Open access • Posted Content • DOI:10.1101/2020.07.16.207365

Early acquisition of conserved, lineage-specific proteins currently lacking functional predictions were central to the rise and diversification of archaea — Source link

Raphaël Méheust, Raphaël Méheust, Cindy J. Castelle, Alexander L. Jaffe ...+1 more authors

Institutions: University of California, Berkeley, Planetary Science Institute

Published on: 17 Jul 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Archaea and Phylogenetics

Early acquisition of conserved, lineage-specific proteins currently lacking functional predictions were central to the rise and diversification of archaea

Raphaël Méheust[+1,4], Cindy J. Castelle[1,2], Alexander L. Jaffe[5] and Jillian F. Banfield[+,1,2,3,4,5]

[+]Corresponding authors: jbanfield@berkeley.edu, raphael.meheust@berkeley,edu

[1]Department of Earth and Planetary Science, University of California, Berkeley, CA
[2]Chan Zuckerberg Biohub, San Francisco, CA
[3]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA
[4]Innovative Genomics Institute, University of California, Berkeley, CA
[5]Department of Plant and Microbial Biology, University of California, Berkeley, CA

Abstract

Recent genomic analyses of Archaea have profoundly reshaped our understanding of their distribution, functionalities and roles in eukaryotic evolution. Within the domain, major supergroups are Euryarchaeota, which includes many methanogens, the TACK, which includes Thaumarchaeaota that impact ammonia oxidation in soils and the ocean, the Asgard, which includes lineages inferred to be ancestral to eukaryotes, and the DPANN, a group of mostly symbiotic small-celled archaea. Here, we investigated the extent to which clustering based on protein family content recapitulates archaeal phylogeny and identified the proteins that distinguish the major subdivisions. We also defined 10,866 archaeal protein families that will serve as a community resource. Clustering based on these families broadly recovers the archaeal phylogenetic tree. Interestingly, all major groups are distinguished primarily by the presence of families of conserved hypothetical proteins that are either novel or so highly diverged that their functions are obscured. Given that these hypothetical proteins are near ubiquitous within phyla, we conclude that they were important in the origin of most of the major archaeal lineages.

Introduction

Until recently, the archaeal domain comprised only two phyla, the Euryarchaeota and the Crenarchaeota, most of which were described from extreme environments (Woese, Kandler, and Wheelis 1990; Woese and Fox 1977). The recovery of genomes from metagenomes without the prerequisite of laboratory cultivation has altered our view of diversity and function across the Archaea domain (Spang, Caceres, and Ettema 2017; Adam et al. 2017; Baker et al. 2020). Hundreds of genomes from little studied and newly discovered archaeal clades have provided new insights into archaeal metabolism and evolution. Now, Archaea include at least four major large groups, the Euryarchaeota (Cluster I and Cluster II; (Spang, Caceres, and Ettema 2017; Adam et al. 2017; Baker et al. 2020)), the TACK (Proteoarchaeota) (Petitjean et al. 2014), the Asgard (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), and the DPANN (Castelle et al. 2015; Rinke et al. 2013), all of which comprise several distinct phylum-level lineages. These archaea are not restricted to extreme habitats, but are widely distributed in diverse ecosystems (Spang, Caceres, and Ettema 2017; Adam et al. 2017; Baker et al. 2020).

Most studies have focused on the metabolic potential of archaea based on analysis of proteins with known functions and revealed roles in the carbon, nitrogen, hydrogen and sulfur biogeochemical cycles. For example, Euryarchaeota includes many methanogens and non-methanogens, including heterotrophs and sulfur oxidizers (Offre, Spang, and Schleper 2013). The TACK includes Thaumarchaeota, most but not all of which oxidize ammonia (Pester, Schleper, and Wagner 2011; Brochier-Armanet et al. 2008), Aigarchaeota that tend to be chemolithotrophs that oxidize reduced sulfur compounds (Hua et al. 2018), Crenarchaeota that include thermophilic sulfur oxidizers (Woese et al. 1984), and Korarchaeota, a highly undersampled group represented by amino acid degraders that anaerobically oxidize methane and also metabolize sulfur compounds (McKay et al. 2019). The Asgard have variable metabolisms and their genomes encode pathways involved in structural components that are normally considered to be eukaryotic signatures (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). The DPANN are an intriguing group that typically have very small genomes and symbiotic lifestyles (Castelle et al. 2018; Dombrowski et al. 2019). Their geochemical roles are difficult to predict, given the predominance of hypothetical proteins.

Previously, the distribution of protein families over bacterial genomes was used to provide a function rather than phylogeny-based clustering of lineages (Méheust et al. 2019). Protein clustering allows the comparison of the gene content between genomes by converting amino acid sequences into units of a common language. The method is agnostic and unbiased by preconceptions about the importance or functions of genes.

Here, we adapted this approach to evaluate the protein family-based coherence of the archaea and to test the extent to which a subdivision of archaea could be resolved based on shared protein family content. The analysis drew upon the large genome dataset that is now available for cultivated as well as uncultivated archaea (3,197 genomes). The observation that hypothetical proteins dominate the sets of co-occurring protein families that distinguish major archaeal groups indicates the importance of these protein sets in the rise of the major archaeal lineages.

Results

Genome reconstruction and collection improved the resolution of the DPANN lineages

We collected 2,618 genomes spanning all the recognized phyla and superphyla of the Archaea domain from the NCBI genome database (**Supplemental Dataset - Table S1**). To enable our analyses, we augmented the relatively limited sampling of the DPANN by adding 569 newly available DPANN draft genomes (Castelle et al. in prep.) from low oxygen marine ecosystems, an aquifer adjacent to the Colorado River, Rifle, Colorado, and from groundwater collected at the Genasci dairy farm, Modesto CA (He et al., n.d.). The 3,197 genomes were clustered at ≥ 95% average nucleotide identity (ANI) to generate 1749 clusters. We removed genomes with <70% completeness or >10% contamination or if there was < 50% of the expected columns in the alignment of 14 concatenated ribosomal proteins (**see Materials and Methods**). To avoid contamination due to mis-binning, we required that these proteins were co-encoded on a single scaffold. The average completeness of the final set of 1,179 representative genomes is 95% and 928 were >90% complete (**Supplementary Dataset - Table S1).** The 1,179 representative genomes comprise 39 phylum-level lineages including 16 phyla that have more than 10 genomes (**Supplementary Dataset - Table S1 and Figure 1**).

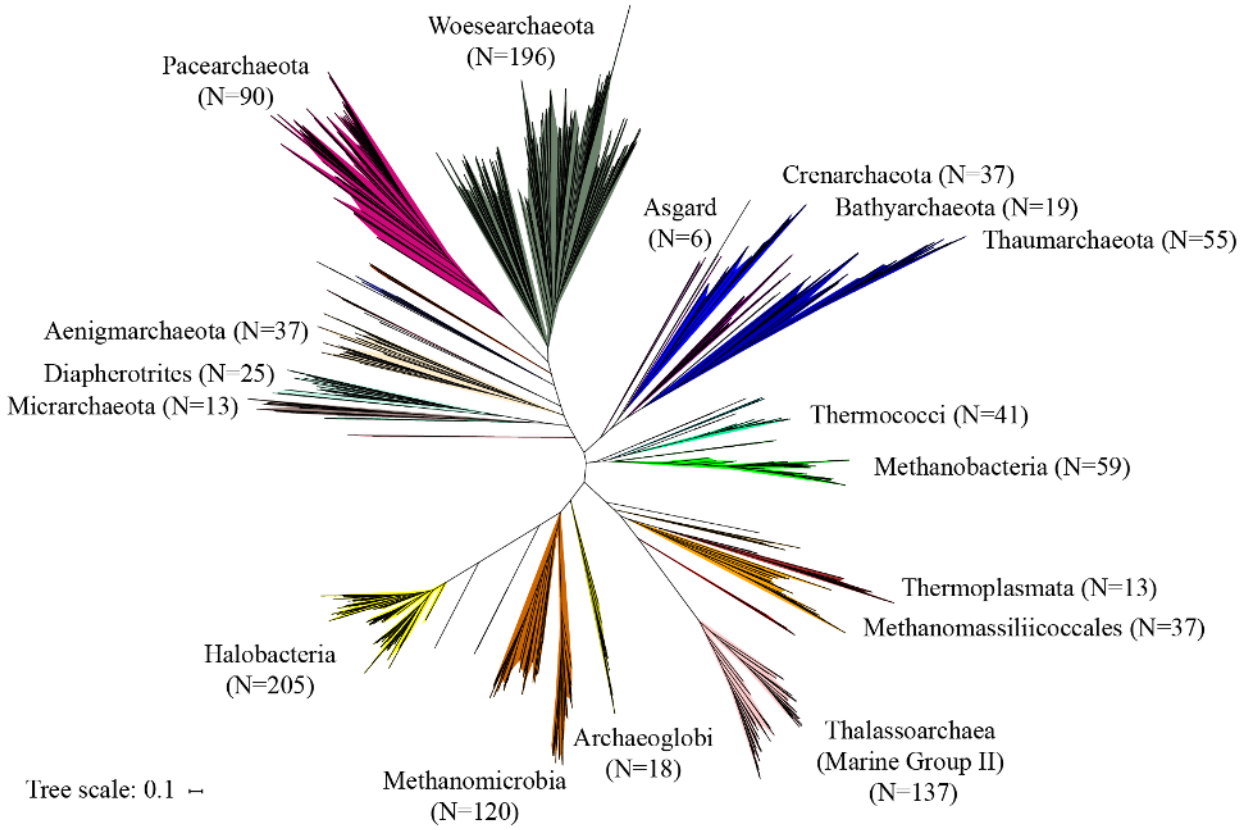


Figure 1. Phylogenetic tree of the 1,179 representative genomes. The maximum-likelihood tree was calculated based on the concatenation of 14 ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L18, L22, L24, S3, S8, S17, and S19) using the LG plus gamma model of evolution. Scale bar indicates the average substitutions per site. The complete ribosomal protein tree is available in rectangular format in Supplementary Figure 1.

Genomic content of representative genomes correlates with the phylogeny of archaea

We clustered the 2,336,157 protein sequences from the representative genomes in a two-step procedure to generate groups of homologous proteins (**Supplementary Figure 2**). This resulted in 10,866 clusters (representing 2,075,863 sequences) that were present in at least five distinct genomes. These clusters are henceforth referred to as protein families.

We assessed the quality of the protein clustering. The rationale was that we expected protein sequences with the same function to cluster into the same protein family. We annotated our protein dataset using the KEGG annotations (Kanehisa et al. 2016) and systematically verified that the protein family groupings approximate functional annotations. The KEGG annotations in our dataset encompass 6,482 unique annotations of various biological processes, including fast-evolving defense mechanisms. For each of these 6,482 annotations, we reported the family that

contains the highest percentage of protein members annotated with that KEGG annotation. Most clusters were of good quality. For 87% of the KEGG annotations (5,627 out of 6,482), one family always contained >80% of the proteins (**Supplementary Figure 3A**). The contamination of each protein family was assessed by computing the percentage of the proteins with KEGG annotations that differ from the dominant annotation (percentage annotation admixture). Most of the families contain only proteins with the same annotation, and 2,654 out of 3,746 families (71%) have <20% annotation admixture (**Supplementary Figure 3B**). Although this metric is useful, we note that it is imperfect because two homologous proteins can have different KEGG annotations and thus cluster into the same protein family, increasing the apparent percentage of annotation admixture. Although we used sensitive HMM-based sequence-comparison methods and assessed the quality of the protein clustering, we cannot completely rule out the possibility that our pipeline failed to retrieve distant homology for highly divergent proteins. Small proteins and fast-evolving proteins are more likely to be affected. This lack of sensitivity would result in the separation of homologous proteins into distinct families and would impact the results. To reduce the incidence of proteins without functional predictions for which annotations should have been achieved we augmented PFAM and KEGG-based annotations by comparing sequences to PDB database and by performing HMM-HMM comparison against the eggNog database (**see Materials and Methods**).

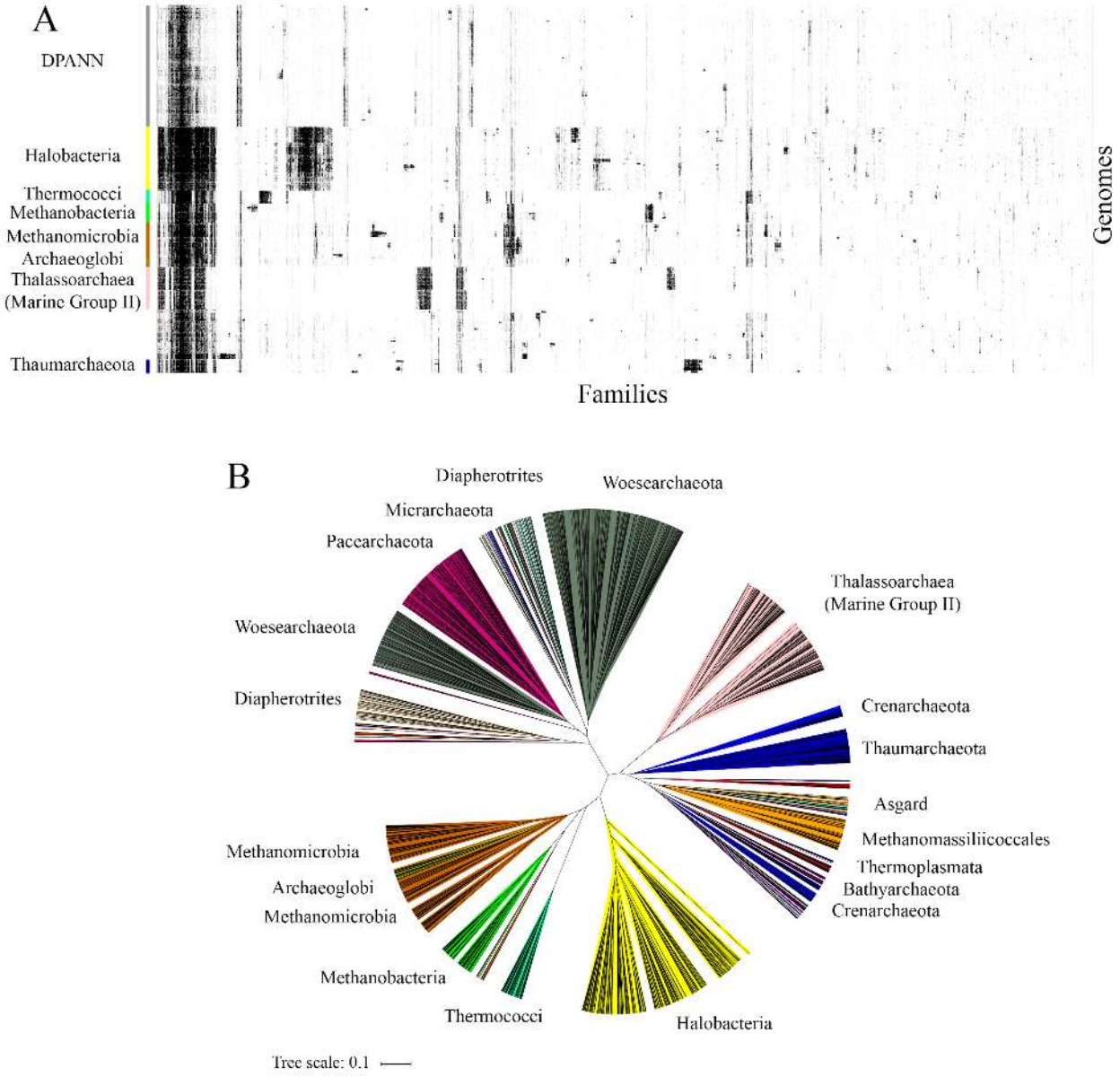


Figure 2. The distribution of the 10,866 families across the 1,179 representative genomes. A. The distribution of 10,866 widely distributed protein families (columns) in 1,179 representative genomes (rows) from Archaea. Data are clustered based on the presence (black) and absence (white) profiles (Jaccard distance, complete linkage). B. Tree resulting from the hierarchical clustering of the genomes based on the distributions of proteins families in panel A.

We visualized the distribution of the families over the genomes by constructing an array of the 1,179 representative genomes (rows) vs. 10,866 protein families (columns) and

hierarchically clustered the genomes based on profiles of protein family presence/absence **(Figure 2A)**. The families were also hierarchically clustered based on profiles of genome presence/absence. As previously reported for bacteria (Snel, Bork, and Huynen 1999; Méheust et al. 2019), the hierarchical clustering tree of the genomes resulting from the protein clustering **(Figure 2B)** correlated with the maximum-likelihood phylogenetic tree based on the concatenation of the 14 ribosomal proteins **(Figure 1)** (the cophenetic correlation based on a complete-linkage method is 0.83, based on average-linkage 0.84, and based on single-linkage, 0.84) (**Supplementary Figure 4**). Although the tree resulting from the protein families correlates with the phylogenetic tree, it does not achieve the resolution of the phylogenetic tree, especially for placement of the deep branches. Interestingly, several phyla, such as the Crenarchaeota or the Woesarchaeota, are resolved into multiple groups (**Figure 2B**). The first clade of Woesearchaeota corresponds to the Woesarchaeota-like I whereas the second clade groups together the Woesarchaeota and Woesarchaeota-like II groups. We could not evaluate the placement of Altiarchaeota relative to the DPANN because no genomes passed our quality control thresholds.

We defined modules as blocks of co-occurring protein families containing at least 20 families (see **Materials and Methods**) (Méheust et al. 2019). Each module was assigned a taxonomic distribution based on the taxonomy of the genomes with the highest number of families (see **Materials and Methods and Supplementary Dataset - Table S2**). A block of 587 protein families that was broadly conserved across the 1,179 genomes (left side in **Figure 2A**) was designed as the module of 'core families' (Module 1) (**Supplementary Figure 5).** Given their widespread distribution, it is unsurprising that most of the families are involved in well-known functions, including replication, transcription and translation, basic metabolism (oxidative phosphorylation chain, nucleotides, amino acids, ribosomal proteins, cofactors and vitamins, transporters, peptidases, DNA repair and chaperones). As expected, many of these easily recognized core families, primarily those involved in energy metabolism and cofactor synthesis, are absent in DPANN genomes (Castelle et al. 2018, 2015) (**Figure 2A and Supplementary Dataset - Table S3**). Another interesting module (module 23) (**Supplementary Figure 5**), composed of ~100 protein families, is widely distributed in most archaeal genomes but was not identified in DPANN and surprisingly, not in the Thalassoarchaea. Module 23 includes functions involved in carbon metabolism, amino-acid synthesis, and many transporter families. For instance, we identified several families for subunits of the Mrp antiporter as widespread in Halobacteria, Methanogens and Thermococci, but they appear to be absent in DPANN and Thalassoarchaea. The Mrp antiporter functions as Na+/H+ antiporter and also contributes to

sodium tolerance in Haloarchaea. Mrp has been reported to be involved in energy conservation in methanogens and in the metabolic system of hydrogen production in Thermococci.

The DPANN are an enigmatic set of lineages, the monophyly of which remains uncertain (Aouad et al. 2018). However, the protein family analysis clearly showed that these lineages group together and are distinct from other Archaea (**Figure 2B**). A detailed protein family analysis of groups within the DPANN is presented elsewhere (Castelle et al. in prep.).

Major non-DPANN groups possess groups of conserved protein families.

We detected 96 modules that are restricted to non-DPANN lineages **(Supplementary Dataset - Table S2)**. Only 9 of the 96 modules were found in multiple phyla and in 8 of these 9 cases, the phyla that possess each module are phylogenetically unrelated (e.g., Crenarchaeota and Halobacteria). The 9th, module 44, is interesting in that it occurs in two phyla and those phyla are monophyletic (Thorarchaeota and Heimdallarchaeota of the Asgard superphylum). Thus, the vast majority of the non-DPANN modules (87) are restricted to a single phylum (**Supplementary Dataset - Table S2**) and, perhaps surprisingly given phylogenetic support for superphyla within Archaea, almost no modules are specific to superphyla.

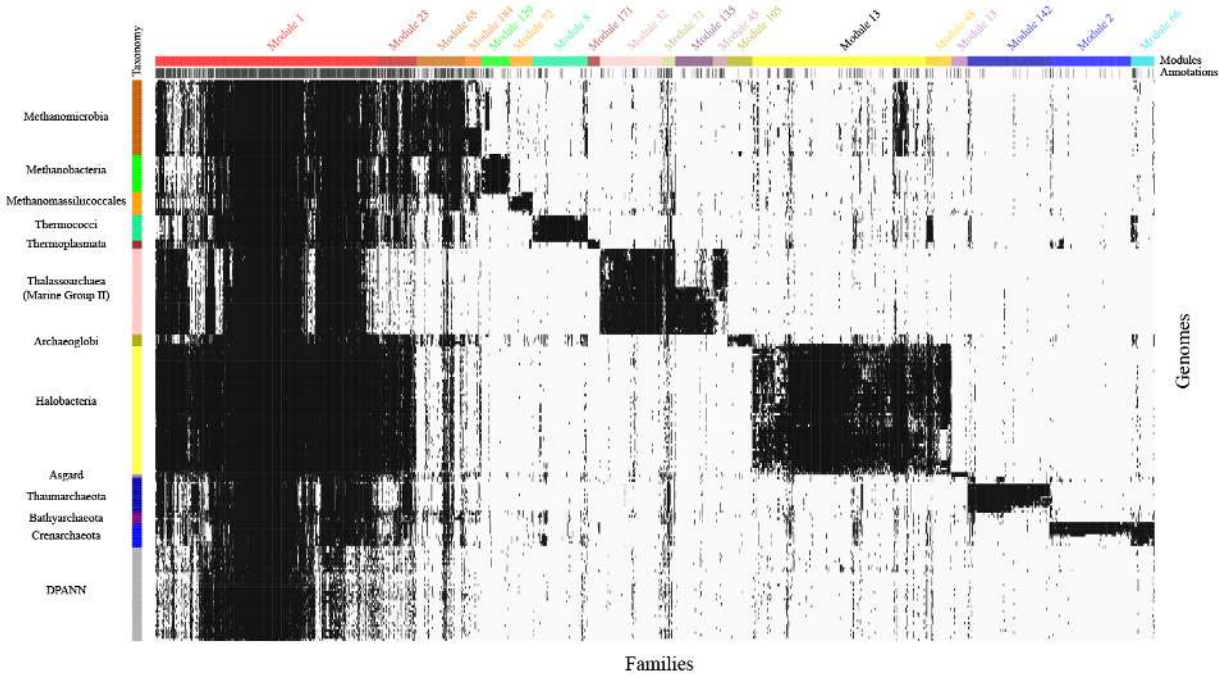


Figure 3. The distribution of the 2,632 families of the 19 modules discussed in this study. Each column represents a protein family and each row represents a genome. Data are

clustered based on the presence (black)/ absence (white) profiles but also based on the taxonomy of the genomes and the module membership. The first colored top bar (annotations) shows the families with (black) / without (white) a predicted annotation whereas the second colored top bar (modules) indicates the module of each family. The colored side bar indicates the taxonomic assignment of each genome.

Visualization of the distribution of protein families highlights the presence of modules that are not only lineage specific but are also well conserved within each lineage (**Figure 2A**). In fact, we identified such archaeal group-specific modules in 10 out of 11 non-DPANN with more than 10 genomes (**Table 1, Figure 3 and Supplementary Figure 5**). For instance, there are two modules (modules 13 and 108) comprising 525 families that are fairly conserved in Halobacteria. On average, each of the 525 families appears in 65% of the halobacterial genomes, yet these families are mostly absent in non-halobacterial genomes **(Supplementary Figure 6)**. These modules are slightly less conserved within each archaeal group than module 1 families (comprising core functions) (**Supplementary Figure 6)**.

Do methanogens cluster together, despite their phylogenetic diversity?

We identified one module of 128 protein families, module 65 (**Figure3 and Supplementary Dataset - Table S3)**, that is common to essentially all methanogens, despite the fact that methanogens are not monophyletic **(Figure 1)**. This module contained *mcrA* (Fam05485), a key gene in methane production (Ermler et al. 1997) all the other subunits (BCDG) of methyl–coenzyme M reductase (Mcr), five subunits of the methyl-tetrahydromethanopterin (methyl-H4MPT): coenzyme M methyltransferase (Mtr), five hypothetical conserved proteins in methanogens (Borrel et al. 2014) and genes for transport of iron, magnesium, cobalt and nickel and for synthesis of key cofactors that are required for growth of methanogens. Details are provided in the **Supplementary Materials**.

Modules 72, 129 and 184 are enriched in subunits of the energy-converting hydrogenase A and B and in enzymes for the utilization of methanol (fam04064 and fam05405), methylamine (fam02336 and fam03937), dimethylamine (fam03076 and fam05873), and trimethylamine (fam04092 and fam21299), which are substrates for methanogenesis (Burke, Lo, and Krzycki 1998) (for details, see the **Supplementary Materials**).

Interestingly, we recovered mcr subunits in lineages that are not considered as canonical methanogenic lineages (Evans et al. 2019). These include two genomes of Bathyarchaeota related to BA1 and BA2 (GCA_002509245.1 and GCA_001399805.1) (Evans et al. 2015), and

one Archaeoglobi genome related to JdFR-42 (GCA_002010305) (Boyd et al. 2019; Wang et al. 2019). These genomes have been described as having divergent MCR genes. It is reassuring that our method is sensitive enough to recover distant homology. Overall, the correspondence between the distribution of protein families linked to methanogenesis and methanogens supports the validity of our protein family delineation method (**Supplementary Figure 7**).

Modules	Lineage(s)	# Families	SignalP (%)	TMHMM (%)	Hypothetical families (%)	Hits to Bacteria (%)
1	Core genome	587	6	20	13	87
13,108	Halobacteria	525	14	36	82	34
66,2	Crenarchaeota	276	9	34	89	11
142	Thaumarchaeota	216	13	31	94	11
32,71	Marine Group II	199	19	55	77	32
8	Thermococci	146	12	32	84	24
65	Methanomicrobia	128	11	22	45	63
129	Methanobacteria	75	16	55	71	17
105	Archaeoglobi	65	3	40	94	12
72	Methanomassiliicoccales	59	22	49	86	27
48	Asgard	42	17	36	79	17
171	Thermoplasmata	32	3	38	97	3

Table 1. A list of the fourteen modules that are lineage specific but also well conserved within eleven major archaeal lineages. A family was counted as having a signal peptide if at least 25% of its protein sequences were predicted to have a signal peptide prediction according to the SignalP software (Almagro Armenteros et al. 2019)**. A family was counted as having a transmembrane helix if more than half of its protein sequences were predicted to have a transmembrane helix according to the TMHMM software** (Krogh et al. 2001)**. Families were considered hypothetical if they have neither PFAM (Domain of Unknown Function domains were excluded) nor KEGG annotations (see the supplementary dataset - Table S3 for the full list of hypothetical families). Finally, a family was considered to have bacterial homologs if the family matched with protein sequences of at least ten distinct bacterial genomes (see Materials and Methods). The core module 1 is included as a comparison**.

Functions specific to Thalassoarchaea

Modules 32 and 71, encompassing 199 families, were consistently associated with genomes of Thalassoarchaea (**Figure 3 and Supplementary Dataset - Table S3)**, which are implicated in protein and saccharide degradation (Tully 2019) (for details, see the **Supplementary Materials**). These modules contain protein degrading enzymes (several different classes of

peptidases and one oligotransporter) previously found in Thalassoarchaea (Tully 2019) and two new Thalassoarchaea-specific families of well-conserved peptidases. As reported by (Tully 2019), peptidase S15 (PF02129; fam03321) and peptidase M60-like (PF13402; fam05454) have a narrow distribution within Thalassoarchaea, and were not assigned to ones of the 96 modules. Interestingly, we identified modules specific to Thalassoarchaea subgroup a (MGIIa) (module_135, containing 99 families) and Thalassoarchaea subgroup b (MGIIb) (module_45, containing 39 families) with calcium-binding domains (**Supplementary Figure 8**). These proteins may be involved in signaling and regulation of protein-protein interactions in the cell (Michiels et al. 2002).

Functions specific to Crenarchaeota

The Crenarchaeota comprises thermophilic organisms that are divided into three main classes, the Thermoproteales, the Sulfolobales and the Desulfurococcales. Two distinct modules with distinct distributions were retrieved. Module 66 (61 families) is widespread in the three classes of Crenarchaeota whereas module 2 (215 families) is specific to the Sulfolobales class **(Figure 3 and Supplementary Dataset - Table S3)**. Interestingly, the subunits of RNA polymerase (Korkhin et al. 2009)), RpoG/Rpb8 (fam03177) are widespread in Crenarchaeota but Rpo13 (fam03159) seems restricted to the *Sulfolobales* class (Korkhin et al. 2009). The Rpo13 protein family of Thermoproteales and Desulfurococcales may be highly divergent from the form described experimentally.

Comparison to PDB enabled annotation of three families with no PFAM and KEGG annotations as having functions related to the DNA replication machinery (**Supplementary Dataset - Table S4)**. We were interested to find that this ubiquitous function is performed by specific protein families in Crenarchaeota, possibly reflecting adaptation to their high temperature habitats. One of these, PolB1-binding protein 2 (PBP2) (fam03141, PDB accession 5n35) (Yan et al. 2017), is a subunit of DNA polymerases B1 (PolB1) that are responsible for initial RNA primer extension with DNA, lagging and leading strand synthesis. The second is a single-stranded DNA-binding protein (DBP) ThermoDBP, which we also found to be conserved in Crenarchaeota and in Thermococci (fam03176, PDB accession 4psl) (Ghalei et al. 2014; Paytubi et al. 2012). Interestingly, however, the third is a Fe-S independent primase subunit PriX (fam03870, PDB accessions: 4wyh and 5of3) specific to Sulfolobales (**Supplementary Figure 9**). PriX is essential for the growth of Sulfolobus cells (Holzer et al. 2017; Liu et al. 2015). These observations point to fundamentally different transcription and replication mechanisms in the major groups within the Crenarchaeota.

Restricted to the Sulfolobales are also two multicopy thermostable acid protease thermopsin families (Lin and Tang 1990) (fam01298 and fam01602 in module 2). Fam01298 is also found in two genomes of Thermoproteales (**Supplementary Figure 9**). Extending a prior report that Crenarchaeota have anomalously large numbers of types I and III CRISPR-Cas systems (Vestergaard, Garrett, and Shah 2014), Crenarchaeota-specific module 66 contains four type I-A Cas families (one of which is the sulfolobales-specific CRISPR-associated protein csaX, fam07252) and four Cas families associated with type III systems (**Supplementary Figure 9**) (**Supplementary Dataset - Table S3)**.

Functions specific to Thaumarchaeota

The phylum Thaumarcheaota mostly contains aerobic ammonia oxidizing archaea (Brochier-Armanet et al. 2008; Adam et al. 2017). Module 142, which contains 216 families, is specific to Thaumarchaeota (**Figure 3 and Supplementary Dataset - Table S3**). Although this module contains protein families for the three subunits of the ammonia monooxygenase, these three families are absent in genomes for two basal Thaumarcheota lineages, as expected based on prior analyses (Adam et al. 2017) (**Supplementary Figure 10**). This module also contains a highly conserved hypothetical family (fam08021), referred to as AmoX (Bartossek et al. 2012), that is known to co-occur with the amoABC genomic cluster (**Supplementary Dataset - Table S5)**. Importantly, essentially all other protein families in Module 142 currently lack functional annotations (**Supplementary Dataset - Table S3)**.

Functions specific to Thermococci

The Thermococci comprises sulfur-reducing hyperthermophilic archaea (Palaeococcus, Thermococcus and Pyrococcus). Module 8 contains 146 families abundant in Thermococci and absent in other archaeal lineages (**Figure 3 and Supplementary Dataset - Table S3)**. For example, 98% of the Thermococci genomes have a group 3b (NADP-reducing) [NiFe] hydrogenase. This hydrogenase, also known as sulfhydrogenase, is likely bidirectional (Schut et al. 2012). Only the subunit beta of the sulfur reductase (fam04571) is present in module 8. Subunits alpha (fam00341), delta (fam00630) and gamma (fam00435) are present in the core module (module 1), probably because they are homologs of other hydrogenases. We also detected hydrogen gas-evolving membrane-bound hydrogenases (MBH) in every Thermococci genome (fam03754 in module 8) (Yu et al. 2018; Schut et al. 2016) (**Supplementary Figure 11)**. The MBH transfers electrons from ferredoxin to reduce protons to form H_2 gas (Sapra, Bagramyan, and Adams 2003). The Na^+-translocating unit of the MBH enables H_2 gas evolution

316 by MBH to establish a Na^+ gradient for ATP synthesis near 100 °C in *Pyrococcus furiosus (Yu et*
317 *al. 2018).* As with the sulfhydrogenase, only the subunit I of the MBH is present in module 8, other
318 subunits of MBH are present in core modules 1 and 23 probably because MBH-type respiratory
319 complexes are evolutionarily and functionally related to the Mrp H+/Na+ antiporter system (Yu et
320 al. 2018).
321 In the Thermococci-specific module 8 we detected the alpha and gamma subunits
322 (represented by fam10869 and fam02435, respectively) of the Na^+-pumping methylmalonyl-
323 coenzyme A (CoA) decarboxylase that performs Na^+ extrusion at the expense of the free energy
324 of decarboxylation reactions (Dimroth 1987; Buckel 2001). The beta and delta subunit, fam02317
325 and fam00273, are present in the core module 1, again probably because they are homologs of
326 proteins that perform different functions.
327 Interestingly, three families from module 8 are encoded adjacent in the Thermococci
328 genomes (fam15060, fam07594 and fam05926) (**Supplementary Dataset - Table S6)**. These
329 are annotated as of fungal lactamase (renamed prokaryotic 5-oxoprolinase A, pxpA) and
330 homologs of allophanate hydrolase subunits (renamed pxpB and pxpC) and are likely to form
331 together an 5-oxoprolinase complex (Niehaus et al. 2017). While oxoproline is a major universal
332 metabolite damage product and oxoproline disposal systems are common in all domains of life,
333 the system encoded by these three families appears to be highly conserved in Thermococci
334 (**Supplementary Figure 11)**.
335 We found the ribosomal protein L41e (fam02171) (Yutin et al. 2012) in 83% of the
336 genomes of Thermococci but sparsely distributed or absent in other lineages. It has previously
337 been noted that the distribution of L41e in Archaea is uncertain (Lecompte et al. 2002).
338 Using PDB, we established annotations for three families in Thermococci-specific module
339 8 that lacked PFAM or KEGG annotations (**Supplementary Dataset - Table S4)**. The first
340 appears to be a small protein that inhibits the proliferating cell nuclear antigen by breaking the
341 DNA clamp in *Thermococcus kodakarensis* (fam09868) (Altieri et al. 2016). The second is the S
342 component of an energy-coupling factor (ECF) transporter (fam02033) likely responsible for
343 vitamin uptake (Zhang, Wang, and Shi 2010). The third (fam01133) is the Valosin-containing
344 protein-like ATPase (VAT) that in *Thermoplasma acidophilum* functions in concert with the 20S
345 proteasome by unfolding substrates and passing them on for degradation (Huang et al. 2016).
346 Finally, three peptidases were detected in module 8 (fam01338, fam26972 and fam05052), thus
347 may be specific to the Thermococci (**Supplementary Figure 11)**.
348
349

Functions specific to Halobacteria

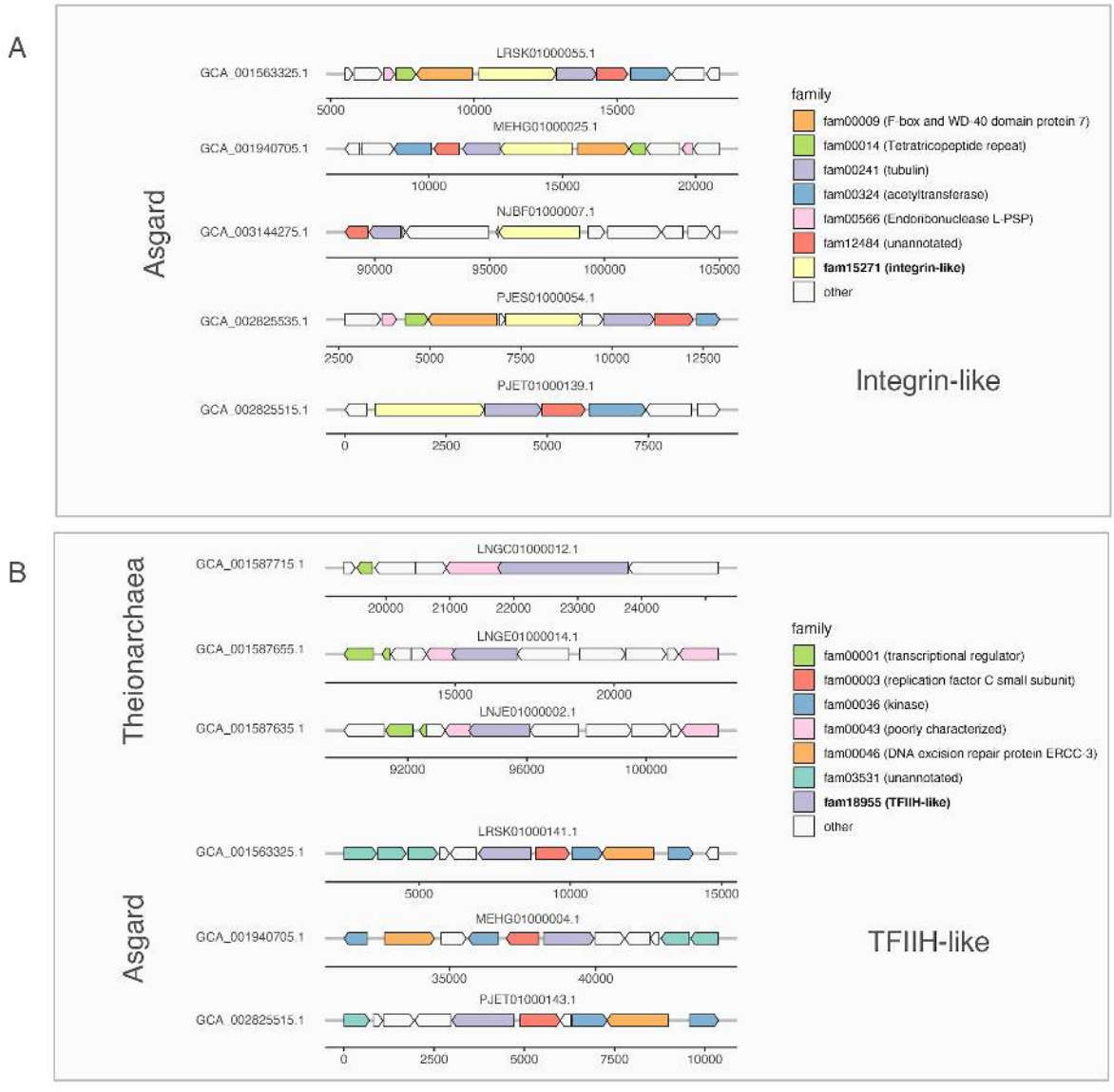
We found that 525 families comprise the Halobacteria-specific modules 13 and 108. Module 108 is composed almost completely of hypothetical proteins (**Figure 3 and Supplementary Dataset - Table S3).**

Module 13 contains the two subunits I (fam02395) and II (fam06634) of the high-affinity oxygen cytochrome *bd* oxidase (module 13) and was identified in half of the genomes (**Supplementary Fig. 12)**. It also contains three families without KEGG and PFAM annotations, but close inspection using HMM-HMM comparison showed that they have distant homology with cytochrome-related proteins (**Supplementary Dataset - Table S4)**. The first, fam02696, has distant homology with the catalytic subunit I of heme-copper oxygen reductases (fam00581) and the genes often colocalize with heme-copper oxygen reductases-related genes such as type C (*cbb*$_3$) subunit I or the nitric oxide reductase subunit B (fam00581) **(Supplementary Dataset - Table S7)**. The two other families are cytochrome c associated proteins (fam01001, cytochrome c biogenesis factor and fam02143, cytochrome C and quinol oxidase polypeptide I). Consistent with the presence of oxygen respiration-related families, a catalase-peroxidase gene is present in 90% (fam02210) of the halobacteria genomes (**Supplementary Fig. 12**). Module 13 also includes proteins for synthesis of proteinaceous gas vacuoles (fam03834, fam03740, fam02854 and fam00889; identified in more than 45% of halobacterial genomes, **Supplementary Dataset - Table S3**) that regulate buoyancy of cells in aqueous environments (DasSarma and Arora 2006). The module also includes bacterioruberin 2'', 3''-hydratase (fam00736, CruF; identified in 97% of the halobacteria genomes). Adjacent in the Halobacteria genomes are two families found in the core module 1 (fam00008 and fam00115) and annotated as digeranylgeranylglycerophospholipid reductase and UbiA prenyltransferases respectively **(Supplementary Dataset - Table S7)**. Closer inspection of these three co-encoded enzymes in *Haloarcula japonica* DSM 6131 (GCA_000336635.1) showed they are identical with the bifunctional lycopene elongase and 1,2-hydratase (LyeJ, fam00115) and the carotenoid 3,4-desaturase (CrtD, fam00008) and the bacterioruberin 2'', 3''-hydratase (CruF, fam00736) genes described in *Haloarcula japonica* JCM 7785^T (Yang et al. 2015). Together, these three enzymes can generate C50 carotenoid bacterioruberin from lycopene in *Haloarcula japonica (Yang et al. 2015)*. Our results showed that C50 carotenoid bacterioruberin is highly conserved in Halobacteria (**Supplementary Figure 12**).

Functions specific to the six Asgard genomes.

The module 48 contains 42 families that are specific and conserved in the six genomes of the superphylum Asgard (four genomes of Thorarchaeota and two genomes of

384 Heimdallarchaeota) (**Figure 3)**. Of these, 33 lack both KEGG and PFAM functional predictions
385 (**Supplementary Dataset - Table S3)**. The Asgard archaea, which affiliate with eukaryotes in the
386 tree of life (Cox et al. 2008), encode many proteins that they share with eukaryotes (Hartman and
387 Fedorov 2002). We detected four eukaryotic signature protein families (ESPs) in module 48 that
388 were described in previous studies (**Supplementary Figure 13 and Supplementary Materials**)
389 (Zaremba-Niedzwiedzka et al. 2017; Spang et al. 2015; Akıl and Robinson 2018).
390



391
392 **Figure 4. Schematic overview of integrin-like and TFIIH-like gene clusters identified in**
393 **archaea. A. Conserved gene clusters comprising archaeal integrin-like genes (fam15271)**

identified in five Asgard genomes. B. Conserved gene clusters comprising archaeal TFIIH-like genes (fam18955) identified in three Theionarchaea and three Asgard genomes. A full gene synteny and genomic context of the genes neighboring the integrin-like (fam15271) and TFIIH-like (fam18955) genes is available in Supplementary Dataset - Table S8.

Interestingly, we found a family in module 48 (fam15271) that shows sequence similarity with the integrin beta 4. To the best we know, integrin genes were never described in archaea and fam15271 may represent a new ESP. The genes of fam15271 are always located next to tubulin genes (fam00241) in the five Asgard genomes **(Figure 4A and Supplementary Dataset - Table S8)**. This is particularly interesting as recent studies have shed light on the crosstalk between integrin and the microtubule cytoskeleton (LaFlamme et al. 2018). Finally, one family in module 48 (fam18955) is annotated as the DNA excision repair protein ERCC-3 in three Asgard genomes and three Theionarchaea genomes. The genes neighboring the genes of fam18955 differ between the two lineages (**Figure 4B and Supplementary Dataset - Table S8**) and the three Asgard sequences only share between 20 and 23% protein identity with the three Theionarchaea sequences. These differences may indicate two distinct functions for this family. Fam18955 shows distant homology with the protein RAD25 of *Saccharomyces cerevisiae*. RAD25 is a DNA helicase required for DNA repair and RNA polymerase II transcription in *S. cerevisiae* (Guzder et al. 1994). RAD25 is also one of the six subunits of the transcription factor IIH (TFIIH) in *S. cerevisiae* (Sung et al. 1996). Consistent with the role of RAD25 in *S. cerevisiae*, the genes of family18955 is found next to replication factor C small subunit genes in the three Asgard genomes (**Figure 4 and Supplementary Dataset - Table S8)**.

Groups without lineage-specific metabolic signatures

The Archaeoglobi and Thermoplasmata lineages are unusual in that they have modules specific to them (modules 105 and 171 respectively), but no specific capacities were identified only in these groups based on functional predictions (**Supplementary Dataset - Table S3**). These lineage-specific modules have the highest percentage of hypothetical families of any lineage-specific module (**Table 1**).

Bathyarchaeota is the only lineage having more than 10 genomes and that does not have a specific module of families that is widespread in the 19 Bathyarchaeota genomes (**Supplementary Dataset - Table S2**). This is intriguing as Bathyarchaeota are widespread across terrestrial marine ecosystems and are known to thrive in diverse chemical environments

(Kubo et al. 2012). Alternatively, Bathyarchaeota may be a superphylum. In this case, the lack of modules shared across a superphylum would reinforce the results noted above.

Hypothetical proteins distinguish the major archaeal groups

Even after augmenting functional predictions using PDB and EggNog databases, families with functional predictions represent a tiny proportion of the protein families that comprise the lineage-specific modules (**Table 1, Figure 5 and Supplementary Dataset - Table S3**). 1053 out of 1411 hypothetical families remain unannotated (**Supplementary Dataset - Table S4**). 358 hypothetical families have small domain matches, but not enough information is available to predict functions. For example, many have domains with matches to zinc finger domains, but such domains occur in proteins with diverse functions (**Supplementary Dataset - Table S4**). We found that the hypothetical proteins are shorter than proteins from the core families of module 1 **(Supplementary Figure 14)** and are more likely to have a transmembrane helix prediction and a signal peptide prediction **(Table 1)**.

Previous studies highlighted the presence of numerous families of proteins with roles in metabolism that are of bacterial origin but occur only in specific archaeal phyla (Nelson-Sathi et al. 2015, 2012). Consequently, we compared all archaeal protein families against a database of bacterial genomes sampled from across the bacterial tree of life to determine the extent to which proteins acquired from bacteria contribute to the archaeal group specific modules (**see materials and methods**). From 3% (Thermoplasmata) to 34% (Halobacteria) of the protein families in modules that are archaeal group specific have homologs in ≥10 distinct bacterial genomes, with the exception of Methanomicrobia, where 63% of the protein families have bacterial homologs (**Table 1**). Thus, for almost all archaeal groups, the majority of the protein families that form modules that separate them from other archaeal groups did not evolve in (or were not acquired from) bacteria. Further, we conclude genes acquired from bacteria only account for a small fraction of the lineage-specific families.

Discussion

We constructed a set of protein families for Domain Archaea, each of which generally corresponds with a set of homologous proteins with the same predicted function (in cases where functions could be assigned). Protein families with functional predictions that are specific to certain archaeal lineages (e.g., genes involved in methanogenesis or ammonia oxidation) well predict functional traits specific to these lineages. These observations indicate that the protein family construction method is robust. The generated set of 10,866 protein families is provided as an important community research resource. The patterns of presence/absence of protein families across genomes highlight sets of co-occurring proteins (modules), and groupings of genomes based on these modules mostly recapitulate archaeal phylogeny.

What is most striking from our analyses is the prominence of families of hypothetical proteins in the sets of highly prevalent lineage-specific proteins. An important consideration is whether (i) divergence of the sequence of these proteins from proteins with known function simply precluded functional annotation or (ii) whether they are novel proteins that serve well known functions, or if (iii) they represent functions that are unique and evolved following the divergence of each lineage from other archaea. Our analyses were designed to avoid case (i) by relying on state-of-the-art HMM-based homology detection methods that appear to well-group proteins with shared functions (**Supplementary Figure 3**). However, the fact that we could identify some probable functions using protein modeling suggests that (i) is correct in at least a subset of cases. For instance, PriX (fam03870) has structural homology with PriL but no sequence similarity was detected between PriX and any other protein in our analysis. Both proteins are distinct components of the primase complex in *Sulfolobus solfataricus* suggesting that PriX evolved from PriL by duplication followed by subfunctionalization (Holzer et al. 2017; Liu et al. 2015). Lineage-specific hypothetical proteins that are actually homologs of known proteins but currently too divergent for functional assignment are interesting, as they may have been under pressure to evolve more rapidly than normal during lineage divergence. It is not possible to distinguish (ii) from (iii) with the data available. In general, the sets of relatively common archaeal proteins without functional assignments provides targets for future biochemical studies.

Overall, the prevalence of transmembrane helices and signal peptides in the hypothetical proteins in lineage-specific modules indicates that they are membrane associated or extracellular, thus possibly involved in cell-cell and cell-environment interactions (some may be transporters). Where the lineages are confined to specific environments (e.g., halophiles), lineage-specific protein families may have evolved to meet the requirements of that environment (case (i) or (iii)). It is important to note that some modules contain many protein families and probably represent

combinations of new functions that, at the present time, cannot be resolved. Regardless of the explanation, or combination of explanations, for the presence of large numbers of lineage-specific proteins, the results clearly indicate the importance of divergence or evolution of a specific subset of proteins during emergence of the major archaeal lineages.

Possibly also informative regarding archaeal evolution is the observation that, despite resolving a Domain-wide core module (module 1), we detected only one case where a clearly defined module is conserved at the superphylum level. It is important to note that, with additional genomes, the two newly recognized Asgard phyla may be reclassified into a single phylum, eliminating this exception. The apparent lack of protein family module support for superphyla may argue against the phyletic gradualism, in which one lineage gradually transforms into another, and favor the theory of cladogenesis, where a lineage splits into two distinct lineages (Gould and Eldredge 1977). We acknowledge that modules containing fewer than 20 protein families (the cutoff used to define modules) may be uniquely associated with superphyla, and that some potentially important archaeal lineages were not included in the current analysis due to lack of a sufficient number of high-quality genomes.

The observation that the patterns of presence/absence of shared protein families groups together archaea that historically have been assigned to the same lineage and separates them from other lineages, in combination with innumerable prior publications on archaeal physiology and taxonomy (Adam et al. 2017; Spang, Caceres, and Ettema 2017; Baker et al. 2020), supports the value of the current taxonomic classifications within Domain Archaea. Overall, the results reinforce the concept that early archaeal evolution rapidly generated the major lineages, the rise of which was linked to acquisition of a set of proteins (recognized here as modules) that were largely retained during subsequent evolution of each lineage.

519 **Methods**

520 Genome collection

521 569 unpublished genomes were added to the 2,618 genomes of Archaea downloaded
522 from the NCBI genome database in September 2018.

523 132 genomes were obtained from metagenomes of sediment samples. Sediment samples
524 were collected from the Guaymas Basin (27°N0.388, 111°W24.560, Gulf of California, Mexico)
525 during three cruises at a depth of approximately 2000 m below the water surface. Sediment cores
526 were collected during two Alvin dives, 4486 and 4573 in 2008 and 2009. Sites referred to as
527 “Megamat” (genomes starting with “Meg”) and “Aceto Balsamico” (genomes starting with “AB” in
528 name), Core sections between 0-18 cm from 4486 and from 0-33 cm 4573 and were processed
529 for these analyses. Intact sediment cores were subsampled under N_2 gas, and immediately frozen
530 at -80 °C on board. The background of sampling sites was described previously (Teske et al.
531 2016). Samples were processed for DNA isolation from using the MoBio PowerMax soil kit
532 (Qiagen) following the manufacturer’s protocol. Illumina library preparation and sequencing were
533 performed using Hiseq 4000 at Michigan State University. Paired-end reads were interleaved
534 using interleav_fasta.py (https://github.com/jorvis/biocode/blob/master/fasta/interleave_fasta.py)
535 and the interleaved sequences were trimmed using Sickle (https://github.com/najoshi/sickle) with
536 the default settings. Metagenomic reads from each subsample were individually assembled using
537 IDBA-UD with the following parameters: --pre_correction --mink 65 --maxk 115 --step 10 --
538 seed_kmer 55 (Peng et al. 2012). Metagenomic binning was performed on contigs with a
539 minimum length of 2000 bp in individual assemblies using the binning tools MetaBAT (Kang et al.
540 2015) and CONCOCT (Alneberg et al. 2014), and resulting bins were combined with using DAS
541 Tool (Sieber et al., n.d.). CheckM lineage_wf (v1.0.5) (Parks et al. 2015) was used to estimate
542 the percentage of completeness and contamination of bins. Genomes with more than 50%
543 completeness and 10% contamination were manually optimized based on GC content, sequence
544 depth and coverage using mmgenome (Karst, Kirkegaard, and Albertsen, n.d.).

545 The remaining 447 genomes came from previous sequencing and binning efforts
546 (genomes starting with “ggkbase”). In brief, 168 genomes were obtained from an aquifer adjacent
547 to the Colorado River near the town of Rifle, Colorado, USA in 2011 (Anantharaman et al. 2016),
548 50 genomes from the Crystal Geyser system in Utah, USA (Probst et al. 2014). For DNA
549 processing and sequencing methods see (Probst et al. 2017; Anantharaman et al. 2016). Forty-
550 one genomes were obtained from (Parks et al. 2017). Additionally, 188 genomes were obtained
551 from groundwater samples from Genasci Dairy Farm, located in Modesto, California (CA) as
552 described in (He et al., n.d.).

553

554 Genome completeness assessment and de-replication.

555 Genome completeness and contamination were estimated based on the presence of single-copy
556 genes (SCGs) as described in (Olm et al. 2017; Anantharaman et al. 2016). Genome
557 completeness was estimated using 38 SCGs. For non-DPANN archaea, genomes with more than
558 26 SCGs (>70% completeness) and less than 4 duplicated copies of the SCGs (<10%
559 contamination) were considered as draft-quality genomes. Due to the reduced nature of DPANN
560 genomes (Castelle et al. 2015), genomes with more than 22 SCGs and less than 4 duplicated
561 copies of the SCGs were considered as draft-quality genomes. Genomes were de-replicated
562 using dRep (Olm et al. 2017) (version v2.0.5 with ANI > 95%). The most complete and less
563 contaminated genome per cluster was used in downstream analyses.

564

565 Concatenated 14 ribosomal proteins phylogeny

566 A maximum-likelihood tree was calculated based on the concatenation of 14 ribosomal proteins
567 (L2, L3, L4, L5, L6, L14, L15, L18, L22, L24, S3, S8, S17, and S19). Homologous protein
568 sequences were aligned using MAFFT (version 7.390) (--auto option) (Katoh and Standley 2016),
569 and alignments refined to remove gapped regions using Trimal (version 1.4.22) (--gappyout
570 option) (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). The protein alignments were
571 concatenated, with a final alignment of 1,179 genomes and 2,388 positions. The tree was inferred
572 using RAxML (Stamatakis 2014) (version 8.2.10) (as implemented on the CIPRES web server
573 (Miller, Pfeiffer, and Schwartz 2010)), under the LG plus gamma model of evolution, and with the
574 number of bootstraps automatically determined via the MRE-based bootstopping criterion. A total
575 of 108 bootstrap replicates were conducted, from which 100 were randomly sampled to determine
576 support values.

577

578 Protein clustering

579 Protein clustering into families was achieved using a two-step procedure as previously described
580 in (Méheust et al. 2019). A first protein clustering was done using the fast and sensitive protein
581 sequence searching software MMseqs2 (version
582 9f493f538d28b1412a2d124614e9d6ee27a55f45) (Steinegger and Söding 2017). An all-vs-all
583 search was performed using e-value: 0.001, sensitivity: 7.5 and cover: 0.5. A sequence similarity
584 network was built based on the pairwise similarities and the greedy set cover algorithm from
585 MMseqs2 was performed to define protein subclusters. The resulting subclusters were defined
586 as subfamilies. In order to test for distant homology, we grouped subfamilies into protein families

using an HMM-HMM comparison procedure as follows: the proteins of each subfamily with at least two protein members were aligned using the result2msa parameter of mmseqs2, and, from the multiple sequence alignments, HMM profiles were built using the HHpred suite (version 3.0.3) (Soding 2005). The subfamilies were then compared to each other using hhblits (Remmert et al. 2011) from the HHpred suite (with parameters -v 0 -p 50 -z 4 -Z 32000 -B 0 -b 0). For subfamilies with probability scores of ≥ 95% and coverage ≥ 0.50, a similarity score (probability × coverage) was used as weights of the input network in the final clustering using the Markov Clustering algorithm (Enright, Van Dongen, and Ouzounis 2002), with 2.0 as the inflation parameter. These clusters were defined as the protein families.

Module definition and taxonomic assignment

Looking at the distribution of the protein families across the genomes, a clear modular organization emerged. Modules of families were defined using a cutoff of 0.95 on the dendrogram tree of the families. The dendrogram tree was obtained from a hierarchical clustering using the Jaccard distance that was calculated based on profiles of protein family presence/absence. The corresponding clusters define the modules.

A phyla distribution was assigned to each module using the method of (Méheust et al. 2019). For each module, the median number of genomes per family (m) was calculated. The genomes were ranked by the number of families they carry. The m genomes that carry the most of families were retained; their phyla distribution defines the taxonomic assignment of the module.

Functional annotation

Protein sequences were functionally annotated based on the accession of their best Hmmsearch match (version 3.1) (E-value cut-off 0.001) (Eddy 1998) against an HMM database constructed based on ortholog groups defined by the KEGG (Kanehisa et al. 2016) (downloaded on June 10, 2015). Domains were predicted using the same hmmsearch procedure against the Pfam database (version 31.0) (Punta et al. 2012). The domain architecture of each protein sequence was predicted using the DAMA software (version 1.0) (default parameters) (Bernardes et al. 2016). SIGNALP (version 5.0) (parameters: -format short -org arch) (Almagro Armenteros et al. 2019) was used to predict the putative cellular localization of the proteins. Prediction of transmembrane helices in proteins was performed using TMHMM (version 2.0) (default parameters) (Krogh et al. 2001). Protein sequences were also functionally annotated based on the accession of their best hmmsearch match (version 3.1) (E-value cut-off 1e-10) against the PDB database (Rose et al. 2017) (downloaded in February 2020).

621

622 HMM-HMM Predictions

623 Subfamilies were used to perform HMM-HMM annotation against arCogs of EggNog (version 5.0)
624 (Huerta-Cepas et al. 2019) using hhblits (Remmert et al. 2011) from the HHpred suite (with
625 parameters -v 0 -p 50 -z 4 -Z 32000 -B 0 -b 0). Subfamilies were subsequently functionally
626 annotated based on the EggNog accessions of their best probability score.

627

628 Sequence similarity analysis

629 The 75,737 subfamilies from the 10,866 families were searched against a bacterial database of
630 2,552 bacterial genomes (Méheust et al. 2019) using hmmsearch (version 3.1) (E-value cut-off
631 0.001) (Eddy 1998). Among them 46,261 subfamilies, comprising 8,300 families, have at least
632 one hit against a bacterial genome.

633

634

Data availability

The newly reconstructed genomes have been deposited at NCBI under BioProject **XX** (**TBA**). The genomes of the herein analysed archaea have been made publicly available on the ggkbase database (**TBA**). Detailed annotations of the families are provided in the **Supplementary Dataset - Table S3** accompanying this paper. Raw data files (phylogenetic tree and fasta sequences of the families) are made available via figshare under the following link: **TBA**.

Author contributions

R.M., C.J.C. and J.F.B. designed the study. R.M. and C.J.C. created the dataset. C.J.C performed the phylogenetic analysis. A.L.J. performed the bacterial analysis. R.M. performed the protein family, the module detection, the genome annotations and HMMs analyses. R.M., C.J.C. and J.F.B. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

J.F.B. is a founder of Metagenomi. The authors declare that they have no conflict of interest.

Materials and Correspondence

Correspondence and material requests should be addressed to jbanfield@berkeley.edu and raphael.meheust@berkeley.edu.

Acknowledgements

We thank Dr. Brett Baker, Dr Kiley Seitz and Dr Xianzhe Gong for the permission to use the metagenomic datasets from the Guaymas Basin in this study. We thank Dr. Christine He for the permission to use the metagenomic dataset from Genasci. We acknowledge funding support from the Chan Zuckerberg Biohub and the Innovative Genomics Institute at UC Berkeley.

References

Adam, Panagiotis S., Guillaume Borrel, Céline Brochier-Armanet, and Simonetta Gribaldo. 2017. "The Growing Tree of Archaea: New Perspectives on Their Diversity, Evolution and Ecology." *The ISME Journal* 11 (11): 2407–25.

Akıl, Caner, and Robert C. Robinson. 2018. "Genomes of Asgard Archaea Encode Profilins That Regulate Actin." *Nature* 562 (7727): 439–43.

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. "SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks." *Nature Biotechnology* 37 (4): 420–23.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. "Binning Metagenomic Contigs by Coverage and Composition." *Nature Methods* 11 (11): 1144–46.

Altieri, Amanda S., Jane E. Ladner, Zhuo Li, Howard Robinson, Zahur F. Sallman, John P. Marino, and Zvi Kelman. 2016. "A Small Protein Inhibits Proliferating Cell Nuclear Antigen by Breaking the DNA Clamp." *Nucleic Acids Research* 44 (20): 10015.

Anantharaman, Karthik, Christopher T. Brown, Laura A. Hug, Itai Sharon, Cindy J. Castelle, Alexander J. Probst, Brian C. Thomas, et al. 2016. "Thousands of Microbial Genomes Shed Light on Interconnected Biogeochemical Processes in an Aquifer System." *Nature Communications* 7 (1): 13219.

Aouad, Monique, Najwa Taib, Anne Oudart, Michel Lecocq, Manolo Gouy, and Céline Brochier-Armanet. 2018. "Extreme Halophilic Archaea Derive from Two Distinct Methanogen Class II Lineages." *Molecular Phylogenetics and Evolution* 127 (October): 46–54.

Baker, Brett J., Valerie De Anda, Kiley W. Seitz, Nina Dombrowski, Alyson E. Santoro, and Karen G. Lloyd. 2020. "Diversity, Ecology and Evolution of Archaea." *Nature Microbiology* 5 (7): 887–900.

Bartossek, Rita, Anja Spang, Gerhard Weidler, Anders Lanzen, and Christa Schleper. 2012. "Metagenomic Analysis of Ammonia-Oxidizing Archaea Affiliated with the Soil Group." *Frontiers in Microbiology* 3 (June): 208.

Bernardes, J. S., F. R. J. Vieira, G. Zaverucha, and A. Carbone. 2016. "A Multi-Objective Optimization Approach Accurately Resolves Protein Domain Architectures." *Bioinformatics* 32 (3): 345–53.

Borrel, Guillaume, Nicolas Parisot, Hugh M. B. Harris, Eric Peyretaillade, Nadia Gaci, William Tottey, Olivier Bardot, et al. 2014. "Comparative Genomics Highlights the Unique Biology of Methanomassiliicoccales, a Thermoplasmatales-Related Seventh Order of Methanogenic Archaea That Encodes Pyrrolysine." *BMC Genomics* 15 (August): 679.

Boyd, Joel A., Sean P. Jungbluth, Andy O. Leu, Paul N. Evans, Ben J. Woodcroft, Grayson L. Chadwick, Victoria J. Orphan, Jan P. Amend, Michael S. Rappé, and Gene W. Tyson. 2019. "Divergent Methyl-Coenzyme M Reductase Genes in a Deep-Subseafloor Archaeoglobi." *The ISME Journal*. https://doi.org/10.1038/s41396-018-0343-2.

Brochier-Armanet, Céline, Bastien Boussau, Simonetta Gribaldo, and Patrick Forterre. 2008. "Mesophilic Crenarchaeota: Proposal for a Third Archaeal Phylum, the Thaumarchaeota." *Nature Reviews. Microbiology* 6 (3): 245–52.

Buckel, W. 2001. "Sodium Ion-Translocating Decarboxylases." *Biochimica et Biophysica Acta* 1505 (1): 15–27.

Burke, S. A., S. L. Lo, and J. A. Krzycki. 1998. "Clustered Genes Encoding the Methyltransferases of Methanogenesis from Monomethylamine." *Journal of Bacteriology* 180 (13): 3432–40.

710 Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool
711 for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics*
712 25 (15): 1972–73.
713 Castelle, Cindy J., Christopher T. Brown, Karthik Anantharaman, Alexander J. Probst, Raven H.
714 Huang, and Jillian F. Banfield. 2018. “Biosynthetic Capacity, Metabolic Variety and Unusual
715 Biology in the CPR and DPANN Radiations.” *Nature Reviews. Microbiology* 16 (10): 629–
716 45.
717 Castelle, Cindy J., Kelly C. Wrighton, Brian C. Thomas, Laura A. Hug, Christopher T. Brown,
718 Michael J. Wilkins, Kyle R. Frischkorn, et al. 2015. “Genomic Expansion of Domain
719 Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling.”
720 *Current Biology: CB* 25 (6): 690–701.
721 Cox, Cymon J., Peter G. Foster, Robert P. Hirt, Simon R. Harris, and T. Martin Embley. 2008.
722 “The Archaebacterial Origin of Eukaryotes.” *Proceedings of the National Academy of*
723 *Sciences of the United States of America* 105 (51): 20356–61.
724 DasSarma, Shiladitya, and Priya Arora. 2006. “Genetic Analysis of the Gas Vesicle Gene
725 Cluster in Haloarchaea.” *FEMS Microbiology Letters*. https://doi.org/10.1111/j.1574-
726 6968.1997.tb10456.x.
727 Dimroth, P. 1987. “Sodium Ion Transport Decarboxylases and Other Aspects of Sodium Ion
728 Cycling in Bacteria.” *Microbiological Reviews* 51 (3): 320–40.
729 Dombrowski, Nina, Jun-Hoe Lee, Tom A. Williams, Pierre Offre, and Anja Spang. 2019.
730 “Genomic Diversity, Lifestyles and Evolutionary Origins of DPANN Archaea.” *FEMS*
731 *Microbiology Letters* 366 (2). https://doi.org/10.1093/femsle/fnz008.
732 Eddy, S. R. 1998. “Profile Hidden Markov Models.” *Bioinformatics* 14 (9): 755–63.
733 Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. “An Efficient Algorithm for Large-
734 Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84.
735 Ermler, U., W. Grabarse, S. Shima, M. Goubeaud, and R. K. Thauer. 1997. “Crystal Structure of
736 Methyl-Coenzyme M Reductase: The Key Enzyme of Biological Methane Formation.”
737 *Science* 278 (5342): 1457–62.
738 Evans, Paul N., Joel A. Boyd, Andy O. Leu, Ben J. Woodcroft, Donovan H. Parks, Philip
739 Hugenholtz, and Gene W. Tyson. 2019. “An Evolving View of Methane Metabolism in the
740 Archaea.” *Nature Reviews. Microbiology* 17 (4): 219–32.
741 Evans, Paul N., Donovan H. Parks, Grayson L. Chadwick, Steven J. Robbins, Victoria J.
742 Orphan, Suzanne D. Golding, and Gene W. Tyson. 2015. “Methane Metabolism in the
743 Archaeal Phylum Bathyarchaeota Revealed by Genome-Centric Metagenomics.” *Science*
744 350 (6259): 434–38.
745 Garrett, Roger A., Gisle Vestergaard, and Shiraz A. Shah. 2011. “Archaeal CRISPR-Based
746 Immune Systems: Exchangeable Functional Modules.” *Trends in Microbiology* 19 (11):
747 549–56.
748 Ghalei, Homa, Holger von Moeller, Detlef Eppers, Daniel Sohmen, Daniel N. Wilson, Bernhard
749 Loll, and Markus C. Wahl. 2014. “Entrapment of DNA in an Intersubunit Tunnel System of a
750 Single-Stranded DNA-Binding Protein.” *Nucleic Acids Research*.
751 https://doi.org/10.1093/nar/gku259.
752 Gould, Stephen Jay, and Niles Eldredge. 1977. “Punctuated Equilibria: The Tempo and Mode of
753 Evolution Reconsidered.” *Paleobiology*. https://doi.org/10.1017/s0094837300005224.
754 Guzder, Sami N., Patrick Sung, Véronique Bailly, Louise Prakash, and Satya Prakash. 1994.
755 “RAD25 Is a DMA Helicase Required for DNA Repair and RNA Polymerase II
756 Transcription.” *Nature*. https://doi.org/10.1038/369578a0.
757 Hartman, Hyman, and Alexei Fedorov. 2002. “The Origin of the Eukaryotic Cell: A Genomic
758 Investigation.” *Proceedings of the National Academy of Sciences of the United States of*
759 *America* 99 (3): 1420–25.
760 He, Christine, Ray Keren, Michael Whittaker, Ibrahim F. Farag, Jennifer Doudna, Jamie H. D.

761 Cate, and Jillian Banfield. n.d. “Huge and Variable Diversity of Episymbiotic CPR Bacteria
762 and DPANN Archaea in Groundwater Ecosystems.”
763 https://doi.org/10.1101/2020.05.14.094862.
764 Holzer, Sandro, Jiangyu Yan, Mairi L. Kilkenny, Stephen D. Bell, and Luca Pellegrini. 2017.
765 “Primer Synthesis by a Eukaryotic-like Archaeal Primase Is Independent of Its Fe-S
766 Cluster.” *Nature Communications* 8 (1): 1718.
767 Huang, Rui, Zev A. Ripstein, Rafal Augustyniak, Michal Lazniewski, Krzysztof Ginalski, Lewis E.
768 Kay, and John L. Rubinstein. 2016. “Unfolding the Mechanism of the AAA+ Unfoldase VAT
769 by a Combined Cryo-EM, Solution NMR Study.” *Proceedings of the National Academy of*
770 *Sciences of the United States of America* 113 (29): E4190–99.
771 Hua, Zheng-Shuang, Yan-Ni Qu, Qiyun Zhu, En-Min Zhou, Yan-Ling Qi, Yi-Rui Yin, Yang-Zhi
772 Rao, et al. 2018. “Genomic Inference of the Metabolism and Evolution of the Archaeal
773 Phylum Aigarchaeota.” *Nature Communications*. https://doi.org/10.1038/s41467-018-
774 05284-4.
775 Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K.
776 Forslund, Helen Cook, Daniel R. Mende, et al. 2019. “eggNOG 5.0: A Hierarchical,
777 Functionally and Phylogenetically Annotated Orthology Resource Based on 5090
778 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14.
779 Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016.
780 “KEGG as a Reference Resource for Gene and Protein Annotation.” *Nucleic Acids*
781 *Research* 44 (D1): D457–62.
782 Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an Efficient Tool
783 for Accurately Reconstructing Single Genomes from Complex Microbial Communities.”
784 *PeerJ* 3 (August): e1165.
785 Karst, Søeren M., Rasmus H. Kirkegaard, and Mads Albertsen. n.d. “Mmgenome: A Toolbox for
786 Reproducible Genome Extraction from Metagenomes.” https://doi.org/10.1101/059121.
787 Katoh, Kazutaka, and Daron M. Standley. 2016. “A Simple Method to Control over-Alignment in
788 the MAFFT Multiple Sequence Alignment Program.” *Bioinformatics* 32 (13): 1933–42.
789 Korkhin, Yakov, Ulug M. Unligil, Otis Littlefield, Pamlea J. Nelson, David I. Stuart, Paul B. Sigler,
790 Stephen D. Bell, and Nicola G. A. Abrescia. 2009. “Evolution of Complex RNA
791 Polymerases: The Complete Archaeal RNA Polymerase Structure.” *PLoS Biology* 7 (5):
792 e1000102.
793 Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. “Predicting
794 Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete
795 Genomes.” *Journal of Molecular Biology* 305 (3): 567–80.
796 Kubo, Kyoko, Karen G. Lloyd, Jennifer F. Biddle, Rudolf Amann, Andreas Teske, and Katrin
797 Knittel. 2012. “Archaea of the Miscellaneous Crenarchaeotal Group Are Abundant, Diverse
798 and Widespread in Marine Sediments.” *The ISME Journal*.
799 https://doi.org/10.1038/ismej.2012.37.
800 LaFlamme, Susan E., Shomita Mathew-Steiner, Neetu Singh, Diane Colello-Borges, and
801 Bethsaida Nieves. 2018. “Integrin and Microtubule Crosstalk in the Regulation of Cellular
802 Processes.” *Cellular and Molecular Life Sciences: CMLS* 75 (22): 4177–85.
803 Lecompte, Odile, Raymond Ripp, Jean-Claude Thierry, Dino Moras, and Olivier Poch. 2002.
804 “Comparative Analysis of Ribosomal Proteins in Complete Genomes: An Example of
805 Reductive Evolution at the Domain Scale.” *Nucleic Acids Research* 30 (24): 5382–90.
806 Lin, X., and J. Tang. 1990. “Purification, Characterization, and Gene Cloning of Thermopsin, a
807 Thermostable Acid Protease from Sulfolobus Acidocaldarius.” *The Journal of Biological*
808 *Chemistry* 265 (3): 1490–95.
809 Liu, Bing, Songying Ouyang, Kira S. Makarova, Qiu Xia, Yanping Zhu, Zhimeng Li, Li Guo,
810 Eugene V. Koonin, Zhi-Jie Liu, and Li Huang. 2015. “A Primase Subunit Essential for
811 Efficient Primer Synthesis by an Archaeal Eukaryotic-Type Primase.” *Nature*

Communications 6 (June): 7300.
McKay, Luke J., Mensur Dlakić, Matthew W. Fields, Tom O. Delmont, A. Murat Eren, Zackary J. Jay, Korinne B. Klingelsmith, Douglas B. Rusch, and William P. Inskeep. 2019. "Co-Occurring Genomic Capacity for Anaerobic Methane and Dissimilatory Sulfur Metabolisms Discovered in the Korarchaeota." *Nature Microbiology* 4 (4): 614–22.
Méheust, Raphaël, David Burstein, Cindy J. Castelle, and Jillian F. Banfield. 2019. "The Distinction of CPR Bacteria from Other Bacteria Based on Protein Family Content." *Nature Communications* 10 (1): 4173.
Michiels, Jan, Chuanwu Xi, Jan Verhaert, and Jos Vanderleyden. 2002. "The Functions of Ca(2+) in Bacteria: A Role for EF-Hand Proteins?" *Trends in Microbiology* 10 (2): 87–93.
Miller, M. A., W. Pfeiffer, and T. Schwartz. 2010. "Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees." In *2010 Gateway Computing Environments Workshop (GCE)*, 1–8.
Nelson-Sathi, Shijulal, Tal Dagan, Giddy Landan, Arnold Janssen, Mike Steel, James O. McInerney, Uwe Deppenmeier, and William F. Martin. 2012. "Acquisition of 1,000 Eubacterial Genes Physiologically Transformed a Methanogen at the Origin of Haloarchaea." *Proceedings of the National Academy of Sciences of the United States of America* 109 (50): 20537–42.
Nelson-Sathi, Shijulal, Filipa L. Sousa, Mayo Roettger, Nabor Lozada-Chávez, Thorsten Thiergart, Arnold Janssen, David Bryant, et al. 2015. "Origins of Major Archaeal Clades Correspond to Gene Acquisitions from Bacteria." *Nature* 517 (7532): 77–80.
Niehaus, Thomas D., Mona Elbadawi-Sidhu, Valérie de Crécy-Lagard, Oliver Fiehn, and Andrew D. Hanson. 2017. "Discovery of a Widespread Prokaryotic 5-Oxoprolinase That Was Hiding in Plain Sight." *The Journal of Biological Chemistry* 292 (39): 16360–67.
Offre, Pierre, Anja Spang, and Christa Schleper. 2013. "Archaea in Biogeochemical Cycles." *Annual Review of Microbiology* 67 (June): 437–57.
Olm, Matthew R., Christopher T. Brown, Brandon Brooks, and Jillian F. Banfield. 2017. "dRep: A Tool for Fast and Accurate Genomic Comparisons That Enables Improved Genome Recovery from Metagenomes through de-Replication." *The ISME Journal* 11 (12): 2864–68.
Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55.
Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. "Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life." *Nature Microbiology* 2 (11): 1533–42.
Paytubi, Sonia, Stephen A. McMahon, Shirley Graham, Huanting Liu, Catherine H. Botting, Kira S. Makarova, Eugene V. Koonin, James H. Naismith, and Malcolm F. White. 2012. "Displacement of the Canonical Single-Stranded DNA-Binding Protein in the Thermoproteales." *Proceedings of the National Academy of Sciences of the United States of America* 109 (7): E398–405.
Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. 2012. "IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts174.
Pester, Michael, Christa Schleper, and Michael Wagner. 2011. "The Thaumarchaeota: An Emerging View of Their Phylogeny and Ecophysiology." *Current Opinion in Microbiology*. https://doi.org/10.1016/j.mib.2011.04.007.
Petitjean, Céline, Philippe Deschamps, Purificación López-García, and David Moreira. 2014. "Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota." *Genome Biology and Evolution* 7 (1): 191–204.
Probst, Alexander J., Cindy J. Castelle, Andrea Singh, Christopher T. Brown, Karthik

Anantharaman, Itai Sharon, Laura A. Hug, et al. 2017. “Genomic Resolution of a Cold Subsurface Aquifer Community Provides Metabolic Insights for Novel Microbes Adapted to High CO2concentrations.” *Environmental Microbiology*. https://doi.org/10.1111/1462-2920.13362.

Probst, Alexander J., Thomas Weinmaier, Kasie Raymann, Alexandra Perras, Joanne B. Emerson, Thomas Rattei, Gerhard Wanner, et al. 2014. “Biology of a Widespread Uncultivated Archaeon That Contributes to Carbon Fixation in the Subsurface.” *Nature Communications* 5 (November): 5497.

Punta, Marco, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, et al. 2012. “The Pfam Protein Families Database.” *Nucleic Acids Research* 40 (Database issue): D290–301.

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. 2011. “HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment.” *Nature Methods* 9 (2): 173–75.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, et al. 2013. “Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.” *Nature* 499 (7459): 431–37.

Rose, Peter W., Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R. Bradley, Cole H. Christie, Luigi Di Costanzo, et al. 2017. “The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information.” *Nucleic Acids Research* 45 (D1): D271–81.

Sapra, Rajat, Karine Bagramyan, and Michael W. W. Adams. 2003. “A Simple Energy-Conserving System: Proton Reduction Coupled to Proton Translocation.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (13): 7545–50.

Schut, Gerrit J., Gina L. Lipscomb, Diep M. N. Nguyen, Robert M. Kelly, and Michael W. W. Adams. 2016. “Heterologous Production of an Energy-Conserving Carbon Monoxide Dehydrogenase Complex in the Hyperthermophile Pyrococcus Furiosus.” *Frontiers in Microbiology* 7 (January): 29.

Schut, Gerrit J., William J. Nixon, Gina L. Lipscomb, Robert A. Scott, and Michael W. W. Adams. 2012. “Mutational Analyses of the Enzymes Involved in the Metabolism of Hydrogen by the Hyperthermophilic Archaeon Pyrococcus Furiosus.” *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2012.00163.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. n.d. “Recovery of Genomes from Metagenomes via a Dereplication, Aggregation, and Scoring Strategy.” https://doi.org/10.1101/107789.

Snel, B., P. Bork, and M. A. Huynen. 1999. “Genome Phylogeny Based on Gene Content.” *Nature Genetics* 21 (1): 108–10.

Soding, J. 2005. “Protein Homology Detection by HMM-HMM Comparison.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti125.

Spang, Anja, Eva F. Caceres, and Thijs J. G. Ettema. 2017. “Genomic Exploration of the Diversity, Ecology, and Evolution of the Archaeal Domain of Life.” *Science* 357 (6351). https://doi.org/10.1126/science.aaf3883.

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E. Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs J. G. Ettema. 2015. “Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes.” *Nature* 521 (7551): 173–79.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology* 35 (11): 1026–28.

914 Sung, P., S. N. Guzder, L. Prakash, and S. Prakash. 1996. “Reconstitution of TFIIH and
915 Requirement of Its DNA Helicase Subunits, Rad3 and Rad25, in the Incision Step of
916 Nucleotide Excision Repair.” *The Journal of Biological Chemistry* 271 (18): 10821–26.
917 Teske, Andreas, Dirk de Beer, Luke J. McKay, Margaret K. Tivey, Jennifer F. Biddle, Daniel
918 Hoer, Karen G. Lloyd, et al. 2016. “The Guaymas Basin Hiking Guide to Hydrothermal
919 Mounds, Chimneys, and Microbial Mats: Complex Seafloor Expressions of Subsurface
920 Hydrothermal Circulation.” *Frontiers in Microbiology* 7 (February): 75.
921 Tully, Benjamin J. 2019. “Metabolic Diversity within the Globally Abundant Marine Group II
922 Euryarchaea Offers Insight into Ecological Patterns.” *Nature Communications* 10 (1): 271.
923 Vestergaard, Gisle, Roger A. Garrett, and Shiraz A. Shah. 2014. “CRISPR Adaptive Immune
924 Systems of Archaea.” *RNA Biology*. https://doi.org/10.4161/rna.27990.
925 Wang, Yinzhao, Gunter Wegener, Jialin Hou, Fengping Wang, and Xiang Xiao. 2019.
926 “Expanding Anaerobic Alkane Metabolism in the Domain of Archaea.” *Nature Microbiology*
927 4 (4): 595–602.
928 Woese, C. R., and G. E. Fox. 1977. “Phylogenetic Structure of the Prokaryotic Domain: The
929 Primary Kingdoms.” *Proceedings of the National Academy of Sciences of the United States*
930 *of America* 74 (11): 5088–90.
931 Woese, C. R., R. Gupta, C. M. Hahn, W. Zillig, and J. Tu. 1984. “The Phylogenetic
932 Relationships of Three Sulfur Dependent Archaebacteria.” *Systematic and Applied*
933 *Microbiology* 5: 97–105.
934 Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. “Towards a Natural System of Organisms:
935 Proposal for the Domains Archaea, Bacteria, and Eucarya.” *Proceedings of the National*
936 *Academy of Sciences of the United States of America* 87 (12): 4576–79.
937 Yang, Ying, Rie Yatsunami, Ai Ando, Nobuhiro Miyoko, Toshiaki Fukui, Shinichi Takaichi, and
938 Satoshi Nakamura. 2015. “Complete Biosynthetic Pathway of the C50 Carotenoid
939 Bacterioruberin from Lycopene in the Extremely Halophilic Archaeon Haloarcula Japonica.”
940 *Journal of Bacteriology* 197 (9): 1614–23.
941 Yan, Jiangyu, Thomas R. Beattie, Adriana L. Rojas, Kelly Schermerhorn, Tamzin Gristwood,
942 Jonathan C. Trinidad, Sonja V. Albers, et al. 2017. “Identification and Characterization of a
943 Heterotrimeric Archaeal DNA Polymerase Holoenzyme.” *Nature Communications* 8 (May):
944 15075.
945 Yu, Hongjun, Chang-Hao Wu, Gerrit J. Schut, Dominik K. Haja, Gongpu Zhao, John W. Peters,
946 Michael W. W. Adams, and Huilin Li. 2018. “Structure of an Ancient Respiratory System.”
947 *Cell* 173 (7): 1636–49.e16.
948 Yutin, Natalya, Pere Puigbò, Eugene V. Koonin, and Yuri I. Wolf. 2012. “Phylogenomics of
949 Prokaryotic Ribosomal Proteins.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0036972.
950 Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina
951 Juzokaite, Emmelien Vancaester, Kiley W. Seitz, et al. 2017. “Asgard Archaea Illuminate
952 the Origin of Eukaryotic Cellular Complexity.” *Nature* 541 (7637): 353–58.
953 Zhang, P., J. Wang, and Y. Shi. 2010. “Structure and Mechanism of the S Component of a
954 Bacterial ECF Transporter.” https://doi.org/10.2210/pdb3p5n/pdb.

955