGUDNUGTHINLKV-EHAL-4 127 B.cepa 202 SLIVINEFE | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I S.cere E2N A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 Human E2G2 Human E2G4 S.cere E2G2 Human E2G4 S.cere E2G2 Human E2G5 S.cere E2G2 Human E2G4 S.cere E2G2 Human E2G5 S.cere E2G2 Human E2G5 C.cere E2M C.cere E2M C.cere E2M C.cere E2J2 Human E2J A.thal E2J Human E2J A.thal E2J Human E2J Human E2J A.thal E2J Human E2J Human E2J Human E2J Human E2J S.cere E2J2 | 76 133 74 74 76 75 81 107 75 81 107 75 81 78 78 77 79 76 75 79 70 70 75 70 70 70 70 70 70 70 70 81 107 70 81 107 70 81 102 70 81 70 81 70 81 70 81 70 81 70 70 81 70 70 81 70 70 81 70 70 81 70 70 81 70 70 70 81 70 70 70 81 70 70 70 70 70 70 70 70 70 70 70 70 70 | MFEIN IYHEN | VYAN | | LDILQ | | WILE TY WS AI WS | DVASS TTISK TTISK TTISK TTISK TTISK TTISK QIRT TVKQ QIRT TVKQ UTLKQ QIRT TVKQ TTKQ TVKQ TTKQ TVKQ TTKQ SVEK TTKS SVEK TTISS SVEK TTISS SVEK TTISS SVEK TTISS TTISS SVEK TTISS TTISS SVEK TTISS TTISS SVEK TTISS SVEK TTISS TTISS SVEK TTISS TTISS SVEK TTISS TTISS SVEK TTISS TTIS TTI | ILTSIQ. TLSIQ. TLSIC. TLSIC. TLLSIC | SILD SLID SLIT SLIT SLIT SLIT SLIT SLID ALIN ALIN DLID SLIG SLIG SMLS - | = = PN = | PNS PAN- PAD PAD PU- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PUA- PAD PAO- PND PLA- PAD PAO- PND PLA- PAD PAO- PND PAO- PDD PO- PDD PO- PDD PO- PDD PO- PDD PD- PDD PD- PD- PDD PD- PDD PDD PDD PDD PD- PDD PDD PDD PDD PDD PDD PDD PDD PDD PDD | | KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- STRK- STRK- SRNK- | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NYIRTVK - QQAIETAR - QQAIETAR - QQAIETAR - QQAIETAR - QQAIETAR - QUEYKKVK - AEYYKKVK - EYYKKVK - DEYERQVK - NGEFKRVR - DEFERQVK - DEFERQVK - ALSSKSWD - EYYKKTAK - EYYKTAK - FLFEQNVQ - KAFEADVK - FINSTYP - YSLSGS - GAYYNSYA | ETV QWT EEWT AVT KYT QWT EWT LVX QWT LLYX VET LUX VET LUX VET LUX VET LUX VET LUX VET LUY VET LUY VET LUX VET LUX VET LUX SC SC SC SC SC SC SC SC SC SC SC SC SC | $\begin{array}{c} 1452\\ 2022\\ 1433\\ 1771\\ 1452\\ 144\\ 1532\\ 176\\ 150\\ 160\\ 159\\ 149\\ 145\\ 162\\ 145\\ 162\\ 145\\ 145\\ 162\\ 152\\ 152\\ 155\\ 168\\ 147\\ 152\\ 154\\ 168\\ 147\\ 152\\ 154\\ 168\\ 147\\ 152\\ 152\\ 152\\ 152\\ 152\\ 152\\ 152\\ 152$ |
| A.vari 146 PPLENWWEDSSIG FGGNSLS-MCSAATISOVWDLF-WKSP-10-SKTHEDNIGNQLIKLHESKA-KSYPSSDLVPV 223 B.thet 149 APPFNVAGRSWGLGSSSPKREQDMDPLEFQSYWEKR-FMMS-1FSHLGGNRNPTRSNLVSVTEHARINNPPDYSELQGS 244 Synech 84 TEARQOFNRSWGNSRHEREMARGUDGIWTMLKRV-EHAL-4 | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C Human E2G S.cere E2G Human E2G S.cere E2G Human E2K C.rein E2K S.cere E2K Human E2H S.cere E2H Human E2M C.cere E2H Human E2M C.cere E2H Human E2J C.cere E2J Luner E2J A.thal E2J A.thal E2J Human Tsg101 Rueger Arthro E.coli S.acid | 76 133 74 74 73 76 76 75 81 107 77 81 103 103 103 82 79 84 77 81 83 83 80 84 83 80 81 83 | MELDN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN CFHDD IYHEN CFHDD IYHEN IYHEN IYHEN IWHEN IYHEN IWHEN IYHEN IWHEN IYHEN IWHEN I IWHEN IWHEN I IWHEN I IWHEN I I I IWHEN I I I I I I I I I I I I I I I I I I I | VYAN | | LDILQ | | $ \begin{split} & \forall \mathbf{U} \\ \forall \mathbf{U} \\ \forall \mathbf{U} \\ U$ | IDVASS ITISKU TISKU TISKU TISKU TISKU TISKU QIRT TIVKQ QIRT TIVKQ QIRT TIVKQ QIRT TIVKQ QIRT TIVKQ QIRT TIVKQ QIRT TIVKQ VDVT VDVT TIVKQ VDVT VDVT VDVT VDVT VDVT VDVT VDVT VDV | ILTSIQ. ILTSIQ. //LSIC. //L | SLLD SLFN - SLT - SLT - SLT - SLT - SLT - SLT - SLT - ALN - ALN - ALN - SLT - ALN - SLT - SL | | PNSPAN- PADPUV- PDDFUV- PDDFUV- PDDFUV- PDDFIA- PDDFIA- PADPIA- PNDFIA- PADFAQ- SISPAQ- SISPAQ- SISPAQ- SISPAQ- SISPAQ- SISPAQ- SISPAQ- SISPAQ- PADFQD- PDSPQD- PDDFQD- PDDFQD- PDDFQD- PDDFQD- PDDFQD- PDDFQD- PDDFQD- PDDFIN- PDDFIN- PDDFIN- PDDFIN- PDDFIN- PDDFIN- SISPAQ | | KDHK- LTNR- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- SNOE- SNOE- SNOE- SNOE- SNOE- SNOE- RENP- RENP- RENP- RENR- RENR- RENR- RENR- RENR- RENR- RENR- LHRP- LHRP- LHRP- LHRP- LHRP- RONR- RENR- | - SQVYKRVK - ABHDRIAR - PKYEATAR - PKYEATAR - RKHDETAR - TKYHQTAK - ENYIRTVK - AQALETAR - QGAKAKAR - VEYKRVK - BEYRKWEK - BEYRKWEK - BEYRKWEK - EEYRKWEK - EEYRKWEK - EEYRKWEK - EEYRKWEK - EEYRKWEK - EEYRKWEK - EEYRKVEK - KAFEADVK - VOSLAPNI - KKSETPP - PFNTSVM - PSLLSCS - GAYYNSYA - AAVWRSE | ETV QWT EWT AWT KAWT LUX LUX LUX LUX LUX LUX LUX LUX LUX LUX | $\begin{array}{c} 1452\\ 2022\\ 1433\\ 1732\\ 1711\\ 1455\\ 152\\ 1560\\ 1599\\ 1465\\ 1453\\ 1609\\ 1455\\ 1453\\ 1609\\ 1455\\ 1453\\ 1609\\ 1455\\ 1452\\ 1552\\ $ |
| B.thet 149_APFBWAGANVGLGSSSPKKEQCMDGLEFOGYWEKR-FWMS1FSHLGGNRNPTRSNLVSVTEHARNNPFDYSELQQS 224 Synech 84_TEARCOGFNGRSWQRWSRHEREMRRGVDGIWTMIKRV-EHAL4 | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I A.thal E2I A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 K.cere E2G2 Human E2K C.rein E2K S.cere E2K Human E2H S.cere E2K Human E2H S.cere E2K C.cere E2K Human E2H C.cere E2M C.cere E2M C.cere E2M C.cere E2J Human E2J A.thal E2J Human E2J A.thal E2J Human E2J A.thal E2J Human E2J A.thal E2J Human E2J A.thal E2J Human E3D C.cere E2J Human E2J A.thal E2J Human E2J A.thal E2J Human E3D A.thal E2J Human E3D A.thal E3D Kacid R.spha C.perf | 76373747878787878787877777797757747779777577477797977577477279991033103277998487777777577448138880888808880888088808880888088808880 | MFHINN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN MFHEN IYHEN IFHEN IWHEN IYHEN IFHEN IYHEN IFHEN IFHEN IGRFK IWHEN IFHEN MGRFK IWHEN IFHEN MGRFK IWHEN IFHEN MGRFK IFTGHH MFF INGRFK IFTGHH MFF INGRFK IFTGHH MFF IVHEN IFFF INGRFK IFTGH IWHEN IFFF INGRFK IFTGH IWHEN IFFF INGRFK IFTGH IWHEN IFFF INGRFK IFTGH IWHEN IFFF INGRFK IFTGH I HEFF I I I HEFF I I I I I I I I I I I I I I I I I I | VYAN | | LD LLQ | | WII P - T WII P - AI WII P - VII WII P - VII WII P - VII WII P - VII VII WII P - VII VIII VIII WIII VIII - VIII VIII VIII VIIII VIIII VIIII VIIIII VIIIII VIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII | IDVASS ITISKU TISKU TISKU TISKU TISKU TISKU QIRT TIAK QIRT TVKQ QIRT TVXQI QIRT TVKQ QIRT TVXQI TVXQ | LITSIQ. LITSIQ | SLLD SLLT SLT | = = PN = | PNS PAN- PAD PAD PAD PAD PU- PDD PUV- PDD PUV- PDD PUV- PDD PUV- PDD PLA- PAD PAD PAD PAD PAD- PND PAD- PND PAD- PND PAD- PDS PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- DES PAN- PDD PD- PDD PD- PDD PD- PND PDD- PDD PD- PND PDD- PDD PD- PND PDD- PDD PLN- PND | | KDHK- LTNR- KTDR- LTNR- KTDR- KTDK- KTDK- KTDK- KTDK- SNOE- SNOE- SNOE- SNOE- SNOE- RDNR- | - SQVYKRYK - ABHDRIAR - PKYEATAR - PKYEATAR - PKYEATAR - PKYEATAR - NCHDETAR - NCHDETAR - QAXALETAQ - QGAKAKAR - VEYKRWYK - DEYYKRWYK - DEYYKRWYK - DEYYKRWYK - DEYYKWYK - DEFYKRYR - DEFYKRYR - KEFACA - KEFACAR - KEFACAR - KEFACAR - KEFACAR - VQSLAFNL - KKSRETP - VQSLAFNL - KKSRETP - PSLLSGS - GAYYNSYA | ETV QWT EWT AWT LQS LQA KYT LQS LQA KAVT LQS LQA KAVT LQS LQA KAVT LQS LQA KAVT LQS LQA KAVT LQS LQA KAVT LQS LQA KAVT LQS KAVT LQS KAVT LQS CANT CANT CANT CANT CANT CANT CANT CANT | $\begin{array}{c} 1452\\ 2022\\ 1433\\ 1421\\ 1731\\ 1455\\ 1441\\ 1532\\ 1502\\ 1600\\ 1599\\ 1465\\ 1600\\ 1599\\ 1445\\ 1600\\ 1599\\ 1445\\ 1600\\ 1599\\ 1455\\ 1522\\ 1554\\ 1600\\ 1554\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1552\\ 1554\\ 1554\\ 1552\\ 1554\\ 1554\\ 1552\\ 1554\\ 1554\\ 1552\\ 1554$ 1554 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 15555 1555 1555 1555 1555 1555 1555 1555 1555 1555 1555 |
| Synech 84 TEARQOFNCRSWQRWSRHEREMRRGVDGIWTMLKRV-EHAL4 | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I S.cere E2I A.thal E2C Human E2G1 S.cere E2C Human E2G1 S.cere E2C Human E2G2 S.cere E2C Human E2M C.crei E2C Human E2J A.thal E2J Human E2J A.thal E2J Human Tsg101 Rueger Arthro E.coli S.acid R.spha C.perf D.arom | 76 133 73 76 76 76 81 78 78 78 78 77 77 77 77 79 75 74 75 74 75 74 75 74 81 83 82 79 84 103 82 79 84 84 83 80 84 81 83 89 90 91 54 | MEEDM ITHEA | VYAN | | LDILQ | | $ \begin{split} & \forall I \ I \ P &= - T \ W \ I \ P &= - T \ W \ S \ P &= - A \ I \ W \ S \ S \ S \ S \ S \ S \ S \ S \ S$ | IDVASS ITISKU TISKU TISKU TISKU TISKU TISKU QIRT TIVKQ UDVRT TUKQ UDVRT TUKQ UDVRT TUKQ UDVRT TUKQ UDVRT TUKQ UDVRT TUKQ SVEK SVEK SVEK SVEK SVEK ISVST SVEK ISVST SVET SVET ISVST SVET SVET SVET ISVST SVET SVET ISVST SVET SVET SVET SVET SVET SVET SVET | LITSIQ. LITSIQ | SLLD SLFN SLFN SLT SLT SLT SLT ALMS ALMS ALMS SLLG SL | | PNSPAN- PAD PAD PAD PAD PAD PAD PAD PAD PAD PAD | | KDHK- LTNR- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- STDR- SNRE- SNRE- SNRE- SNRE- SNRE- SNRE- SNRE- SNRE- SNRE- SNRE- SNRE- RENR- | - SQYVKRVK - ABHDRIAR - PKYEATAR - PKYEATAR - RKHDETAR - TKYHQTAK - ENYIRTVK - AQALETAR - QGAKAKAR - VEYKRVK - DEYKKVK - DEYKKVK - DEYKKVK - DEYKKVK - DEFKKRV - DEFKKRV - DEFKKRV - DEFKKRV - ESFNKTAA - ESFNKTAA - ESYNKTAA - KLYEEKIK - KLYEEKIK - KLYEEKIK - KLYEKIK - KLYEKIK - KLYEKIK - KLYEKIK - KLYEKIK - KLYEKIK - KLYEKIK - KLYEKIK - FLFEQNVQ - KAFEADVK - KLYRVHAH - DLFYSRVR - KKSQDFCC - IKSRTPP - PPNTSVM - PSLLSGS - GAYYNSYA - AAYWRSE - VYSINKIT | ETV QWT EWT EWT LQA XAWT LQS LQA XARCY QUV LVX LQA LQA QUV LSI RSM LQA XARCY QUV LSI EYI LWT LWT LWT LWT LWT LWT TLF RSSL LTM KDDY KOF KYD KYD KYD KYD KYD KYD KYD KYD KYD KYD | $\begin{array}{c} 1452\\ 2022\\ 1433\\ 1422\\ 1733\\ 1422\\ 1711\\ 1455\\ 1522\\ 1502\\ 1600\\ 1599\\ 1496\\ 1455\\ 1602\\ 1455\\ 1496\\ 1455\\ 1150\\ 1599\\ 1496\\ 1152\\ 1502\\ 1512\\ 1502\\$ |
| B.cepa 202 <u>SLVTNEFE</u> EKT <u>W</u> QGWSRH3NSPMRQGIDNVGTHLMLV-DDFL6 | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2I A.thal E2C S.cere E2C Human E2G S.cere E2G Human E2G S.cere E2G Human E2K S.cere E2G Human E2K S.cere E2C Human E2K S.cere E2C Human E2M T.ther E2M C.cere E2H Human E2J C.cere E2J2 Human E2J2 C.cere E2J2 Human E2J2 C.cere E2J2 Human E2J3 C.cere E2J2 Human E2J3 C.cere E2J2 Human E2J3 C.cere E2J2 Human E2J3 C.cere E2J2 Human Tsg101 Rueger Arthro E.coli S.acid R.spha C.perf D.arom A.vari | 76 133 73 1024 76 82 81 107 75 75 77 77 76 75 72 99 90 31 03 82 79 90 84 77 78 82 79 90 84 82 79 90 31 03 82 82 79 90 31 03 82 82 79 90 31 02 84 82 82 76 82 82 82 83 82 82 83 82 82 83 82 82 83 84 82 83 83 82 83 84 83 82 83 84 83 83 84 83 83 84 83 83 84 83 84 83 84 83 84 83 84 83 84 83 84 84 84 84 84 84 84 84 84 84 84 84 84 | MEEDN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IWHEN IWHEN IWHEN IWHEN IYHEN IGREFK IYHEN IGREFK IYHEN IGREFK IYHEN IGREFK IYHEN IGREFK IYHEN IGREFK IYHEN IYHEN IGREFK IYHEN IYHEN IYHEN IYHEN IGREFK IYHEN IYHE | VYAN | | LDILQ | | WII P - TX WIS P - AI WIS P - CI WIS P - CI </td <td>IDVASS ITISK ITISK ITISK ITISK ITISK QIRT QIRT QIRT QURT ITISK ITISK ITVET ISVEK IDITN SVEK ITVET ISVEK IDITN SVEK ITVET ISVE SVEK INVET INVET INVET INVET INVET INVET INVET INVET INVET INVET INVE INVE INVE INVE INVE INVE INVE INVE</td> <td>LITSIQ. LITSIQ.</td> <td>SLLD SLTN SLTN SLT SLT SLT SLT SLT SLT SLT SLT SLT SLT ALLN DLD SLT</td> <td></td> <td>PNSPAN- PADPIV- PDDPIV- PDDPIV- PDDPIV- PDDPIV- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PNDPIN- PNSPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PTIGSI- ATTGSI- ATTGSI- GALGSV- VFSRPI KDDGIS- WSOBH- VSATQG</td> <td></td> <td>KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- SNQE- COPP- SNRC SNQE- COP- SNRC SNQE- COP- COP- COP- COP- COP- COP- COP- COP</td> <td>- SQYVKRVK - AEHDRIAR - PKYEATAR - TKYHQTAK - TKYHQTAK - SAYNGTAK - OQAKARAR - VEYKRVK - QQAKARAR - VEYKRVK - QEFKKV - QEFKKV - DEFEROVK - DEFEROVK - ESFNKTAA - ESFNKTAA - ESFNKTAA - ESFNKTAA - EKYQKIK - KLYEKIK - KLYEKIK - KLYEKIK - KEYAEAVR - KEYAEAVR - IKFRHAH - USISYNT - KSSUPT - JKSETPP - PPNTSYMP - PSILSGS - GAYYNSYA - AKYPSSDL - NSYNKIT - NSYNKIT</td> <td>ETV QWT EWT AWT LQS LQA ACY QWT LLYK ACY QUV LSI EVI LVK LLYK LLYK LSI EVI LWT LSI EVI LWT LSI EVI LWT LYK LYK KFQC KGMP KFQC KGMP KFQC KFQC VVVV VVV</td> <td>1452 1433 1422 1733 1422 1731 1455 1443 1552 1456 1599 1496 1455 1455 1455 1455 1455 1455 1455 145</td> | IDVASS ITISK ITISK ITISK ITISK ITISK QIRT QIRT QIRT QURT ITISK ITISK ITVET ISVEK IDITN SVEK ITVET ISVEK IDITN SVEK ITVET ISVE SVEK INVET INVET INVET INVET INVET INVET INVET INVET INVET INVET INVE INVE INVE INVE INVE INVE INVE INVE | LITSIQ. | SLLD SLTN SLTN SLT SLT SLT SLT SLT SLT SLT SLT SLT SLT ALLN DLD SLT | | PNSPAN- PADPIV- PDDPIV- PDDPIV- PDDPIV- PDDPIV- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PNDPIN- PNSPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PTIGSI- ATTGSI- ATTGSI- GALGSV- VFSRPI KDDGIS- WSOBH- VSATQG | | KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- SNQE- COPP- SNRC SNQE- COP- SNRC SNQE- COP- COP- COP- COP- COP- COP- COP- COP | - SQYVKRVK - AEHDRIAR - PKYEATAR - TKYHQTAK - TKYHQTAK - SAYNGTAK - OQAKARAR - VEYKRVK - QQAKARAR - VEYKRVK - QEFKKV - QEFKKV - DEFEROVK - DEFEROVK - ESFNKTAA - ESFNKTAA - ESFNKTAA - ESFNKTAA - EKYQKIK - KLYEKIK - KLYEKIK - KLYEKIK - KEYAEAVR - KEYAEAVR - IKFRHAH - USISYNT - KSSUPT - JKSETPP - PPNTSYMP - PSILSGS - GAYYNSYA - AKYPSSDL - NSYNKIT - NSYNKIT | ETV QWT EWT AWT LQS LQA ACY QWT LLYK ACY QUV LSI EVI LVK LLYK LLYK LSI EVI LWT LSI EVI LWT LSI EVI LWT LYK LYK KFQC KGMP KFQC KGMP KFQC KFQC VVVV VVV | 1452 1433 1422 1733 1422 1731 1455 1443 1552 1456 1599 1496 1455 1455 1455 1455 1455 1455 1455 145 |
| KNIZOD 104 MVCLV@GATR-1-SLK@ITEMIGSDGWAAK-LSRTDPVGSGSSLL-PYGAGAASCFGAANVFRTIFAA 167 | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N S.cere E2N A.thal E2I A.thal E2I A.thal E2C S.cere E2C Human E2G1 S.pomb E2G1 Human E2G2 Human E2G2 Human E2G2 Human E2G4 S.cere E2G2 Human E2H S.cere E2K Human E2H S.cere E2K Human E2H S.cere E2K Human E2H S.cere E2K Human E2J C.cere E2X Human E2J Human E2J Human E2J Human E2J Human E2J Human E2J Human E2J Human E2J Human E2J Human E2J A.thal E2J Human E3 S.ceri E.coli S.acid R.spha C.perf D.arom A.vari B.thet | 763 744 1044 733 766 755 822 811 788 777 714 727 799 90 103 103 20 799 84 813 838 800 811 899 99154 1499 | MFEIN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN MFHEN IYHEN IFHEN IWHEN IYHEN IFHEN IYHEN IFHEN IYHEN IFHEN IGREK IWHEN IFHEN MGREK IWHEN IFHEN MGREK IWHEN IFHEN MGREK IWHEN IFHEN MGREK IWHEN IFHEN MGREK IWHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IWHEN IFHEN IFHEN IWHEN IFHEN IWHEN IFHEN | VYAN | | LDILQ | | WII | IDVASS ITISKU TISKU TISKU TISKU TISKU TISKU QIRT TIKQ QIRT TVKQ UDVT TVKQ UDVT TVKQ VDV TVKQ VDV VDV VDV VDV VDV VDV VDV VDV VDV VD | LITSIQ. LITS | SILD SLEN SLEN SLLT SLLT SLLT ALNS ALNS ALNS ALNS SLLG - | = = PN = DCN = DAN = DAN = DAN = PN = | PNSPAN- PADPIV- PDDPIV- PDDPIV- PDDPIV- PDDPIV- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PNDPIN- PNSPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PDDPIN- PTIGSI- ATTGSI- ATTGSI- GALGSV- VFSRPI KDDGIS- WSOBH- VSATQG | | KDHK- LTNR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- KTDR- SNQE- COPP- SNRC SNQE- COP- SNRC SNQE- COP- COP- COP- COP- COP- COP- COP- COP | - SQYVKRVK - AEHDRIAR - PKYEATAR - TKYHQTAK - TKYHQTAK - SAYNGTAK - OQAKARAR - VEYKRVK - QQAKARAR - VEYKRVK - QEFKKV - QEFKKV - DEFEROVK - DEFEROVK - ESFNKTAA - ESFNKTAA - ESFNKTAA - ESFNKTAA - EKYQKIK - KLYEKIK - KLYEKIK - KLYEKIK - KEYAEAVR - KEYAEAVR - IKFRHAH - USISYNT - KSSUPT - JKSETPP - PPNTSYMP - PSILSGS - GAYYNSYA - AKYPSSDL - NSYNKIT - NSYNKIT | ETVY QWT EWT AWT LQS LQA LQA LQA LQA LQA VET EYI LQA VET EYI LYX VET EYI LYX KCC CV VYWT EYI LYX KCC CV VYWT EYI LYX VVT VYWT FEGC GRF NISE FEGC CV VVV VQQ V VVV VVV VVV VVV VVV VVV V | 1452023 1733 142217 1453217 1453217 1500 1599 1465 1453 1453 1500 1599 1465 1453 16822 1751 1504 1453 15221 1548 1672 152223 1554 167222224 |
| | S.cere E2A Human E2D S.cere E2D C.mero E2D G.lamb E2D Human E2N C.mero E2N S.cere E2N A.thal E2C S.cere E2C Human E2G S.cere E2G Human E2G S.cere E2G Human E2G S.cere E2G Human E2K S.cere E2G Human E2H S.cere E2C Human E2H S.cere E2C Human E2H T.ther E2M T.ther E2M C.cere E2J Human E2J A.thal E2J Human E2J A.thal E2J Human Tsg101 Rueger Arthro E.coli S.acid R.spha C.perf D.arom A.vari B.thet Synech B.cepa | 763 744 1044 732 766 755 821 818 787 777 799 765 744 722 999 103 1033 1033 1033 1033 1033 1033 1 | MFHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IYHEN IWHEN | VYAN | | LDILQ | | WII | IDVASS ITISK ITISK ITISK ITISK ITISK ITISK QIRT QIRT QIRT IVIS ITISK ITISK ITISK ITISK ITISK ISVEK ISVEK ISVEK ISVEK ISVEK ISVEK ISVEK ISVEK ISVE ISVEK ISVE ISVEK ISVE ISVEK ISVE ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVE ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVEK ISVET ISVE ISVET ISVE ISVET ISVE ISVET ISVE ISVET ISVE ISVET ISVE ISVET ISVET ISVE ISVET ISVE ISVET ISVET ISVET ISVET ISVE ISVET | LITSIQ. LIT | SLLD SLTN SLTN SLTT SLTT SLTT SLTT SLTD SLTD SLTD ALLN DLD SLTG SLTG SLTG SLTG SLTG SLTG STLT SLTG - | | PNS PAN- PAD PAD PUV- PDD PLV- PDD PLV- PDD PLV- PDD PLV- PDD PLN- PDD PLN- PDD PLN- PDD PLN- PND PLA- PND PLA- PND PLN- PNS PAQ- PS PAQ- PS PAQ- PS PAQ- PS PAQ- PS PAQ- PS PAQ- PS PAQ- PS PLS PAQ- PS PLS PLS PLS PLS PLS PLS PLS PLS PLS P | | KDHK- LTNR- LTNR- KTDR- KTDR- KKTDR- KKTDK- KKTDK- KKTA- SNQE- COPP- SNQE- SNQE- COP- SNQE- SNQE- LTNR- RENR- RENR- RENR- RENR- RENR- RENR- RENR- RUNR- LHRP- LHRP- LHRP- LHRP- LHRP- LHRP- LHRP- CTSO | - SQYVKRVK - AEHDRIAR - PKYEATAR - PKYEATAR - TKYHQTAK - NKHDETTR - ARAKEIAQ - QAVARAR - VEYYKRVK - QQAIETTR - QAKAKAR - VEYKKVK - BEYKKVK - BEYKKVK - BEYKKVK - ESFNKVA - ESFNKVA - ESFNKVA - ESFNKVA - ESFNKVA - ESFNKVA - FLFEQNVQ - KAFEADVK - KAFEADVK - KAFEADVK - KAFEADVK - KAFEADVR - IKFYHAH - ULFYSRVR - KSISYNT - KSSQDFCC - IKSRETPP - PPNTSYMP - PSLSGS - GAYYNSYA - AXYWRSE - NYSYNKT - KSYSSDL - NYSYNKT | ETVY QWT EWT AWT LQS QWT LQS LLYK AWT LLQS RLV VET EVI LLQA VET EVI EYI LQS RLV VET EYI LQS RLV VET EYI LQS RLV VET EYI EYI LQS KSL VV VET EVI KSL VV VV VV VV VV VV VV VV VV VV VV VV VV | 145 2022 173 173 142 174 5 145 145 145 152 176 160 160 160 160 160 160 160 160 160 16 |

Figure 2. Continued.

| B | | Active core |
|--|---|--|
| | | |
| | CKVDGETMIIMDSFALPVEGTETRVNAQAAAYEYMAAYIENA CEFVDDYTVRVIDVFAMPQSGTGVSVEAVDPVFQAKMLDML | |
| A.thal CSN5A 62 SALALLKMVVHARSGGTIDIMCLM | CETKODTIVKVIDVFAMPQSGIGVSVEAVDFVFQAMHDMD- | KLAGRLENVVGWYHSHPGYGCWLSGTDVSTOMLNO 163 |
| S.cere RPN11 30 SSIALLKMLKHGRAGVEMEVMELM | EEFVDDYTVNVVDVFAMPOSGTGVSVEAVDDVFOAKMMDML | KOTGRDOMVVEWYHSHPEFGCWLSSVDVNTOKSFE 130 |
| T.bruc RPN11 30 SSLALLKMLMHGRAGVPLEVMCLM | CKTEGDTIIVMDAFALPVEGTETRVNAQSDAYEYMVEYSQTS CEFVDDYTVNVVDVFAMPQSGTGVSVEAVDDVFQAKMMDML- ICELIDDYTVRVSDVFSMPQTATGQSVEAVDPEYQVHMLDKL- | SVVGRPEKVVGWYHSHPGFGCWLSGEDVMTASSYE 130 |
| T.bruc CSN5 62SDGQAILSSPQTTTDTQRRENWFEVMCLL | GHFRENELIVTSTFALPVDASEVECSMNEASQMYMLEYLQYHQRTGF- | 15-EEIEEAECCVGWYHSHPGYTCFLSGTDVATQRVGQ 188 |
| G.lamb RPN11 41 SHVALIKMLRHCKOGIEI | GTFVDKYTVYVSDCFSMPOVGOADSVDSVDEVFOAEMMEML | KKVNVPENCVGWYHSHPGYFAWLSHIDONTHKSFE 141 |
| S.pomb AMSHP 268 LLKKVFLDVVKPNTKKNLPTCCIL | CEKLR-1NAFFITHLVIPLQEATSDTCGTTDEASLFE | FQDKHNLLTIGWIHTHPTQTCFMSSVDLHTHCSYQ 362 |
| CSTB C1473 6 VDLALAKINKHAASSIOPBVACIL | REKSACKVIETWDAVTCFOVCTPAVVOLDEMVMAKVAFFIS | KSDKNILVILGWYHSHDGLDVFISPTTTDTOKRYO 104 |
| A.fulg JAB 4 SRGLLKTILEAAKSAHEDDFIALL | SSKGSVSAVIHL | DMLPIGMKVFCTVHSHPSPSCRPSEEDLSLFTRFG 88 |
| P.aeru JAB 7 TEHALSVIYRHACRTYERECCEFV | ADAKVKEGTNIQDELHMA-DPRRYPRTAANGYTFSVTDTVFLN- | SSFKTCSPVSVIVISHPDVGAYFSREDIDKALYAG 108 |
| Py.aer JAB 1 MPKAFLEEARKKCAPEAECVALI P.hori JAB 36 LPKNIIEEIITRSRESKIEICEFI | GISD-1ALSWRWMKNVAASPVFFKLDPEEVYK- | -3-EAEERGEELLAIFFITHECPP-TPSWEDVRHMRL 90 |
| E.coli RadC 39 STRAAREWLILINMAGLERBEFRVL | GTKNGERFIGKEVE-FIRNRLNSSVEFEMDPEEMINALE YLNNONOLIAGETXFTGTINRTEVHPREVIK | PALVENDANATATI AMATHE CONTROL AND THE PALA |
| B.subt RadC 112 SPEDGANLVMEDMRFLTCPHFVCL | YLNTKNQVIHKRTVFIGSLNSSIVHPREVFK | EAFKRSAASFTCVHNHPSGDPTPSREDTEVTRRI- 200 |
| | YLDTKNQILKEEVVSIGSLNASIVHPREVFK | |
| | | |
| D.radi 6 PAPLRRALWAQVRREIPRECVCAL | LDTKLNVIGENTLTVGTSDRSLIHPRDVFR GwVRGEQVQAHALYPLPNVAADPEREYLADPGDLLRVVR | AMQREGLDLVALYHSHPHGPAAPSASDRRLAA 101 |
| A.aeol 5 KKEVLEKMIKQAERDYEYETCELL | ISKSEGGIRIAYEAFET-PNANPDRKHDRYEIAPKDYMRAED | YAISKGMEIVGVYHSHEDHPDRPSOFDLORAFP 102 |
| P.puti 6 TAQALEQVRHLAQAAHPIBACCLI | AASGEPLAHRVVPMRNQAASPTWFSFDPREQLQVWR CTVT-2-GDNRVAALHRATNRRSEQRTRRYELTADDYRAADA | ELDQRDEDCRVI <mark>YH</mark> SHTASEAWPS <mark>RED</mark> IALASDP- 100 |
| S.rubb 4 TPDILDQIRVHGADAYPEEGCCFL | GTVT-2-GDNRVAALHRATNRRSEQRTRRYELTADDYRAADA | AAQEQGLDVVCVYHSHPDHPARPSATDLEEATFP- 103 |
| M.tube 14 RADLVNAMVAHARRDHPDEACGVL | A <mark>G</mark> PEGSDRPERHIPMTNAERSPTFYRLDSGEQLKVWR | AMEDADEVPVVIYHSHTATEAYPSRTDVKLATEP- 108 |
| N.farc 5 KSDLVAAMVAHARADHPDBACCVI. | AGPEGSDRPERFIAMINAERSPTFYRFDSGEQLKVWR | EMDAADEEPVVIYHSHTATEAYPSRTDISYASEP- 99 |
| | AGDPAEQFSAFKDPAEQFSAFK | |
| G.meta 6 -RAIHAELIAHAQADA <mark>F</mark> IEAC <mark>C</mark> IL | GIDGAVSAIFRMANTDQSDEHFMMDPKEQFAVVK | ELRNRGLAMLAI <mark>YH</mark> S <mark>H</mark> PETPAR <mark>PS</mark> EEDIRLALTP-97 |
| | GAFRTYPKLNKIED | |
| | IDAFRLINANMMVL-GHE <mark>P</mark> RQTTSNL <mark>G</mark> HLNKPSI(| |
| A.thal SCN5A 164 QYQEPFLAVVIDPTRTVSAGKVE | IGAFRTYPELNKIED | GVHCKQYYSLDITYFKS- 235 |
| | IDAFRLIDTLLNKANI | |
| T.bruc RPNII 131 QLTPRSVSVVIDPIQSV1-GKVV | IDAFRTTKDPHTCPRIM-FQEERQTTSNICWLTRPSP: /RAFRTFPECAVGDGTESTSADSTGICAAPRQCGFHD! | LALTRGLDRDYYSLPITFRKK- 209 |
| C lamb RDN11 142 REDVRCTATVID | IEAFRIFPEASMGLSFCISFGSSTDTRVITSDKCFMRPKNP | PELINEIGANGHCIIELPIILVKS- 271 |
| S pomb AMSHP 363 LM-LPEATATVMAPSKNTSGTFR | LLDEGLQTIVKCRKI | PGLEHP 405 |
| CSUB C1473 105 AMFSKAVALVVDEVDYAKTRRIS | SLKFKVFQISKE | GRVVSLPVSIG 150 |
| | YNRKGEEVELEVVEKD 121 | |
| P.aeru JAB 109 EPMLPVDYLVVDVAAGNVRG-4-AWRN | | |
| Py.aer JAB 91WPVTWIIANVFDWHI-1-AWRI | DGGLKTIPLEFI 122 | |
| P.hori JAB 134IPWLIVSLKGDMKAF | | |
| E.coli RadC 128VQALGLVDIRVPDHLI | | |
| B.subt RadC 201FECGNLIGIELLDHLV | | |
| M.acet RadC 198VEGGKLLGIDILDHII | | |
| T.mari RadC 192KQAGEIIGVSLVDHVT | | |
| D.radi 102YPVPYLIAD <mark>P</mark> AAEVLRA A.aeol 103DLSYIIFSVOKGKVASYRSWEL | | |
| A.aeol 103DLSYIIFSVQKGKVASYRSWEL P.puti 101QVHYLIVSTWGEARHAARSFRI | | |
| S.rubb 104GFTYVIVSVRDGAPEALTAWAL | | |
| M.tube 109DAHYVLVSTRDPHRHELRSYRI | | |
| N.farc 100NAHYVLISTRDPEOHELRSYRI | | |
| | KEG-VVTPENIEVI 130 | |
| G.meta 98GVSY <mark>VIA</mark> SL-AGAEPDVKAFRI | IDG-VVEPEPIDIVE 132 | |
| | | |
| | INADYTTGQVFDLSEKLEQSEAQLGRGSF-MLGLETHDRKSEDKLAKATRD | |
| | EHCKHNESVVKEMLELAKNYNKAVEEEDKMTPEQLAIKNVGKQD-PKRHLEEH | |
| | GNGDYVAGQISDLAEKLEQAESQLANSRYGGIAPAGHQRRKEDEPQLAKITRD | |
| | EKEESNLAATKSMVKIAEQYSKRIEEEKELTEEELKTRYVGRQD-PKKHLSET | |
| | RFDRNTVREKMRALASLAVQSERFIVQGLDEDDVGNVGRAN-PIAHLQSE | |
| | HDAVQQIQQITALLEGVSPSHEGKDGSGSRTRELHRQQNNREGGRRGATAVTD | |
| | LTIDEESKADDRATHTVERPSDFEMYADRTPLETLQHLQSALKLATTSNVSRP GHVREINSKLQVVDLRVKGHVREINSKLQVVDLRVK | |
| | STFDFMHILGESSG-KTRDKPLSEEQESL | |
| COD_CT413 ISI AUKAMUTESILHAT | 91576/101105990-VIKDV57976/597 | ZUZ |

Figure 2. Continued.



operon-like gene cluster for eukaryotic ubl system

Figure 3. The gene cluster of the Ub-like protein modifier system in *C. subterraneum.* CDSs without gene annotation encode hypothetical proteins. CDSs; *rpn111* (CSUB_C1473), *ubl* (CSUB_C1474), *e21* (CSUB_C1475), *e11* (CSUB_C1476) and *srfp* (CSUB_C1477) encode eukaryotic RPN11, Ubl, E21 and E11 and small RING finger protein, respectively.

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

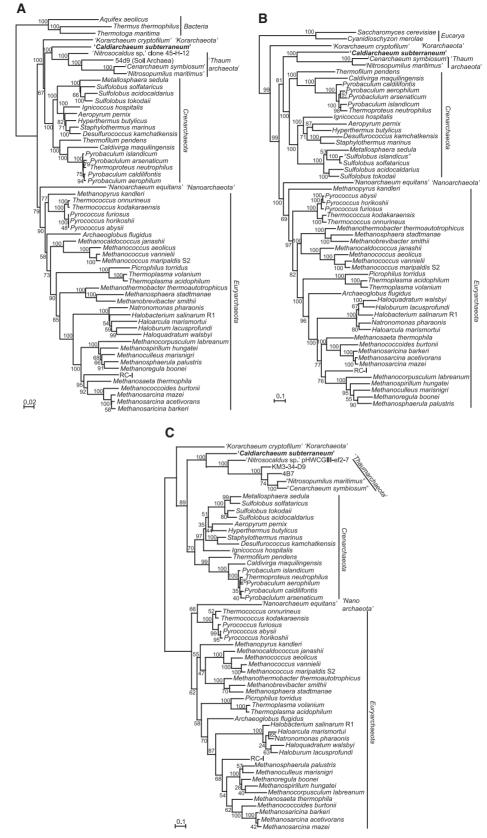


Figure 4. Phylogenetic analyses of *Archaea* including *C. subterraneum.* (A) Maximum likelihood phylogenetic tree of concatenated (SSU+LSU) rRNA genes using 3063 identical nucleotide positions. Bacterial sequences were used as out-group. Numbers indicate bootstrap values from 100 replications. (B) Maximum 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.

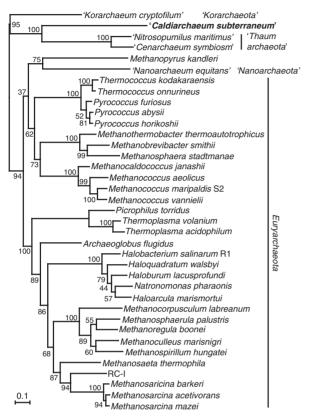


Figure 5. Maximum likelihood phylogenetic tree of concatenated (SSU+LSU) DNAP. Number of identical amino acid residues used were 829. Numbers indicate bootstrap values (%) from 200 replications.

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 $35^{\circ}C$ to $60^{\circ}C$) (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 $(74^{\circ}C)$ (7) and mesophilic thaumarchaeotes, and the branch lengths of C. subterraneum and thaumarchaeal sequences in the

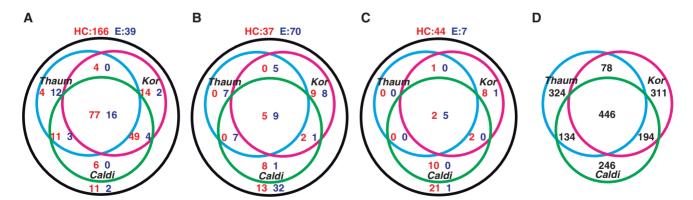


Figure 6. Venn diagrams presenting number of arCOGs among crenarchaeotic lineages; *Caldiarchaeum, Korarchaeum* and *Thaumarchaeota*. (A, B and C) Venn diagrams presenting number of arCOGs represents genome core genes of hyperthermophilic *Crenarchaeota* (HC: red) and *Euryarchaeota* (E: blue) in the genomes of the novel crenarchaeal lineages; *Caldiarchaeum subterraneum* (*Caldi*), *Thaumarchaeota* (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 eurvarchaeotes (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 Eurvarchaeota 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 between the relationship 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 groups; Miscellaneous Crenarchaeotic crenarchaeal 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 avyn '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, E21 and Ell 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 RPN111 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, E11, E21 and adjacent RPN111 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 prokarvotes, also occurred in the last common ancestor of Eucarva 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 E11) 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, E11, E21 and RPN111 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 pHWCGIII-ef2-7; AB543518. (HWCGIII), sp. 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 (Grantin-Aid No. 22370066 to A.K., in part); Scientific Research 'Comprehensive Priority Areas Genomics' on (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.

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Hochstrasser, M. (2009) Origin and function of ubiquitin-like proteins. *Nature*, 458, 422–429.
- İyer,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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Nunoura, T., Ĥirayama, 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.
- 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.
- DeLong, E.F. (1992) Archaea in coastal marine environments. Proc. Natl Acad. Sci. USA, 89, 5685–5689.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, 17, 540–552.
- 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.
- Stamatakis, A. (2006) RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22, 2688–2690.

- 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.
- Hochstrasser, M. (2000) Evolution and function of ubiquitin-like protein-conjugation systems. *Nat. Cell Biol.*, 2, E153–157.
- 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.
- 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.
- Filée, J., Siguier, P. and Chandler, M. (2007) Insertion sequence diversity in archaea. *Microbiol. Mol. Biol. Rev.*, 71, 121–157.
- 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.
- Cann, I.K. and Ishino, Y. (1999) Archaeal DNA replication: identifying the pieces to solve a puzzle. *Genetics*, 152, 1249–1267.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.

- 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.
- 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 CO₂ 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.
- 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.
- Ramos-Vera, W.H., Berg, I.A. and Fuchs, G. (2009) Autotrophic carbon dioxide assimilation in *Thermoproteales* revisited. *J. Bacteriol.*, **191**, 4286–4297.
- 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.
- Grochowski, L.L., Xu, H. and White, R.H. (2005) Ribose-5phosphate biosynthesis in *Methanocaldococcus jannaschii* occurs in the absence of a pentose-phosphate pathway. *J. Bacteriol.*, 187, 7382–7389.
- 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.
- 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.
- 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.
- Murata,S., Yashiroda,H. and Tanaka,K. (2009) Molecular mechanisms of proteasome assembly. *Nat. Rev. Mol. Cell. Biol.*, 10, 104–115.
- Amerik, A.Y. and Hochstrasser, M. (2004) Mechanism and function of deubiquitinating enzymes. *Biochim. Biophys. Acta*, 1695, 189–207.
- 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.

- 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 Cull. *Science*, **298**, 608–611.
- 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.
- 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.
- Jurica, M.S. and Stoddard, B.L. (1999) Homing endonucleases: structure, function and evolution. *Cell. Mol. Life Sci.*, 55, 1304–1326.
- 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.
- 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.

- 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.
- Bernander, R. (2000) Chromosome replication, nucleoid segregation and cell division in archaea. *Trends Microbiol.*, 8, 278–283.
- 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.