Life **2015**, *5*, 818-840; doi:10.3390/life5010818

OPEN ACCESS

life

ISSN 2075-1729
www.mdpi.com/journal/life

Article

Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales

Kira S. Makarova *, Yuri I. Wolf and Eugene V. Koonin

National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, MD 20894, USA; E-Mails: wolf@ncbi.nlm.nih.gov (Y.I.W.); koonin@ncbi.nlm.nih.gov (E.V.K.)

* Author to whom correspondence should be addressed; E-Mail: makarova@ncbi.nlm.nih.gov; Tel.: +1-301-435-5913; Fax: +1-301-435-7793.

Academic Editors: Hans-Peter Klenk, Michael W. W. Adams and Roger A. Garrett

Received: 12 January 2015 / Accepted: 28 February 2015 / Published: 10 March 2015

Abstract: With the continuously accelerating genome sequencing from diverse groups of archaea and bacteria, accurate identification of gene orthology and availability of readily expandable clusters of orthologous genes are essential for the functional annotation of new genomes. We report an update of the collection of archaeal Clusters of Orthologous Genes (arCOGs) to cover, on average, 91% of the protein-coding genes in 168 archaeal genomes. The new arCOGs were constructed using refined algorithms for orthology identification combined with extensive manual curation, including incorporation of the results of several completed and ongoing research projects in archaeal genomics. A new level of classification is introduced, superclusters that untie two or more arCOGs and more completely reflect gene family evolution than individual, disconnected arCOGs. Assessment of the current archaeal genome annotation in public databases indicates that consistent use of arCOGs can significantly improve the annotation quality. In addition to their utility for genome annotation, arCOGs also are a platform for phylogenomic analysis. We explore this aspect of arCOGs by performing a phylogenomic study of the Thermococci that are traditionally viewed as the basal branch of the Euryarchaeota. The results of phylogenomic analysis that involved both comparison of multiple phylogenetic trees and a search for putative derived shared characters by using phyletic patterns extracted from the arCOGs reveal a likely evolutionary relationship between the Thermococci, Methanococci, and Methanobacteria. The arCOGs are expected to be instrumental for a comprehensive phylogenomic study of the archaea.

Keywords: arCOGs; genome annotation; phylogenomics; Thermococci; methanogens

1. Introduction

The quality of genome annotation and the utility of comparative genomic analysis critically depend on the accuracy of sequence comparison and reconstruction of gene and genome evolution. Evolutionary reconstruction is complicated by frequent events of gene duplication, gene loss and gain, typically, via horizontal gene transfer (HGT), gene fusions and fissions, and changing rates of gene evolution. Numerous genomic and metagenomic projects brought about an avalanche of genome sequences. Comparative analysis of these sequences increasingly relies on expert-curated databases of orthologous genes in which the common problems of automatic clustering and sequence comparison are systematically addressed. Otherwise, many errors are propagated any time when the most common sequence comparison approaches are used on the new set of genomes. Several sophisticated methods of automatic sequence clustering for the construction of orthologous gene groups have been proposed [1–4]. However, to attain accurate orthology identification and functional prediction, clusters produced by each of these approaches require continuous expert curation and additional methods such as phylogenetic analysis [5–9].

The database of archaeal Clusters of Orthologous Genes (arCOGs) was first developed in 2007 and initially covered 41 archaeal genomes [10]. Over the past years, a comprehensive update has been released, and methods for assigning arCOG membership have been refined in an attempt to achieve higher sensitivity and selectivity [11]. Importantly, manual curation of the arCOGs incorporated results of relevant research projects on archaeal genomics and evolution (e.g., [12–22]). The arCOGs database has become an important tool for archaeal genome analysis and annotation [19,23–34]. In particular, arCOGs have been included as the main resource for protein function annotation in archaeal genomes in the UCSC Archaeal Genome Browser [35].

Here we present a substantial update of the arCOGs that includes an extensive revision of the arCOG annotations based on comparison with the pfam, Conserved Domain Database (CDD), tigrfams, and recently updated COGs databases, literature searching, and data from several comparative genomic and phylogenomic studies. This effort resulted in the manual correction of both the composition and the annotation of many arCOGs. We also demonstrate how the arCOGs database can be efficiently used to address important evolutionary questions, by exploring shared features between the Thermococci and class I methanogenes (Methanococci, Methanobacteria, and *Methonopyrus*) [36]. The results of this analysis put into question the currently accepted position of the Thermococci in the archaeal phylogeny.

2. Materials and Methods

2.1. Genome Sequences and Basic Sequence Analysis

Genome sequences were downloaded from the NCBI FTP site [37]. New proteomes were assigned to the previous version of arCOGs using a RPS-BLAST search with a set of position-specific scoring matrices that were generated from the multiple alignments of the protein sequences of each arCOG [11]. Proteins that were not included in the existing arCOGs were clustered to create new arCOGs, as described previously [10]. Annotations of arCOGs that were linked to COGs were adopted when appropriate from recent updates in COG annotation [38]. A PSI-BLAST search [39] against the profiles available in the current version of the CDD database [40] was used to find the closest hits from arCOGs to pfam [41], conserved domain families [40], and tigrfams [42]. All annotations from these sources were compared with each other and the existing arCOG annotation and transferred to the respective arCOGs when appropriate. For the "uncharacterized" arCOGs remaining after the above approach, publicly available annotation of respective proteins has been examined and verified using PSI-BLAST and HHpred [43] programs and/or the literature search.

Transmembrane segments in protein sequences were predicted using the TMMHMM v. 2.0c program with default parameters [44]. Signal peptides were predicted using the SignalP v. 4.1c program; the union of the three predictions (gram-negative, gram-positive, and eukaryotic models) was used [45]. Uncharacterized protein families for which two or more transmembrane segments were predicted were renamed to "Uncharacterized membrane protein".

Multiple sequence alignments were constructed using a script combining MUSCLE [46] to align closely related sequences and MAFFT [47] to merge alignments. For phylogenetic reconstruction, sites with a gap character fraction >0.5 and homogeneity <0.1 [48] were removed. The FastTree program [49] with WAG evolutionary model and discrete gamma model with 20 rate categories was used for phylogenetic tree reconstruction.

2.2. Genome Weighting

Estimates that involve aggregation of data from multiple genome sequences are often affected by sampling bias. In Archaea, some phyla (Korarchaeota) are represented by a single genome (*Candidatus* Korarchaeum cryptofilum OPF8), whereas the *Sulfolobus islandicus* species is represented by 10 genomes of different isolates. To mitigate this bias, we introduce relative genome weights.

The concept of relative genome weights is based on two intuitive notions: first, closely related genomes should contribute individually less to the total clade weight than their more distant relatives; second, the relative contribution of the clades should reflect the number of independent evolutionary events that occurred in the history of the clade. To the extent that the genome-level evolutionary scenario can be represented by a tree-like graph, the sum of branch lengths in a (sub-) tree naturally embodies both concepts.

More formally, consider a node in a rooted phylogenetic tree that has several descendant subtrees, each with the sum of its descendant branch lengths of T_i and connected to the parent node by a branch of the length L_i. If the descendant node is a leaf, it has no descendants, *i.e.*, $T_i = 0$. If the total weight assigned to this node is set to W, then it is distributed between the descendant subtrees as $W_i = W(L_i + T_i)/\Sigma(L_k + T_k)$. Given a phylogenetic tree with branch lengths, the sums of branch lengths for each node can be easily computed iteratively in the leaf-to-root direction and the total tree weight can be iteratively distributed between clades and leaves in the root-to-leaf direction.

For the purpose of this work, we used an approximate phylogenetic tree reconstructed from concatenated alignments of ribosomal proteins that are universal among the archaea [12]. This tree was rooted between the common ancestor of Euryarchaeota and Nanoarchaeota and the common ancestor of the TACK superphylum (an assemblage of archaeal phyla that includes Taumarchaeota, Crenarchaeota, Korarchaeota, and Candidatus Caldiarchaeum subterreneum, only known the representative of the potential phylum Aigarchaeota) using a bacterial outgroup. This position of the archaeal root has been recently supported using a different set of markers that are shared between archaea and bacteria [50]. The weights calculated using the above procedure are robust to minor perturbations of tree topology, especially those concerning deep branching events and short internal branches. As long as rearrangements and uncertainty involve only relatively short paths on the tree, the weights are not affected substantially. The total number of the available archaeal genomes (168) was used to define the weight of the root. On this scale, the weights of individual genomes (Supplementary Table S1) varied from 0.016 (*Methanosarcina mazei* Go1) to 3.9 (*Candidatus* Korarchaeum cryptofilum OPF8).

2.3. Clade Representation in arCOGs

For the coarse-grained analysis of the distribution of various features across archaea, we defined 13 clades that roughly correspond to "classes" within archaeal phyla (Supplementary Table S1). Phyla with a low number of available genome sequences (Korarchaeota, Nanoarchaeota and Thaumarchaeota) were not subdivided; Candidatus "*Caldiarchaeum subterraneum*" genome, which represents the candidate phylum Aigarchaeota, was merged into Thaumarchaeota; *Methanopyrus kandleri* AV19, the only known representative of Methanopyri, was merged into Methanobacteria. Clade names used in this work are purely operational, do not represent formally accepted taxonomy, and do not necessarily follow conventional rules of nomenclature.

Representation of clades in arCOGs was calculated by normalizing the sum of the weights of genomes that encompass genes assigned to this particular arCOG by the sum of the weights of all genomes in the given clade.

The arCOGs that display a pattern of (near) exclusive presence in a particular set of clades were identified using the following index: $R_E = (\min C_i + 0.0001)/(\max C_j + 0.0001)$ where C_i is the number of members of the i-th clade in a given arCOG; the minimum is taken across all clades that belong to the analyzed set and the maximum is taken across all clades that do not belong to the set.

2.4. Reconstruction of Archaeal Phylogeny Using arCOGs

One thousand four hundred and two arCOGs with at least four clades represented by more than half of the genomes (representation calculated using genome weights as described above) were selected for phylogenetic analysis. Multiple alignments were constructed from the protein sequences of each arCOG; poorly aligned and fragmented sequences were removed from the alignments and clade representation was re-calculated for the sequences remaining in the alignments. Sequences from clades with less than 0.5 representation were also removed from the alignment. This procedure yielded 1366 arCOGs with at least four well-represented clades. Approximate ML phylogenetic trees were reconstructed from these alignments and rooted using the modified midpoint procedure [51].

Phylogenetic affinities of archaeal clades were assessed in these trees using a customized, automated procedure. For each clade present in these trees, all tree nodes that satisfy the following criteria were identified:

- weights of leaves descending from a node include at least 0.75 of the total weight of the given clade;
- among the leaves descending from a node, at least one other clade is represented by at least 0.75 of the total weight of this clade;
- among the internal nodes descending from a node, there is no other node satisfying the two criteria described above.

Under this procedure, the node selected to represent the phylogenetic affinity of the given clade is the shallowest node that covers most of the genomes from that clade and includes at least one of the clades that are its closest relative in this tree. The existence of such a node is guaranteed by the fact that the leaves descending from the root represent 100% of the clades in the given tree, of which there are at least four. If a clade is represented in the given arCOG by deep paralogs, there could be more than one node satisfying the above criteria; in these cases, the results are averaged across all such nodes.

The following formalism is used to calculate the phylogenetic affinity of a clade at the selected node of the tree. Any internal node of the tree $\boldsymbol{N}$ (including the root node $\boldsymbol{R}$) defines a set of descendant leaves [52]. An i-th leaf ($i\in[52]$) belongs to a clade c_i and has a weight of w_i^{ci}. For the target clade c^* and the selected node $\boldsymbol{N}^*$ one can calculate the representation of sister clades as $W_i^* = \Sigma w_j^i/\Sigma w_k^i$, where $j\in[52]$, $k\in\{\boldsymbol{L_R}\}$ and $i \neq c^*$. Normalization of W_i^* across the entire set of sister clades ($R_i^* = W_i^*/\Sigma W_k^*$, $k \neq c^*$) shows the phylogenetic affinities of the given clade in terms of other clades. Numbers close to 1 indicate that most neighbors of the given clade belong to a particular clade, suggesting sister relationships; series of smaller, almost equal, numbers suggest that the given clade is nearly basal to the subtree that includes these clades.

3. Results and Discussion

3.1. Update of the arCOGs

The updated arCOG database includes proteins from 168 complete archaeal genomes [53]. Altogether, 369,725 protein-coding genes (94% of the total gene complement) were assigned to 13,443 clusters. Furthermore, 99 protein sequences were added to the database by translating the genomic sequences, 50 of these (mostly ribosomal proteins) during this revision. High coverage of individual genomes had already been attained in the 2012 version of the arCOGs (91% of the protein-coding genes, on average) [11] and increased to 94% in the present update. The increase in most cases was due to the appearance of closely related genomes for several species. The biggest gainers are *Vulcanisaeta* species (+9%); *Aeropyrum* (+9%), *Methanocella* (+8%), *Nitrosoarchaeum* and *Nitrosopumilus* (about +7% on average) (see Supplementary Table S1). Although two Nanoarchaeal genomes are currently available and notwithstanding the tiny size of these genomes, their coverage remains the lowest at ~74%. Anomalous low coverage was also observed for the genome of *Methanosarcina mazei* Tuc01 (coverage 84%), which is closely related to *Methanosarcina mazei* Go1 (coverage 96%) but encompasses numerous small ORFs, most of which are products of multiple frameshifts of longer proteins present in *Methanosarcina mazei* Go1 and could result from pseudogenization and/or sequencing errors.

Analysis of gene frequencies (or more precisely, the representation of orthologous gene clusters) in prokaryotes reveals a universal pattern. At all evolutionary depths, the distribution of gene frequencies can be decomposed into three distinct components: the small core of (nearly) universal genes, the much larger "shell" of moderately conserved genes, and the large, incompletely defined "cloud" of rare genes [54]. In the previous update of the arCOGs, we concluded that the core and shell gene sets have already mostly saturated [11]. Thus, not surprisingly, the new estimates are very close to those obtained in the previous work, ~207 and ~2200 families, respectively (Supplementary Figure S1). Predictably, the cloud is still growing, and is now estimated to contain ~8900 families, up from ~7400 in 2012.

In the new version of the arCOGs, many clusters that encompassed only a few genomes and consisted of genes that had paralogs in other, larger arCOGs were merged into the latter, resulting in elimination of 397 old arCOGs (Figure 1). Addition of 48 new genomes resulted in the generation of 3527 new arCOGs, several of which were manually rearranged on the basis of phylogenetic analysis. Notably, these changed arCOGs included the archaeal DNA replicative polymerases of the PolB family that were now manually divided into five arCOGs according to recent comparative genomics and phylogenetic analysis [22]: arCOG15271, PolB1; arCOG00329, PolB2; arCOG00328, PolB3; arCOG04926, DNA polymerase elongation subunit, PolB family; and arCOG15272, Casposon-associated protein-primed PolB family polymerase. Similarly, another family of essential proteins involved in DNA replication, the GINS, were split into separate clusters GINS15 (arCOG00551) and GINS23 (arCOG00552). Thus, the updated arCOG database reflects the evolutionary relationships and functional diversification of these key proteins as accurately as is currently feasible.

In addition to the previously defined 20 functional categories, the updated arCOGs include a new functional group "Mobilome" [38]. The mobilome includes proteins derived from viruses, plasmids, and transposable elements. Many of the arCOGs that comprise the mobilome were transferred from the "Replication, recombination, and repair" category but most were identified in the course of our recent work in which we analyzed (putative) viruses and other mobile elements integrated into archaeal genomes [19] (Figure 1). Another major change is the annotation of many previously uncharacterized arCOGs associated with diverse secretion systems in archaea (the respective arCOGs were moved to category N, "Motility, secretion, and vesicular transport"). These functional predictions were obtained in the course of a comprehensive comparative genomic analysis of Type IV-like secretion system genes and other conserved membrane proteins in archaea ([55] and KSM, YIW, and EVK, manuscript in preparation). Furthermore, for many arCOGs that remained in the same functional category, the annotation was refined. For example, numerous diverged components of the CRISPR-Cas system [15,16,56,57] are now provided with specific annotations in the arCOGs, and the same applies to several other defense systems that have recently been investigated in detail [20,57].

Several other new features were added to the release of the arCOG database file set [53]. These include the links from arCOGs to confidently matching pfam, CDD, and tigrfam IDs, prediction of transmembrane segments and signal peptides for each protein, and more (see README file in [53] for details). Furthermore, the arCOGs now include a higher hierarchical level of classification. In cases of significant sequence similarity between two or more arCOGs, they are combined into a supercluster. These superclusters are designed to be on the COG level (less than a family) and often include paralogous families in archaea that would be assigned to one COG if combined with bacteria. In these cases, the superclusters are given the respective COG name. Currently, the largest supercluster is COG0500, S-adenosyl-methionine-dependent methyltransferases (63 arCOGs), with only a few methylases with established substrate specificity. The largest superclusters among proteins without counterparts in the

COGs are SC.00069, HEPN domain-containing proteins (5 arCOGs), SC.00027, Uncharacterized membrane protein DUF1673 family expanded in Methanosarcina species, and SC.00178, Uncharacterized protein, expanded in Halobacteria (4 arCOGs each). In addition, some superclusters include arCOGs with low sequence similarity that nevertheless have been shown to be highly diverged orthologs by additional analysis. Such examples include derived orthologs of the essential DNA replication protein CDC45 in crenarachaea [14], the OxaA/SpoJ/YidC translocases/secretases [55], and others.

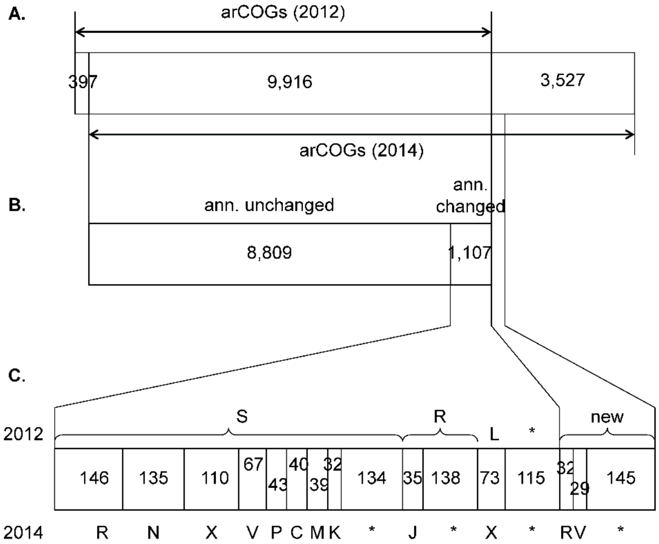

Figure 1. Changes in arCOGs between 2012 and 2014. (**A**) arCOGs added to and removed from the 2012 set; (**B**) Annotation changes and additions; (**C**) Breakdown of annotation changes into arCOG functional categories. 2012 annotations are indicated on top, 2014 annotations on the bottom. Functional categories are as follows. Information storage and processing: **J**, Translation, ribosomal structure, and biogenesis; **K**, Transcription; **L**, Replication, recombination, and repair. Cellular processes and signaling: **V**, Defense mechanisms; **M**, Cell wall/membrane/envelope biogenesis; **N**, Cell motility, secretion, and vesicular transport; **X**, Mobilome. Metabolism: **C**, Energy production and conversion; **P**, Inorganic ion transport and metabolism. Poorly characterized: **R**, General function prediction only; **S**, Function unknown. Asterisks indicate other categories combined.

Considering that genome annotation is one of the main applications of the arCOGs, we have assessed the state of annotation of genes that belong to specifically annotated arCOGs (*i.e.*, arCOGs in all categories other than "S, Uncharacterized proteins") in the respective genome submissions in the RefSeq database. We found that ~22% of proteins in these arCOGs lack meaningful annotation in the databases (*i.e.*, are listed as "hypothetical" or "uncharacterized"). Furthermore, there are many examples when more than 90% of the proteins that belong to a functionally annotated arCOG are marked as "hypothetical" in public protein databases. These poorly annotated proteins included essential components of the translation apparatus such as large ribosomal subunit proteins L45 and L47 [58], pre-rRNA processing, and ribosome biogenesis "NOL1/NOP2/fmu family" (pfam13636 and [59]). Similarly, certain essential components of the replication system often remain annotated such as members of arCOGs that include GINS proteins and RecJ/CDC45 superfamily proteins in crenarchaea, which are the components of the CMG complex [13,14,60]. The same poor state of annotation was noticed for most of the components of the ESCRT system involved in archaeal cell division and membrane remodeling [21]. The common theme among these poorly annotated proteins is the weak sequence similarity between orthologs (often associated with the small size of the respective proteins), which means the annotations are not transferred by the commonly used genome annotation pipelines. The annotation of such diverged proteins would substantially benefit from consistent use of the arCOGs, and in particular arCOG-associated PSSMs, for genome annotation.

3.2. Analysis of the Shared Features of the Thermococci with Methanococci, Methanobacteria, and Methonopyrus

There is a continuing and still intense debate on whether or not a tree of life (species tree) is an adequate representation of the evolution of archaea and bacteria ([61–69] and references therein). The underlying cause of these persistent disagreements is the difference between the topologies of the phylogenetic trees for different genes. Nevertheless, although there are very few phylogenetic trees with identical topologies, the phylogenies of the conserved and particularly (nearly) universal gene trees in archaea and bacteria are substantially congruent, suggesting that a consensus topology of such trees could represent a central trend in prokaryotic evolution. The existence of such a central trend justifies attempts to construct a "statistical tree of life" by phylogenetic analysis of subsets of the nearly universal genes that have been selected using varying criteria and different phylogenetic methods [10,12,50,69–73].

The putative global species trees of archaea and bacteria show several points of instability for which it is hard to determine a reliable branching order, with different approaches often yielding conflicting results. When it comes to the archaeal phylogeny, such persistent uncertainties include the position of the Nanoarchaea [24,50,74–76], the position of *Methanopyrus kandleri*, the monophyly of class I methanogens [36,72,73,77–79], and the validity of the TACK superphylum (an assemblage of archaeal phyla that includes Taumarchaeota, Crenarchaeota, Korarchaeota, and *Candidatus* Caldiarchaeum subterreneum, only known the representative of the potential phylum Aigarchaeota) [12,50,80,81]. Furthermore, the phylogenetic affinities of several derived archaea with small genomes, such as ARMANs and Nanosalina, remain uncertain [72,82,83].

Despite being the primary approach for the study of gene evolution, phylogenetic analysis is fraught with several notorious methodological shortcomings, such as sensitivity to the number and quality of the alignment positions, species sampling bias, and long or short branch attraction [84–86]. Different methods of tree reconstruction often produce incompatible topologies, and alternative tree topologies are often statistically indistinguishable. A complement to the traditional, sequence-based molecular phylogeny is the search for shared derived characters (synapomorphies) that has the potential to circumvent the above caveats. Here we apply a combination of phylogenetic analysis and the search for shared genes by using the phyletic patterns (patterns of gene presence/absence in archaeal lineages) of the arCOGs to shed light on the phylogenetic affinity of the Thermococci.

Until recently, the basal position of Thermococci in the euryarchaeal part of the archaeal tree remained undisputed. This position of the Thermococcal branch had been reproduced with a variety of phylogenetic methods, gene sets, and species samples [12,24,50,73,87]. However, all these analyses overlapped in that each of them included a significant fraction of genes encoding components of the translational apparatus. In a recently reported phylogeny of archaeal genes involved in DNA replication, the Thermococci were confidently grouped with class I methanogens [72]. The authors suspected that this clustering could be a tree reconstruction attraction artifact but several alternative reconstructions (e.g., different combinations of taxa) and the use of different amino acid similarity matrices have consistently produced the same clustering of these lineages [72]. A separate phylogenetic tree of family B DNA polymerases only (specifically, the PolB3 branch) also confidently groups these lineages [22]. Furthermore, Thermococci and Methanococci share two intein insertions in the same positions in this DNA polymerase gene, whereas Methanobacteria possess two split genes that most likely emerged via trans-splicing of ancestral inteins [22]. However, inteins are also present in several halobacterial PolB3 which have three intein insertions, two in positions shared with Thermococci and Methanococci—although the halobacterial inteins substantially differ from those in the other groups in sequence and length. Taken together, these observations are better compatible with the monophyly of Thermococci, Methanococci, and Methanobacteria, although alternative scenarios, such as HGT of the PolB3 genes between these archaeal groups, or differential loss of inteins in other groups of archaea, cannot be ruled out.

Prompted by these observations, we sought evidence in support of either the basal position of the Thermococci in the Euryarchaea tree (topology I) or their clustering with class I methanogens (topology II) (Figure 2). First, we reconstructed 1366 phylogenetic trees for arCOGs that have at least four well-represented clades (see Methods for details; the trees are available on [88]); 951 of these trees contained the Thermococci clade. For each of these 951 trees, the affinity of the Thermococci to other clades was quantified (see Methods for details). Ranked by the average affinity, the most common neighbors of the Thermococci were Methanococci (0.49), followed by Korarchaeota (0.48) and Nanoarchaeota (0.44) (see Supplementary Table S2).

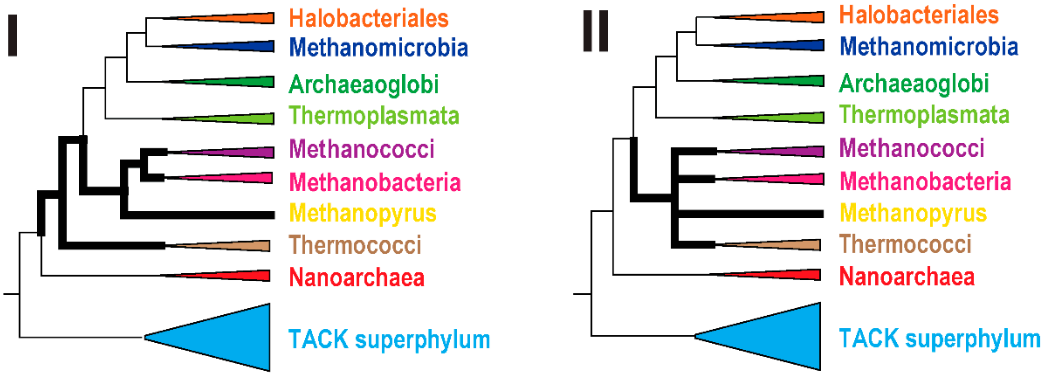

Figure 2. Alternative phylogenies for the Themococci Thick lines show the relationships between Thermococci and their closest relatives, Methanococci, Methanobacteria, and Methanopyrus. Alternative topologies of Thermococci in the Euryarchaeota tree are indicated with Roman numerals (I and II). The topology II is shown as multifurcation because the specific order of branching of these lineages has not been addressed or analyzed in this work.

Among the 219 arCOGs that belong to the archaeal core (defined here as arCOGs with all clades, excluding Nanoarchaeota, represented by at least 0.75 genomes, under the weighting scheme described in the Methods), 47 trees support topology I (basal position of Thermococci), 69 support topology II (a clade of Thermococci with Methanococci and/or Methanobacteria) (Figure 3A), and the remaining

103 trees are not fully compatible with either (e.g., Thermococci associated with non-euryarchaeal lineages). Among the trees that support topology II, 36 belong to the translation system, indicating that phylogenetic analysis of this system yields a mixed signal. Some earlier studies (e.g., [79,89]) produced trees, supporting topology II using a limited taxonomic sample that was available at the time. Nowadays this topology is often considered to be a long branch attraction artifact given that topology I appears more consistently. There is no single clearly dominant topology in the "topology X" category (Figure 3A).

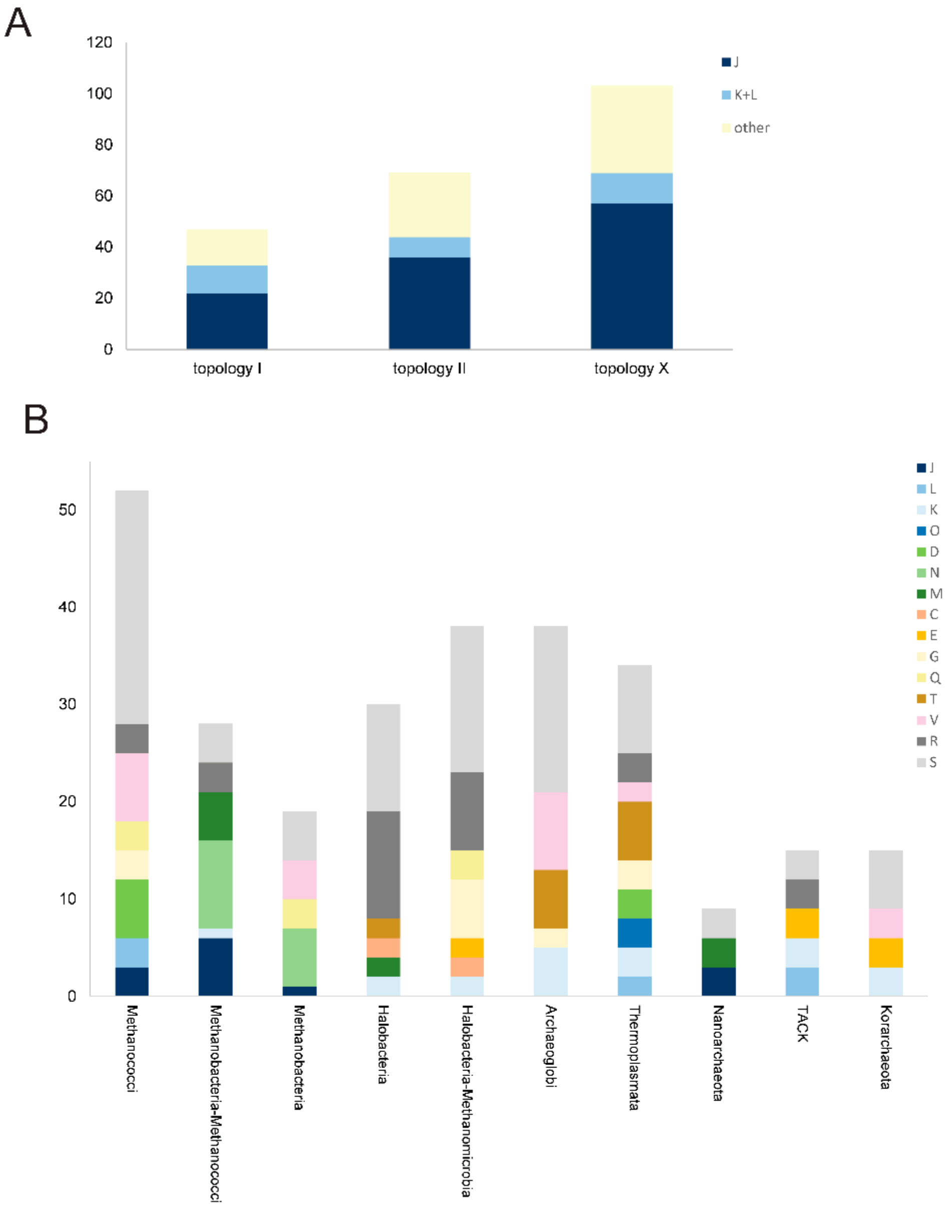

Figure 3. Phylogenomic analysis of the Thermococci. (**A**) Breakdown of the phylogenies for the 219 core arCOGs. Letters indicate the functional categories for the respective arCOGs; (**B**) Breakdown of the arCOGs that are exclusively shared between Thermococci and other archaeal lineages. The vertical axis indicates the contribution of shared arCOGs. Individual arCOGs contribute 1 to 3 units depending on their exclusivity ratio (see Methods and Supplementary Table S3). Letters indicate functional categories of arCOGs. Thermococci were compared with individual major lineages and, in two cases, with pairs of lineages related according to phylogenetic analyses.

We then analyzed the genes that are (almost) exclusively shared by Thermococci and each of the other major archaeal lineages and higher order groups, such as Halobacteria-Methanomicrobia, Methanobacteria-Methanococci, the TACK superphylum, and others (Supplementary Table S3). We selected only the arCOGs with the exclusivity ratio $E_R > 10$ (see Methods). There are only few such arCOGs for each pair of the considered lineages, and most of them belong to functional categories R and S (poorly characterized genes). The largest number of such genes are shared by Thermococci and Methanococci (Figure 3B). Among the better characterized arCOGs, Thermococci exclusively share with Methanococci slightly more genes from the informational functional categories (J, L, K, and O) and division-, secretion-, and membrane biogenesis-related genes (N, D, M) than with other lineages. This result could be explained by a single event of gene loss in the branch that includes all euryarchaea, to the exclusion of the Thermococci (topology I), rather than by the emergence of these genes in the Thermococci/class I methanogens branch (topology II). Nevertheless, we examined in greater detail the features of these genes, especially those that have the highest exclusivity ratio.

A well characterized protein in this group is enolase, a key glycolytic enzyme that additionally has been identified as a subunit of the bacterial RNA degradosome [90,91]. The enolase is often encoded in a conserved gene neighborhood within the ribosomal superoperon [92] and in the majority of the archaeal and bacterial genomes is present in a single copy. In arCOGs, enolase is represented by arCOG01169 and arCOG01170, which together belong to the supercluster COG0148. All Methanococci and Thermococci, several Metanomicrobia, and *Candidatus* Caldiarchaeum subterraneum have two enolase paralogs. We reconstructed the phylogenetic tree for the enolases of both arCOGs and overlaid the genomic contexts of the respective genes (Figure 4).

Phylogenetic analysis reveals four major branches in the enolase tree (Figure 4, branches 1–4). Three of these (branches 1, 3, and 4) share a mostly conserved context within the ribosomal superoperon. Specifically, these enolase genes are located next to genes coding for DNA-directed RNA polymerase subunit Rpo6 and ribosomal protein S2, each of which is present in a single copy per genome in most archaea. In contrast, branch 2 genes are scattered across the genome without any contextual conservation outside of the narrow clades. However, branch 2 proteins clearly cluster in the tree with branch 1, to the exclusion of branches 3 and 4. Branch 2 shows Methanococci and Thermococci as sister groups, although overall the topology in branches 1 and 2 is poorly compatible with the core phylogeny of Archaea.

Several evolutionary scenarios as well as persistent tree reconstruction artifacts can be invoked to explain the observed relationships among archaeal enolases. Evolution of enolases in archaea clearly involved at least two ancient duplications or acquisitions via HGT that produced (pseudo)paralogs in some groups and several likely xenologous gene displacements that shuffled the clades in branches 1 and 2. The disagreement between the gene context similarity and sequence conservation also make genomic rearrangements a distinct possibility.

One group of evolutionary scenarios is based on the premise that branches 1 and 2 represent the ancestral enolase, whereas branches 3 and 4 are (pseudo)paralogs present in Methanococci and Thermococci. If the tree topology is taken at face value, one would have to postulate an acquisition of a distinct enolase gene from a source outside of the available archaeal tree. Because genes in branches 3 and 4 share the context with branch 1 genes, this scenario would require subsequent intragenomic rearrangement, which would place the acquired genes into the (supposedly preferred) location within the ribosomal superoperon, displacing the original gene copies to a different location. Clearly, this is not a parsimonious scenario.

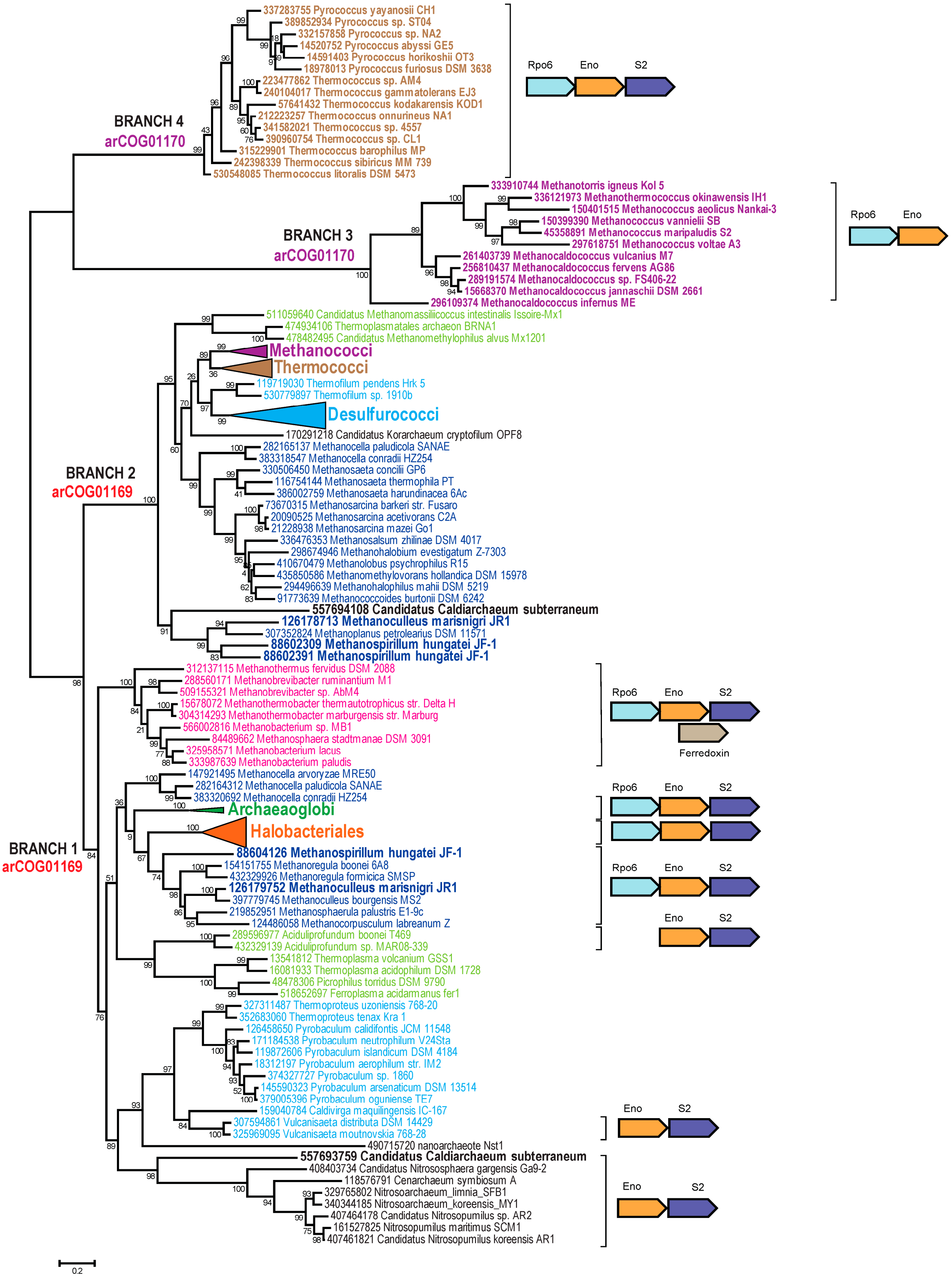
337283755 Pyrococcus yayanosii CH1
389852934 Pyrococcus sp. ST04
332157858 Pyrococcus sp. NA2
14520752 Pyrococcus abyssi GE5
14591403 Pyrococcus horikoshii OT3
18978013 Pyrococcus furiosus DSM 3638
223477862 Thermococcus sp. AM4
240104017 Thermococcus gammatolerans EJ3
57641432 Thermococcus kodakarensis KOD1
212223257 Thermococcus onnurineus NA1
341582021 Thermococcus sp. 4557
390960754 Thermococcus sp. CL1
315229901 Thermococcus barophilus MP
242398339 Thermococcus sibiricus MM 739
530548085 Thermococcus litoralis DSM 5473
BRANCH 4
arCOG01170
Rpo6
Eno
S2
333910744 Methanotorris igneus Kol 5
336121973 Methanothermococcus okinawensis IH1
150401515 Methanococcus aeolicus Nankai-3
150399390 Methanococcus vannielii SB
45358891 Methanococcus maripaludis S2
297618751 Methanococcus voltae A3
261403739 Methanocaldococcus vulcanius M7
256810437 Methanocaldococcus fervens AG86
289191574 Methanocaldococcus sp. FS406-22
15668370 Methanocaldococcus jannaschii DSM 2661
296109374 Methanocaldococcus infernus ME
BRANCH 3
arCOG01170
Rpo6
Eno
511059640 Candidatus Methanomassiliicoccus intestinalis Issoire-Mx1
474934106 Thermoplasmatales archaeon BRNA1
478482495 Candidatus Methanomethylophilus alvus Mx1201
Methanococci
Thermococci
119719030 Thermofilum pendens Hrk 5
530779897 Thermofilum sp. 1910b
Desulfurococci
170291218 Candidatus Korarchaeum cryptofilum OPF8
282165137 Methanocella paludicola SANAE
383318547 Methanocella conradii HZ254
330506450 Methanosaeta concilii GP6
116754144 Methanosaeta thermophila PT
386002759 Methanosaeta harundinacea 6Ac
73670315 Methanosarcina barkeri str. Fusaro
20090525 Methanosarcina acetivorans C2A
21228938 Methanosarcina mazei Go1
336476353 Methanosalsum zhilinae DSM 4017
298674946 Methanohalobium evestigatum Z-7303
410670479 Methanolobus psychrophilus R15
435850586 Methanomethylovorans hollandica DSM 15978
294496639 Methanohalophilus mahii DSM 5219
91773639 Methanococcoides burtonii DSM 6242
BRANCH 2
arCOG01169
557694108 Candidatus Caldiarchaeum subterraneum
126178713 Methanoculleus marisnigri JR1
307352824 Methanoplanus petrolearius DSM 11571
88602309 Methanospirillum hungatei JF-1
88602391 Methanospirillum hungatei JF-1
312137115 Methanothermus fervidus DSM 2088
288560171 Methanobrevibacter ruminantium M1
509155321 Methanobrevibacter sp. AbM4
15678072 Methanothermobacter thermautotrophicus str. Delta H
304314293 Methanothermobacter marburgensis str. Marburg
566002816 Methanobacterium sp. MB1
84489662 Methanosphaera stadtmanae DSM 3091
325958571 Methanobacterium lacus
333987639 Methanobacterium paludis
Rpo6
Eno
S2
Ferredoxin
147921495 Methanocella arvoryzae MRE50
282164312 Methanocella paludicola SANAE
383320692 Methanocella conradii HZ254
Archaeaoglobi
Halobacteriales
Rpo6
Eno
S2
88604126 Methanospirillum hungatei JF-1
154151755 Methanoregula boonei 6A8
432329926 Methanoregula formicica SMSP
126179752 Methanoculleus marisnigri JR1
397779745 Methanoculleus bourgensis MS2
219852951 Methanosphaerula palustris E1-9c
124486058 Methanocorpusculum labreanum Z
Rpo6
Eno
S2
Eno
S2
BRANCH 1
arCOG01169
289596977 Aciduliprofundum boonei T469
432329139 Aciduliprofundum sp. MAR08-339
13541812 Thermoplasma volcanium GSS1
16081933 Thermoplasma acidophilum DSM 1728
48478306 Picrophilus torridus DSM 9790
518652697 Ferroplasma acidarmanus fer1
327311487 Thermoproteus uzoniensis 768-20
352683060 Thermoproteus tenax Kra 1
126458650 Pyrobaculum calidifontis JCM 11548
171184538 Pyrobaculum neutrophilum V24Sta
119872606 Pyrobaculum islandicum DSM 4184
18312197 Pyrobaculum aerophilum str. IM2
374327727 Pyrobaculum sp. 1860
145590323 Pyrobaculum arsenaticum DSM 13514
379005396 Pyrobaculum oguniense TE7
159040784 Caldivirga maquilingensis IC-167
307594861 Vulcanisaeta distributa DSM 14429
325969095 Vulcanisaeta moutnovskia 768-28
Eno
S2
490715720 nanoarchaeote Nst1
557693759 Candidatus Caldiarchaeum subterraneum
408403734 Candidatus Nitrososphaera gargensis Ga9-2
118576791 Cenarchaeum symbiosum A
329765802 Nitrosoarchaeum_limnia_SFB1
340344185 Nitrosoarchaeum_koreensis_MY1
407464178 Candidatus Nitrosopumilus sp. AR2
161527825 Nitrosopumilus maritimus SCM1
407461821 Candidatus Nitrosopumilus koreensis AR1
Eno
S2
0.2

Figure 4. Phylogenetic analysis of the archaeal enolase family. The MUSCLE program [46] was used for construction of sequence alignment. The approximate maximum likelihood tree was reconstructed using the FastTree program [49] (178 sequences and 410 aligned positions). The sequences are denoted by their GI numbers and species names. Several

branches are collapsed and shown as triangles denoted by the respective lineage taxonomy name. The complete tree is available in Supplementary file S4. Color code is the same as in Figure 2. Species or lineages that have paralogs elsewhere in the tree are in bold. The conserved neighborhoods (if any) are shown on the right side of the tree for the respective branches. Homologous genes are shown by arrows of the same color. The arCOG numbers are provided for major branches. Abbreviations: Rpo6: DNA-directed RNA polymerase subunit K/omega and S2: RpsB, ribosomal protein S2.

Another group of scenarios posits that the (pseudo)paralogs evolved by duplication within the respective ancestral genome or by acquisition from a closely related organism so that the basal position of branches 3 and 4 is a tree reconstruction artifact caused by acceleration of evolution. Under this scenario, the paralog retaining the ancestral context underwent a period of fast evolution probably associated with a functional shift whereas the second, slowly evolving paralog most likely retained the original function while losing the original context. In more general terms, this parsimonious scenario of enolase evolution corresponds to the subfunctionalization route of evolution whereby after gene duplication, the two paralogs retain complementary subsets of the ancestral gene functions while undergoing evolutionary acceleration caused by partial loss of constraints [93]. In the case of enolase, it appears most likely that proteins in branch 2 retained the glycolytic function but lost the putative ancillary function linked to ribosomal biogenesis and dependent on the ancestral genomic context. Conversely, the enzymes in branches 3 and 4 retain only the ancillary function but not the role in glycolysis.

With respect to the evolutionary relationships between Methanococci, Methanobacteria, and Thermococci, these scenarios prompt the question how the similarity of the inferred histories of Methanococci and Thermococci, to the exclusion of Methanobacteria, came to be. If Thermococci are basal to all Euryarchaeaota (topology I), these events had to have happened in the respective ancestors of Methanococci and Thermococci independently. Alternatively, under topology II, the duplication, acquisition, or rearrangement event(s) either occurred in the common ancestor of Methanococci, Methanobacteria, and Thermococci with Methanobacteria subsequently losing the novel features; or Methanococci and Thermococci share a more recent common ancestor compared to Methanobacteria. The latter route of evolution appears especially likely under the subfunctionalization scenario outlined above, which involves dramatic acceleration of evolution in the *in situ* paralog of enolase because in this case no loss of this gene in Methanobacteria is required.

The COG00037 supercluster consists of three arCOGs (arCOG00042, arCOG00044, and arCOG00046). All these proteins belong to the TilS/MesJ family of tRNA(Ile)-lysidine synthases, members of the *N*-type ATP pyrophosphatase superfamily, found in all three domains of life [94], and are involved in an essential modification of tRNA(Ile) [95]. Two of these arCOGs are exclusively shared by Thermococci with class I methanogens. The phylogenetic tree for this supercluster (Figure 5) revealed an even more complicated picture than the enolase tree. Apparently, evolution of this family was affected by a number of evolutionary events including HGT, duplications, and accelerations of evolutionary rate. However, some interpretations appear straightforward. Most arCOG00042 representatives belong to two major branches (branches 1 and 2) that probably evolved via an ancestral duplication, although a scenario with multiple HGT events cannot be ruled out (Figure 5). Methanococci are present in both of these branches but also form three additional branches, often together with representatives of other lineages of class I methanogens that apparently encompass fast-evolving variants of this gene. Branch 3 and branch 4 correspond to arCOG00046 and arCOG00044, respectively. The clustering of Thermococci with Methanococci is observed in three branches (2, 3, and 4) (Figure 5). Under topology I, these observations imply that these three genes were acquired by Thermococci from class I methanogens via three independent HGT events. Alternative hypotheses compatible with topology I would involve a long branch attraction artifact for several paralogs that would have to be assumed to experience acceleration of evolution independently in both Thermococci and class I methanogens or multiple gene losses of 4 (or more) ancestral paralogs in all other branches of archaea. Scenarios based on topology II imply that the duplications occurred in the common ancestor of Thermococci and class I methanogens and gave rise to two or three fast-evolving TilS/MesJ family paralogs. An apparent distant homolog of the TilS/MesJ family, arCOG00045, is also shared between Thermococci and class I methanogens, to the exclusion of other archaea. This arCOG belongs to a different COG (COG01365), and, in addition to Thermococci and class I methanogens, is found only in a few bacterial genomes [38]. The functions of these proteins are not known but most of them are fused to a KH RNA-binding domain [96], implicating these enzymes in RNA modification. This family is likely to represent yet another divergent paralog resulting from duplications in the common ancestor of the same clade (Figure 5).

Many genes (Figure 3B, Supplementary Table S3) are specifically shared by the Thermococci and class I methanogens that belong to functional categories D (Cell cycle control, cell division, chromosome partitioning) and N (Motility, secretion, and vesicular transport). This group includes several genes that are associated with the ESCRT-III system, predicted to be involved in cell division and/or membrane remodeling [21]. The details on the organization of the respective loci and the presence of highly similar gene neighborhoods in Thermococci and Methanococci have been discussed previously [21]. Furthermore, Thermococci and class I methanogens share a distinct type IV-like secretion system that is centered around specific pili assembly ATPase (arCOG01819) [97] and prepilin signal peptidase EppA (arCOG02300) [98]. Notably, the ATPases of arCOG01819 in both class I methanogens and Thermococci share an intein insertion in the same position.

Several genes that are shared by Thermococci and class I methanogens do not pass the exclusivity threshold, being present in some other archaea (often Korarchaeum), but share similar features and group together in the respective phylogenetic trees. Two such genes are involved in DNA replication, namely RPA14, a subunit of the ssDNA-binding protein RPA complex (arCOG05741) and Orc1/Cdc6 homolog, an AAA superfamily ATPase (arCOG00472) [18,72]. Both of these genes show increased evolutionary rates in both Thermococci and class I methanogens.

Identification of arCOGs with phyletic patterns supporting topology I is more problematic because this topology includes Thermococci as a basal branch. Nevertheless, we searched for genes that are highly conserved in Thermococci and are shared with all other euryarchaeal lineages except for class I methanogens; this pattern implies a single gene loss in the class I methanogen branch and thus is best compatible with topology I. No genes with this pattern that would pass the exclusivity threshold were identified. Alternatively, we searched for Thermococci genes shared with the TACK superphylum but missing in other euryarchaea; this pattern implies a single loss for topology I and at least two losses for topology II depending on the branching pattern of the four major lineages within the euryarchaeal clade. We found five genes with this pattern (Figure 3B, Supplementary Table S3). Only one of these genes encodes an essential informational protein, the DNA replication initiation complex subunit of the GINS23 family (arCOG00552) [13,60].

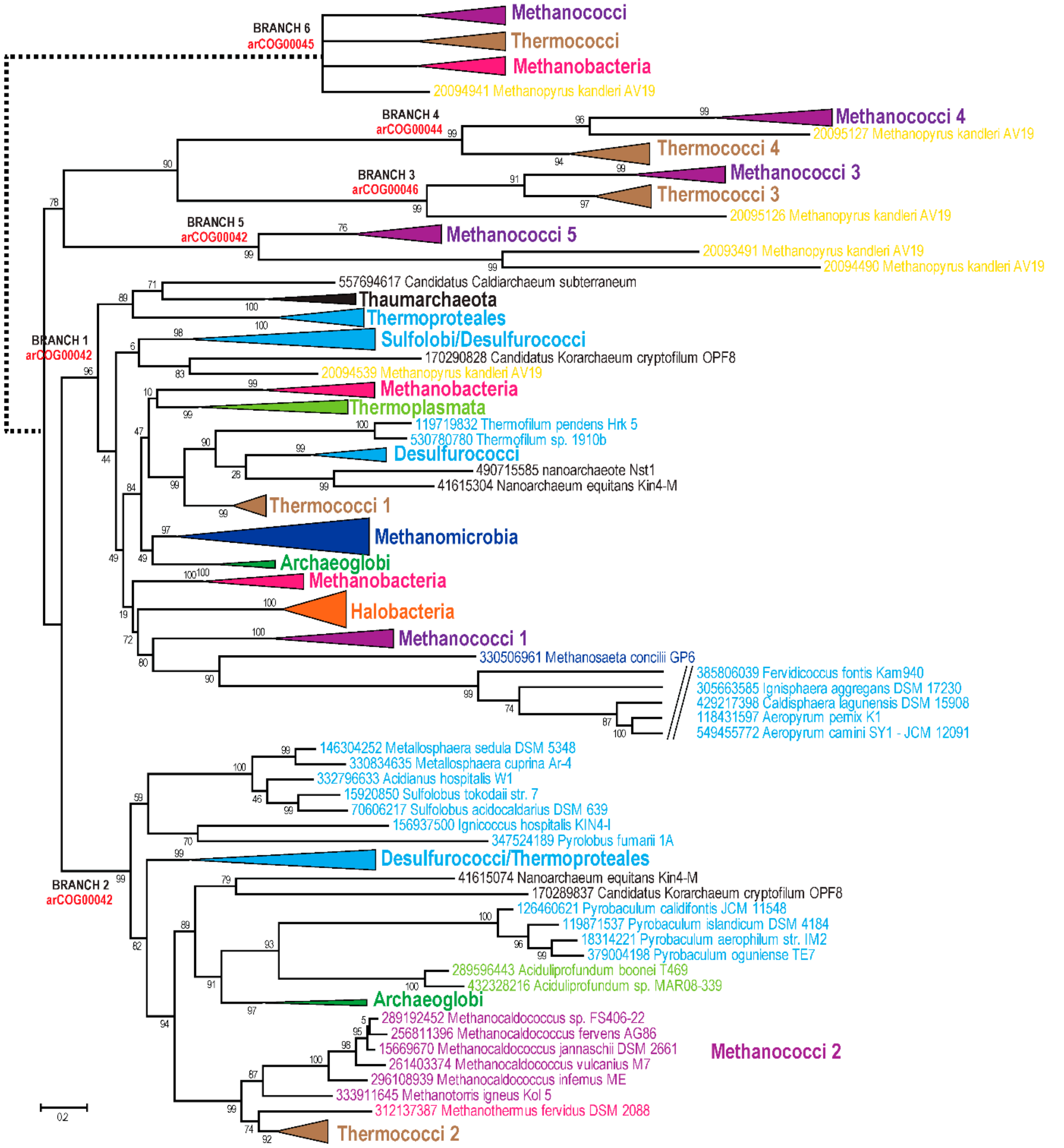
BRANCH 6
arCOG00045
Methanococci
Thermococci
Methanobacteria
20094941 Methanopyrus kandleri AV19
BRANCH 4
arCOG00044
Methanococci 4
20095127 Methanopyrus kandleri AV19
Thermococci 4
BRANCH 3
arCOG00046
Methanococci 3
Thermococci 3
20095126 Methanopyrus kandleri AV19
BRANCH 5
arCOG00042
Methanococci 5
20093491 Methanopyrus kandleri AV19
20094490 Methanopyrus kandleri AV19
557694617 Candidatus Caldiarchaeum subterraneum
Thaumarchaeota
Thermoproteales
BRANCH 1
arCOG00042
Sulfolobi/Desulfurococci
170290828 Candidatus Korarchaeum cryptofilum OPF8
20094539 Methanopyrus kandleri AV19
Methanobacteria
Thermoplasmata
119719832 Thermofilum pendens Hrk 5
530780780 Thermofilum sp. 1910b
Desulfurococci
490715585 nanoarchaeote Nst1
41615304 Nanoarchaeum equitans Kin4-M
Thermococci 1
Methanomicrobia
Archaeoglobi
Methanobacteria
Halobacteria
Methanococci 1
330506961 Methanosaeta concilii GP6
385806039 Fervidicoccus fontis Kam940
305663585 Ignisphaera aggregans DSM 17230
429217398 Caldisphaera lagunensis DSM 15908
118431597 Aeropyrum pernix K1
549455772 Aeropyrum camini SY1 - JCM 12091
146304252 Metallosphaera sedula DSM 5348
330834635 Metallosphaera cuprina Ar-4
332796633 Acidianus hospitalis W1
15920850 Sulfolobus tokodaii str. 7
70606217 Sulfolobus acidocaldarius DSM 639
156937500 Ignicoccus hospitalis KIN4-I
347524189 Pyrolobus fumarii 1A
Desulfurococci/Thermoproteales
BRANCH 2
arCOG00042
41615074 Nanoarchaeum equitans Kin4-M
170289837 Candidatus Korarchaeum cryptofilum OPF8
126460621 Pyrobaculum calidifontis JCM 11548
119871537 Pyrobaculum islandicum DSM 4184
18314221 Pyrobaculum aerophilum str. IM2
379004198 Pyrobaculum oguniense TE7
289596443 Aciduliprofundum boonei T469
432328216 Aciduliprofundum sp. MAR08-339
Archaeoglobi
289192452 Methanocaldococcus sp. FS406-22
256811396 Methanocaldococcus fervens AG86
15669670 Methanocaldococcus jannaschii DSM 2661
261403374 Methanocaldococcus vulcanius M7
296108939 Methanocaldococcus infernus ME
333911645 Methanotorris igneus Kol 5
Methanococci 2
312137387 Methanothermus fervidus DSM 2088
Thermococci 2
02

Figure 5. Phylogenetic analysis of archaeal TilS/MesJ family. The tree was reconstructed as described in Figure 4; 284 sequences and 285 aligned positions were used. The complete tree is available in Supplementary file S4. Coloring scheme is the same as in Figure 2. Sequences and collapsed branches are shown as in Figure 4.

4. Conclusions

With the accelerating accumulation of genome sequences from diverse groups of archaea and bacteria, nearly all genes in most organisms can be included in clusters of orthologs, which makes collections of such clusters increasingly efficient tools for genome annotation and phylogenomic analysis. Clearly, an essential requirement of such tools is accuracy of orthology identification and careful functional annotation of the clusters. With the arCOGs, a consistent attempt to achieve these goals was made through a combination of amended procedures for automatic clustering and intensive manual curation, aided by incorporation of the results of multiple research projects. The new arCOGs also include hierarchy in the form of superclusters that more realistically capture the evolutionary relationships between genes compared to disconnected arCOGs. Further developments are expected to include continued refinement of arCOGs, in particular by systematically applying phylogenetic analysis to resolve complex relationships within gene families with multiple paralogs. Assessment of the current archaeal genome annotation in public databases indicates that consistent use of arCOGs can significantly improve the annotation quality.

We are particularly interested in the phylogenomic applications of arCOGs aimed at comprehensive reconstruction of the evolutionary relationships between different groups of archaea and other organisms. Clearly, these relationships cannot be adequately captured by a phylogenetic tree of a single universal gene (e.g., 16S rRNA) or even a small group of topologically consistent phylogenetic trees such as those of multiple translation system components. Even if such trees show non-random topological similarity to the trees of many other genes [99], they remain “trees of 1%” [68]. In order to actually reconstruct genome and organismal evolution, phylogenomic analysis is essential. Here we employed the arCOGs in their capacity as a phylogenomic platform to reassess the evolutionary status of the Thermococci, which are traditionally viewed as the basal branch of the Euryarchaeota. Two features of arCOGs (and other collections of clusters of orthologs) underlie their utility for phylogenomics: (i) accurate identification of orthologous gene sets, which is a prerequisite for informative phylogenetic analysis; and (ii) straightforward extraction of phyletic patterns of any gene, which provides for identification of putative derived shared characters (synapomorphies). We applied both these approaches to the phylogenomic study of the Thermococci and in both cases detected a signal of evolutionary affinity with the Methanococci and Methanobacteria although a weaker signal corresponding to a basal position of the Thermococci in the Euryarchaea tree was detected as well. Clearly, all genomes have mixed legacy so that their evolution cannot be represented by a single position in a tree (even if it is a consensus of numerous trees), as emphasized, for instance, by recent observations indicating that the emergence of several major groups of archaea could have been triggered by an influx of bacterial genes [70]. Nevertheless, the findings reported here strongly suggest that the Thermococci share substantial common history with class I methanogens, which could involve shared ancestry, with vertical descent, extensive HGT, or both. We expect arCOGs to be instrumental for phylogenomic dissection of the evolution of other groups of archaea as well.

Acknowledgments

We thank Galperin M.Y. for help with annotation of several arCOGs. Kira S. Makarova, Yuri I. Wolf and Eugene V. Koonin are supported by the NIH Intramural Research Program at the National Library of Medicine, US Department of Health and Human Services.

Author Contributions

Kira S. Makarova collected the data; Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin analyzed the data and wrote the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Halachev, M.R.; Loman, N.J.; Pallen, M.J. Calculating orthologs in bacteria and Archaea: A divide and conquer approach. *PLoS One* **2011**, *6*, doi:10.1371/journal.pone.0028388.
2. Kristensen, D.M.; Wolf, Y.I.; Mushegian, A.R.; Koonin, E.V. Computational methods for Gene Orthology inference. *Brief. Bioinform.* **2011**, *12*, 379–391.
3. Altenhoff, A.M.; Dessimoz, C. Inferring orthology and paralogy. *Methods Mol. Biol.* **2012**, *855*, 259–279.
4. Trachana, K.; Larsson, T.A.; Powell, S.; Chen, W.H.; Doerks, T.; Muller, J.; Bork, P. Orthology prediction methods: A quality assessment using curated protein families. *Bioessays* **2011**, *33*, 769–780.
5. Poux, S.; Magrane, M.; Arighi, C.N.; Bridge, A.; O'Donovan, C.; Laiho, K. Expert curation in UniProtKB: A case study on dealing with conflicting and erroneous data. *Database* **2014**, doi:10.1093/database/bau016.
6. Matsuya, A.; Sakate, R.; Kawahara, Y.; Koyanagi, K.O.; Sato, Y.; Fujii, Y.; Yamasaki, C.; Habara, T.; Nakaoka, H.; Todokoro, F.; *et al*. Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.* **2008,** *36*, D787–D792.
7. Mazandu, G.K.; Mulder, N.J. The use of semantic similarity measures for optimally integrating heterogeneous Gene Ontology data from large scale annotation pipelines. *Front. Genet.* **2014**, *5*, doi:10.3389/fgene.2014.00264.
8. Trachana, K.; Forslund, K.; Larsson, T.; Powell, S.; Doerks, T.; von Mering, C.; Bork, P. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One* **2014**, *9*, doi:10.1371/journal.pone.0111122.
9. Bocs, S.; Danchin, A.; Medigue, C. Re-annotation of genome microbial coding-sequences: Finding new genes and inaccurately annotated genes. *BMC Bioinform.* **2002**, *3*, doi:10.1186/1471-2105-3-5.
10. Makarova, K.S.; Sorokin, A.V.; Novichkov, P.S.; Wolf, Y.I.; Koonin, E.V. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct* **2007**, *2*, doi:10.1186/1745-6150-2-33.
11. Wolf, Y.I.; Makarova, K.S.; Yutin, N.; Koonin, E.V. Updated clusters of orthologous genes for Archaea: A complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct* **2012**, *7*, doi:10.1186/1745-6150-7-46.
12. Yutin, N.; Puigbo, P.; Koonin, E.V.; Wolf, Y.I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS One* **2012**, *7*, doi:10.1371/journal.pone.0036972.
13. Marinsek, N.; Barry, E.R.; Makarova, K.S.; Dionne, I.; Koonin, E.V.; Bell, S.D. GINS, a central nexus in the archaeal DNA replication fork. *EMBO Rep.* **2006**, *7*, 539–545.

14. Makarova, K.; Kelman, Z.; Koonin, E.V. The archaeal CMG (CDC45/RecJ, MCM, GINS) complex is a conserved component of the DNA replication system in all archaea and eukaryotes. *Biol. Direct* **2012**, *13*, doi:10.1186/1745-6150-7.
15. Makarova, K.S.; Anantharaman, V.; Grishin, N.V.; Koonin, E.V.; Aravind, L. CARF and WYL domains: Ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **2014**, *5*, doi:10.3389/fgene.2014.00102.
16. Makarova, K.S.; Aravind, L.; Wolf, Y.I.; Koonin, E.V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct* **2011**, *6*, doi:10.1186/1745-6150-6-38.
17. Makarova, K.S.; Koonin, E.V. Two new families of the FtsZ-tubulin protein superfamily implicated in membrane remodeling in diverse bacteria and archaea. *Biol. Direct* **2010**, *5*, doi:10.1186/1745-6150-5-33.
18. Makarova, K.S.; Koonin, E.V. Archaeology of eukaryotic DNA replication. *Cold Spring Harb. Perspect. Biol.* **2013**, *5*, doi:10.1101/cshperspect.a012963.
19. Makarova, K.S.; Wolf, Y.I.; Forterre, P.; Prangishvili, D.; Krupovic, M.; Koonin, E.V. Dark matter in archaeal genomes: A rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles* **2014**, *18*, 877–893.
20. Makarova, K.S.; Wolf, Y.I.; Snir, S.; Koonin, E.V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* **2011**, *193*, 6039–6056.
21. Makarova, K.S.; Yutin, N.; Bell, S.D.; Koonin, E.V. Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat. Rev. Microbiol.* **2010**, *8*, 731–741.
22. Makarova, K.S.; Krupovic, M.; Koonin, E.V. Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. *Front. Microbiol.* **2014**, *5*, doi:10.3389/fmicb.2014.00354.
23. Esser, D.; Pham, T.K.; Reimann, J.; Albers, S.V.; Siebers, B.; Wright, P.C. Change of carbon source causes dramatic effects in the phospho-proteome of the archaeon Sulfolobus solfataricus. *J. Proteome Res.* **2012**, *11*, 4823–4833.
24. Podar, M.; Makarova, K.S.; Graham, D.E.; Wolf, Y.I.; Koonin, E.V.; Reysenbach, A.L. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* **2013**, *8*, doi:10.1186/1745-6150-8.
25. Siebers, B.; Zaparty, M.; Raddatz, G.; Tjaden, B.; Albers, S.V.; Bell, S.D.; Blombach, F.; Kletzin, A.; Kyrpides, N.; Lanz, C.; *et al*. The complete genome sequence of Thermoproteus tenax: A physiologically versatile member of the Crenarchaeota. *PLoS One* **2011**, *6*, doi:10.1371/journal.pone.0024222
26. Podar, M.; Anderson, I.; Makarova, K.S.; Elkins, J.G.; Ivanova, N.; Wall, M.A.; Lykidis, A.; Mavromatis, K.; Sun, H.; Hudson, M.E.; *et al*. A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. *Genome Biol.* **2008**, *9*, doi:10.1186/gb-2008-9-11-r158.
27. 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*. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 8102–8107.

28. Borrel, G.; Parisot, N.; Harris, H.M.; Peyretaillade, E.; Gaci, N.; Tottey, W.; Bardot, O.; Raymann, K.; Gribaldo, S.; Peyret, P.; *et al.* Comparative genomics highlights the unique biology of Methanomassiliicoccales, a Thermoplasmatales-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* **2014**, *15*, doi:10.1186/1471-2164-15-679.
29. Reimann, J.; Esser, D.; Orell, A.; Amman, F.; Pham, T.K.; Noirel, J.; Lindas, A.C.; Bernander, R.; Wright, P.C.; Siebers, B.; *et al.* Archaeal signal transduction: Impact of protein phosphatase deletions on cell size, motility, and energy metabolism in Sulfolobus acidocaldarius. *Mol. Cell. Proteomics* **2013**, *12*, 3908–3923.
30. Goncearenco, A.; Berezovsky, I.N. Exploring the evolution of protein function in Archaea. *BMC Evol. Biol.* **2012**, *12*, doi:10.1186/1471-2148-12-75.
31. Leahy, S.C.; Kelly, W.J.; Altermann, E.; Ronimus, R.S.; Yeoman, C.J.; Pacheco, D.M.; Li, D.; Kong, Z.; McTavish, S.; Sang, C.; *et al.* The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. *PLoS One* **2010**, *5*, doi:10.1371/journal.pone.0008926.
32. Csuros, M.; Miklos, I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.* **2009**, *26*, 2087–2095.
33. Hooper, S.D.; Anderson, I.J.; Pati, A.; Dalevi, D.; Mavromatis, K.; Kyrpides, N.C. Integration of phenotypic metadata and protein similarity in Archaea using a spectral bipartitioning approach. *Nucleic Acids Res.* **2009**, *37*, 2096–2104.
34. Lindas, A.C.; Karlsson, E.A.; Lindgren, M.T.; Ettema, T.J.; Bernander, R. A unique cell division machinery in the Archaea. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18942–18946.
35. Chan, P.P.; Holmes, A.D.; Smith, A.M.; Tran, D.; Lowe, T.M. The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Res.* **2012**, *40*, D646–D652.
36. Anderson, I.; Ulrich, L.E.; Lupa, B.; Susanti, D.; Porat, I.; Hooper, S.D.; Lykidis, A.; Sieprawska-Lupa, M.; Dharmarajan, L.; Goltsman, E.; *et al.* Genomic characterization of methanomicrobiales reveals three classes of methanogens. *PLoS One* **2009**, *4*, doi:10.1371/journal.pone.0005797.
37. NCBI FTP site. Available online: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ (accessed on 6 March 2015).
38. Galperin, M.Y.; Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **2014**, *43*, D261–D269.
39. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
40. Marchler-Bauer, A.; Anderson, J.B.; Chitsaz, F.; Derbyshire, M.K.; DeWeese-Scott, C.; Fong, J.H.; Geer, L.Y.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; *et al.* CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.* **2009**, *37*, D205–D210.
41. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **2014**, *42*, D222–D230.

42. Haft, D.H.; Loftus, B.J.; Richardson, D.L.; Yang, F.; Eisen, J.A.; Paulsen, I.T.; White, O. TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **2001**, *29*, 41–43.
43. Soding, J.; Biegert, A.; Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005**, *33*, W244–W248.
44. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **2001**, *305*, 567–580.
45. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785–786.
46. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
47. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.
48. Yutin, N.; Makarova, K.S.; Mekhedov, S.L.; Wolf, Y.I.; Koonin, E.V. The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **2008**, *25*, 1619–1630.
49. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS One* **2010**, *5*, doi:10.1371/journal.pone.0009490.
50. Petitjean, C.; Deschamps, P.; Lopez-Garcia, P.; Moreira, D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom proteoarchaeota. *Genome Biol. Evol.* **2014**, *7*, 191–204.
51. Wolf, Y.I.; Aravind, L.; Grishin, N.V.; Koonin, E.V. Evolution of aminoacyl-tRNA synthetases—Analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **1999**, *9*, 689–710.
52. Williams, C.H.; Stillman, T.J.; Barynin, V.V.; Sedelnikova, S.E.; Tang, Y.; Green, J.; Guest, J.R.; Artymiuk, P.J. *E. coli* aconitase B structure reveals a HEAT-like domain with implications for protein-protein recognition. *Nat. Struct. Biol.* **2002**, *9*, 447–452.
53. NCBI arCOG database. Available online: ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/ (accessed on 6 March 2015).
54. Koonin, E.V.; Wolf, Y.I. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **2008**, *36*, 6688–6719.
55. Makarova, K.S.; Galperin, M.Y.; Koonin, E.V. Comparative genomic analysis of evolutionarily conserved but functionally uncharacterized membrane proteins in archaea: Prediction of novel components of secretion, membrane remodeling and glycosylation systems. *Biochimie* **2015**, doi:10.1016/j.biochi.2015.01.004.
56. Koonin, E.V.; Makarova, K.S. CRISPR-Cas: Evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **2013**, *10*, 679–686.
57. Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. The basic building blocks and evolution of CRISPR-Cas systems. *Biochem. Soc. Trans.* **2013**, *41*, 1392–1400.

58. Marquez, V.; Frohlich, T.; Armache, J.P.; Sohmen, D.; Donhofer, A.; Mikolajka, A.; Berninghausen, O.; Thomm, M.; Beckmann, R.; Arnold, G.J.; *et al.* Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-specific ribosomal proteins. *J. Mol. Biol.* **2011**, *405*, 1215–1232.
59. Wu, P.; Brockenbrough, J.S.; Paddy, M.R.; Aris, J.P. NCL1, a novel gene for a non-essential nuclear protein in Saccharomyces cerevisiae. *Gene* **1998**, *220*, 109–117.
60. Makarova, K.S.; Wolf, Y.I.; Mekhedov, S.L.; Mirkin, B.G.; Koonin, E.V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **2005**, *33*, 4626–4638.
61. Koonin, E.V. Carl Woese's vision of cellular evolution and the domains of life. *RNA Biol.* **2014**, *11*, 197–204.
62. Puigbo, P.; Wolf, Y.I.; Koonin, E.V. Seeing the Tree of Life behind the phylogenetic forest. *BMC Biol.* **2013**, *11*, doi:10.1186/1741-7007-11-46.
63. Beauregard-Racine, J.; Bicep, C.; Schliep, K.; Lopez, P.; Lapointe, F.J.; Bapteste, E. Of woods and webs: Possible alternatives to the tree of life for studying genomic fluidity in *E. coli*. *Biol. Direct* **2011**, *6*, doi:10.1186/1745-6150-6-3.
64. Bapteste, E.; O'Malley, M.A.; Beiko, R.G.; Ereshefsky, M.; Gogarten, J.P.; Franklin-Hall, L.; Lapointe, F.J.; Dupre, J.; Dagan, T.; Boucher, Y.; *et al.* Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* **2009**, *4*, doi:10.1186/1745-6150-4-34.
65. Koonin, E.V.; Wolf, Y.I. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. *Biol. Direct* **2009**, *4*, doi:10.1186/1745-6150-4-33.
66. Bapteste, E.; Susko, E.; Leigh, J.; Ruiz-Trillo, I.; Bucknam, J.; Doolittle, W.F. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol. Biol. Evol.* **2008**, *25*, 83–91.
67. Doolittle, W.F.; Bapteste, E. Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 2043–2049.
68. Dagan, T.; Martin, W. The tree of one percent. *Genome Biol.* **2006**, *7*, doi:10.1186/gb-2006-7-10-118.
69. Williams, T.A.; Embley, T.M. Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol. Evol.* **2014**, *6*, 474–481.
70. Nelson-Sathi, S.; Sousa, F.L.; Roettger, M.; Lozada-Chavez, N.; Thiergart, T.; Janssen, A.; Bryant, D.; Landan, G.; Schonheit, P.; Siebers, B.; *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **2014**, *517*, 77–80.
71. Matte-Tailliez, O.; Brochier, C.; Forterre, P.; Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **2002**, *19*, 631–639.
72. Raymann, K.; Forterre, P.; Brochier-Armanet, C.; Gribaldo, S. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol. Evol.* **2014**, *6*, 192–212.
73. Brochier, C.; Forterre, P.; Gribaldo, S. An emerging phylogenetic core of Archaea: Phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol. Biol.* **2005**, *5*, doi:10.1186/1471-2148-5-36.
74. Huber, H.; Hohn, M.J.; Rachel, R.; Fuchs, T.; Wimmer, V.C.; Stetter, K.O. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **2002**, *417*, 63–67.

75. Waters, E.; Hohn, M.J.; Ahel, I.; Graham, D.E.; Adams, M.D.; Barnstead, M.; Beeson, K.Y.; Bibbs, L.; Bolanos, R.; Keller, M.; *et al*. The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12984–12988.
76. Brochier, C.; Gribaldo, S.; Zivanovic, Y.; Confalonieri, F.; Forterre, P. Nanoarchaea: Representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol.* **2005**, *6*, doi:10.1186/gb-2005-6-5-r42.
77. Brochier, C.; Forterre, P.; Gribaldo, S. Archaeal phylogeny based on proteins of the transcription and translation machineries: Tackling the Methanopyrus kandleri paradox. *Genome Biol.* **2004**, *5*, R17.
78. Luo, H.; Sun, Z.; Arndt, W.; Shi, J.; Friedman, R.; Tang, J. Gene order phylogeny and the evolution of methanogens. *PLoS One* **2009**, *4*, doi:10.1371/journal.pone.0006069.
79. Slesarev, A.I.; Mezhevaya, K.V.; Makarova, K.S.; Polushin, N.N.; Shcherbinina, O.V.; Shakhova, V.V.; Belova, G.I.; Aravind, L.; Natale, D.A.; Rogozin, I.B.; *et al*. The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 4644–4649.
80. Guy, L.; Ettema, T.J. The archaeal “TACK” superphylum and the origin of eukaryotes. *Trends Microbiol.* **2011**, *19*, 580–587.
81. Williams, T.A.; Foster, P.G.; Nye, T.M.; Cox, C.J.; Embley, T.M. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. Biol. Sci.* **2012**, *279*, 4870–4879.
82. Rinke, C.; Schwientek, P.; Sczyrba, A.; Ivanova, N.N.; Anderson, I.J.; Cheng, J.F.; Darling, A.; Malfatti, S.; Swan, B.K.; Gies, E.A.; *et al*. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **2013**, *499*, 431–437.
83. Baker, B.J.; Comolli, L.R.; Dick, G.J.; Hauser, L.J.; Hyatt, D.; Dill, B.D.; Land, M.L.; Verberkmoes, N.C.; Hettich, R.L.; Banfield, J.F.; *et al*. Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 8806–8811.
84. Gatesy, J.; Springer, M.S. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evol.* **2014**, *80*, 231–266.
85. Lanier, H.C.; Huang, H.; Knowles, L.L. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. *Mol. Phylogenet. Evol.* **2014**, *70*, 112–119.
86. Martin, W.; Roettger, M.; Lockhart, P.J. A reality check for alignments and trees. *Trends Genet.* **2007**, *23*, 478–480.
87. Brochier-Armanet, C.; Forterre, P.; Gribaldo, S. Phylogeny and evolution of the Archaea: One hundred genomes later. *Curr. Opin. Microbiol.* **2011**, *14*, 274–281.
88. NCBI FTP site. Available online: ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/arCOG2014/ (accessed on 6 March 2015).
89. Ciccarelli, F.D.; Doerks, T.; von Mering, C.; Creevey, C.J.; Snel, B.; Bork, P. Toward automatic reconstruction of a highly resolved tree of life. *Science* **2006**, *311*, 1283–1287.
90. Lehnik-Habrink, M.; Lewis, R.J.; Mader, U.; Stulke, J. RNA degradation in Bacillus subtilis: An interplay of essential endo- and exoribonucleases. *Mol. Microbiol.* **2012**, *84*, 1005–1017.
91. Kaberdin, V.R.; Lin-Chao, S. Unraveling new roles for minor components of the *E. coli* RNA degradosome. *RNA Biol.* **2009**, *6*, 402–405.

92. Rogozin, I.B.; Makarova, K.S.; Murvai, J.; Czabarka, E.; Wolf, Y.I.; Tatusov, R.L.; Szekely, L.A.; Koonin, E.V. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.* **2002**, *30*, 2212–2223.
93. Lynch, M.; Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **2000**, *154*, 459–473.
94. Bork, P.; Koonin, E.V. A P-loop-like motif in a widespread ATP pyrophosphatase domain: Implications for the evolution of sequence motifs and enzyme activity. *Proteins* **1994**, *20*, 347–355.
95. Ikeuchi, Y.; Soma, A.; Ote, T.; Kato, J.; Sekine, Y.; Suzuki, T. molecular mechanism of lysidine synthesis that determines tRNA identity and codon recognition. *Mol. Cell* **2005**, *19*, 235–246.
96. Valverde, R.; Edwards, L.; Regan, L. Structure and function of KH domains. *FEBS J.* **2008**, *275*, 2712–2726.
97. Planet, P.J.; Kachlany, S.C.; DeSalle, R.; Figurski, D.H. Phylogeny of genes for secretion NTPases: Identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 2503–2508.
98. Szabo, Z.; Stahl, A.O.; Albers, S.V.; Kissinger, J.C.; Driessen, A.J.; Pohlschroder, M. Identification of diverse archaeal proteins with class III signal peptides cleaved by distinct archaeal prepilin peptidases. *J. Bacteriol.* **2007**, *189*, 772–778.
99. Puigbo, P.; Wolf, Y.I.; Koonin, E.V. Search for a “Tree of Life” in the thicket of the phylogenetic forest. *J. Biol.* **2009**, *8*, doi:10.1186/jbiol159.