Protein family content uncovers lineage relationships and bacterial pathway maintenance mechanisms in DPANN Archaea

Cindy J. Castelle, Raphaël Meheust, Alexander L. Jaffe, Kiley Seitz, Xianzhe Gong, Brett J. Baker, Jillian F. Banfield

frontiers in Microbiology

ORIGINAL RESEARCH
published: 01 June 2021
doi: 10.3389/fmicb.2021.660052

Protein Family Content Uncovers Lineage Relationships and Bacterial Pathway Maintenance Mechanisms in DPANN Archaea

Cindy J. Castelle[1], Raphaël Méheust[1,2,3], Alexander L. Jaffe[4], Kiley Seitz[5], Xianzhe Gong[5,6], Brett J. Baker[5] and Jillian F. Banfield[1,2,7]**

[1] Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, United States, [2] Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, United States, [3] LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, Evry, France, [4] Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, United States, [5] Department of Marine Science, University of Texas Austin, Port Aransas, TX, United States, [6] Institute of Marine Science and Technology, Shandong University, Qingdao, China, [7] Chan Zuckerberg Biohub, San Francisco, CA, United States

OPEN ACCESS

Edited by:
Rekha Seshadri,
Joint Genome Institute, United States

Reviewed by:
Nikolai Ravin,
Russian Academy of Sciences, Russia
Patrick Forterre,
Institut Pasteur, France

***Correspondence:**
Cindy J. Castelle
cjmtcastelle@gmail.com
Jillian F. Banfield
jbanfield@berkeley.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *28 January 2021*
Accepted: *26 April 2021*
Published: *01 June 2021*

Citation:
Castelle CJ, Méheust R, Jaffe AL, Seitz K, Gong X, Baker BJ and Banfield JF (2021) Protein Family Content Uncovers Lineage Relationships and Bacterial Pathway Maintenance Mechanisms in DPANN Archaea. Front. Microbiol. 12:660052. doi: 10.3389/fmicb.2021.660052

DPANN are small-celled archaea that are generally predicted to be symbionts, and in some cases are known episymbionts of other archaea. As the monophyly of the DPANN remains uncertain, we hypothesized that proteome content could reveal relationships among DPANN lineages, constrain genetic overlap with bacteria, and illustrate how organisms with hybrid bacterial and archaeal protein sets might function. We tested this hypothesis using protein family content that was defined in part using 3,197 genomes including 569 newly reconstructed genomes. Protein family content clearly separates the final set of 390 DPANN genomes from other archaea, paralleling the separation of Candidate Phyla Radiation (CPR) bacteria from all other bacteria. This separation is partly driven by hypothetical proteins, some of which may be symbiosis-related. Pacearchaeota with the most limited predicted metabolic capacities have Form II/III and III-like Rubisco, suggesting metabolisms based on scavenged nucleotides. Intriguingly, the Pacearchaeota and Woesearchaeota with the smallest genomes also tend to encode large extracellular murein-like lytic transglycosylase domain proteins that may bind and degrade components of bacterial cell walls, indicating that some might be episymbionts of bacteria. The pathway for biosynthesis of bacterial isoprenoids is widespread in Woesearchaeota genomes and is encoded in proximity to genes involved in bacterial fatty acids synthesis. Surprisingly, in some DPANN genomes we identified a pathway for synthesis of queuosine, an unusual nucleotide in tRNAs of bacteria. Other bacterial systems are predicted to be involved in protein refolding. For example, many DPANN have the complete bacterial DnaK-DnaJ-GrpE system and many Woesearchaeota and Pacearchaeota possess bacterial group I chaperones. Thus, many DPANN appear to have mechanisms to ensure efficient protein folding of both archaeal and laterally acquired bacterial proteins.

Keywords: DPANN, archaea, protein family, bacterial genes, phylogeny

INTRODUCTION

The first insights into archaeal cells of very small size and limited biosynthetic gene inventories predictive of symbiotic lifestyles were provided by researchers studying co-cultures of thermophilic *Ignicoccus* archaea and associated *Nanoarchaeum equitans* (Huber et al., 2002). Transmission electron microscope images established that these nanoarchaea are episymbionts (i.e., they attach to the surfaces of host *Ignicoccus* cells) (Jahn et al., 2008). Archaea with very small cell sizes and small genomes were discovered in acid mine drainage biofilms (Baker et al., 2006, 2010), apparently episymbiotically associated with Thermoplasmatales archaea (Comolli et al., 2009). Halophilic nanoarchaea were found in archaea-dominated planktonic communities (Narasingarao et al., 2012). Over time, a series of papers investigating the microbiology of sediments and aquatic environments greatly expanded the taxonomic diversity of nanoarchaea (Rinke et al., 2013; Castelle et al., 2015; Dombrowski et al., 2020). One group, groundwater-associated Huberarchaeota, are predicted to be symbionts of the archaeal Altiarchaeota (SM1) based on highly correlated abundance patterns (Probst et al., 2018). Recent work has further illuminated the biology, diversity and distribution of these organisms (Ortiz-Alvarez and Casamayor, 2016; Wurch et al., 2016; Krause et al., 2017; Dombrowski et al., 2019; Hamm et al., 2019), now often referred to as the DPANN (Rinke et al., 2013). Genomic analyses were used to propose that lateral gene transfer, including from bacteria, has shaped the inventories of acidophilic nanoarchaea (Baker et al., 2010). This idea has been reinforced by more recent analyses of other DPANN archaea (Dombrowski et al., 2020; Jaffe et al., 2020).

The DPANN may comprise at least 12 different putative phylum-level lineages, but the phylogeny is still unresolved. For example, it is uncertain whether long branches within these groups are due to rapid evolution vs. undersampling. It also remains unclear which lineages are in vs. outside of the DPANN (Castelle et al., 2018) and whether lineages of archaea with small genomes form a monophyletic radiation (Aouad et al., 2018) analogous to the Candidate Phyla Radiation (CPR) of Bacteria or comprise multiple radiations. Additionally, while it has been suggested that Altiarchaeota may be part of the DPANN (Adam et al., 2017; Spang et al., 2017) there are currently no high quality genomes available for Altiarchaeota and the genomic sampling of this group is insufficient to resolve their placement. If Altiarchaeota are confirmed to be part of the DPANN, the metabolic characteristics of DPANN archaea would be expanded substantially (Probst et al., 2014).

Recently, researchers used HMM-based protein family-based analyses to provide insights regarding the relationships within and among major bacterial lineages, including those that comprise the bacterial CPR (Méheust et al., 2019). This approach has the advantage of being annotation-independent, thus can include consideration of proteins without functional predictions (Méheust et al., 2019). Here, we used a similar approach to investigate DPANN in the context of Domain Archaea. This analysis drew upon new genomes that we reconstructed for undersampled lineages, making use of metagenomic datasets from sediments and water. Our analyses of inventories of predicted DPANN proteins uncovered an intriguing number of bacteria-like sequences for functions typically associated with bacteria. The question of how proteome maintenance is achieved for organisms with a mixture of bacterial and archaeal systems was addressed in part by identification of bacterial-type protein refolding complexes in many DPANN genomes.

RESULTS

DPANN Genome Collection

We collected 2,618 archaeal genomes [262 DPANN genomes based on the Genome Taxonomy DataBase system (Chaumeil et al., 2019)] from the NCBI genome database (**Supplementary Table 1**) and augmented these by reconstructing 569 new DPANN draft genomes 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. The 3,197 genomes were clustered at ≥95% average nucleotide identity (ANI) to generate 1,749 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 section "Materials and Methods"). We required that these proteins were co-encoded on a single scaffold to avoid contamination due to mis-binning of assembled genome fragments. Our analyses were performed on a final set that includes 390 draft-quality, non-redundant DPANN genomes with an average completeness of 82% (**Supplementary Table 1**).

We assessed the taxonomic distribution and affiliation of the 332 (348 if the genomes from Parks et al., 2017 are included) newly reconstructed representative genomes based on the alignment of the syntenic block of 14 ribosomal proteins that is commonly used to infer genome-based phylogenies in previous studies (Castelle et al., 2015; **Figure 1**, and also see **Supplementary Figure 1** for more details). These genomes were classified as DPANN because they cluster phylogenetically with known DPANN lineages or are most closely related to them (**Figure 1**). Addition of these new genome sequences increased the taxon sampling of the DPANN radiation and helped to resolve clades that were clustered together due to undersampling. The "Ca Mamarchaeota" (Castelle and Banfield, 2018), now comprises 12 genome sequences from various ecosystem types and locations and is clearly monophyletic. We distinguished six new potential phylum-level DPANN lineages, the full definition of which will require further genomic sampling (**Figure 1**). The first is related to the Micrarchaeota phylum and is based on three genomes from groundwater collected at two different locations (Genasci and Rifle, United States). Two others, based on one and three genomes respectively, are deep branching with the Nanohaloarchaeota and were sampled from deep-subsurface marine sediment ecosystems. A fourth clade clustering at the root of the Nanoarchaeota and Parvarchaeota phyla includes seven genomes from marine and CO_2-saturated groundwater (Crystal Geyser, UT, United States). For this apparently phylum-level lineage we proposed the name of Candidatus "Kleinarchaeota,"

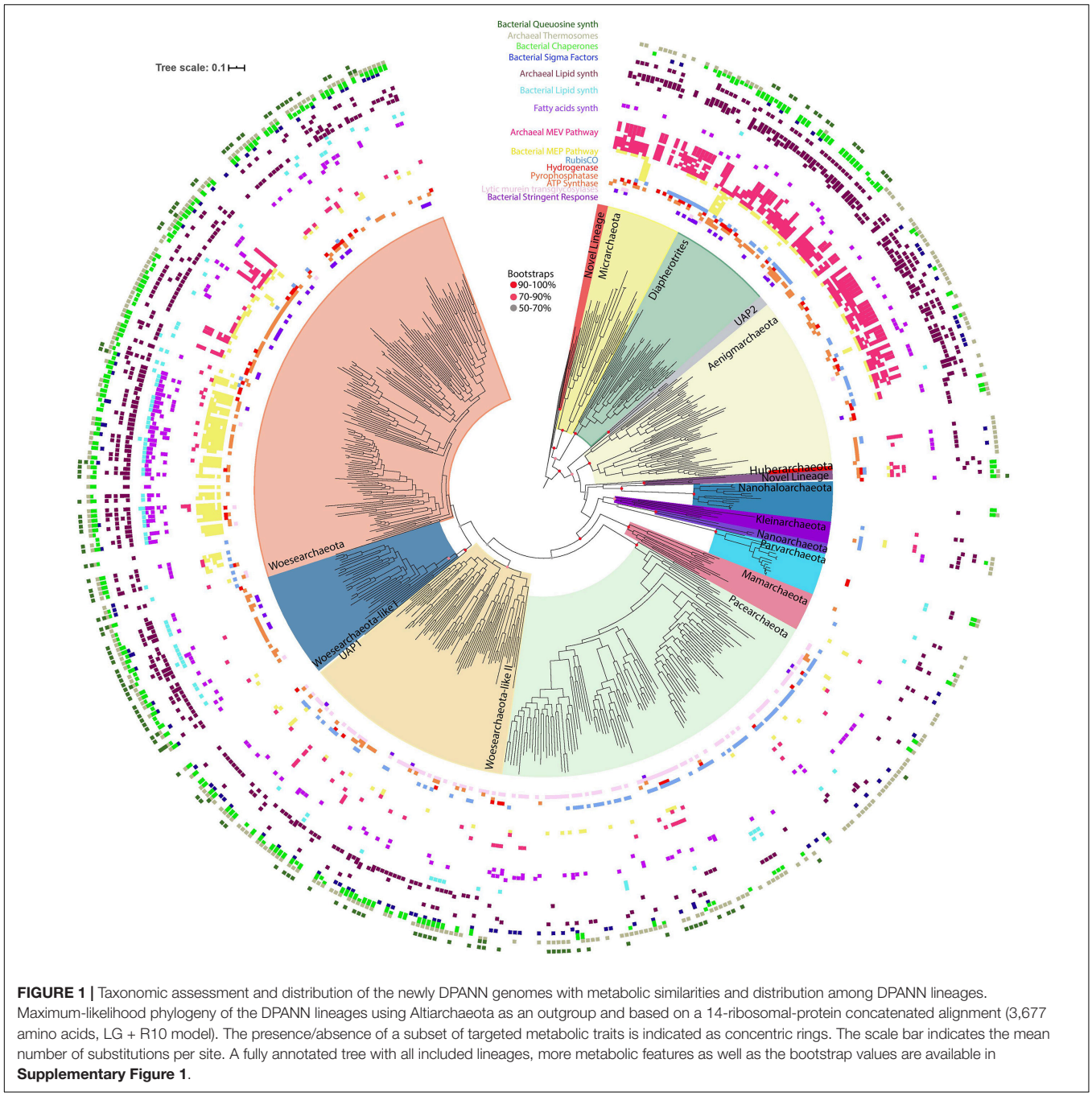

FIGURE 1 | Taxonomic assessment and distribution of the newly DPANN genomes with metabolic similarities and distribution among DPANN lineages. Maximum-likelihood phylogeny of the DPANN lineages using Altiarchaeota as an outgroup and based on a 14-ribosomal-protein concatenated alignment (3,677 amino acids, LG + R10 model). The presence/absence of a subset of targeted metabolic traits is indicated as concentric rings. The scale bar indicates the mean number of substitutions per site. A fully annotated tree with all included lineages, more metabolic features as well as the bootstrap values are available in **Supplementary Figure 1**.

named for the late Dr. Daniel Klein, a scientist who worked at the Joint Bioenergy Institute (**Figure 1**). Fifty new genome sequences further resolved two new clades within the Woesearchaeota phylum. The first clade corresponds to the Woesarchaeota-like II whereas the second clade groups together the Woesearchaeota and Woesarchaeota-like I groups.

Protein Clustering Separates DPANN From Other Archaea

We conducted an analysis of protein families, making use of a newly available set of 10,866 protein families for archaea (Méheust et al., 2020). Analogous to the CPR bacteria, which appear to be the bacterial counterpart to the DPANN based on metabolic limitations, DPANN archaea separate from all other archaea based on their protein family contents (**Figure 2**) (Méheust et al., 2019). This is probably in part due to patterns of gene loss and their consistently minimal metabolic platforms.

The metabolic platforms differ from one DPANN lineage to another. Each module, a block of co-occurring protein families containing at least 20 families, was assigned a taxonomic distribution based on the taxonomy of the genomes with the highest number of families (see section "Materials and Methods" and **Supplementary Table 2**). Overall, 17 modules occurred in at

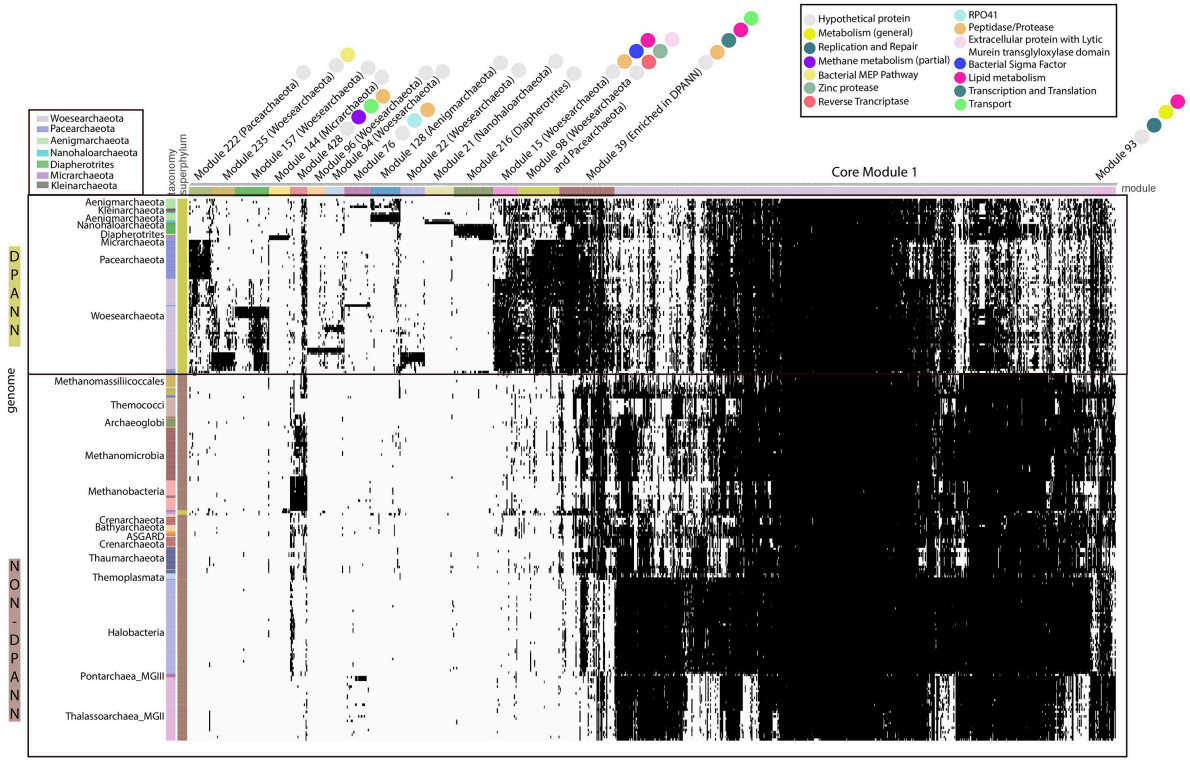

FIGURE 2 | Distribution of 17 modules (1143 Protein families) in DPANN genomes. The distribution of 1143 protein families (columns) in representative genomes (rows) from Archaea. Data are clustered based on the presence (black) and absence (white) profiles (Jaccard distance, complete linkage). Pathway or function for each protein family is indicated by a colored circle.

least one DPANN genome (1,143 protein families), 14 modules are more common in DPANN compared to non-DPANN genomes (504 families), and 11 among the 14 modules are taxonomically assigned to a single DPANN lineage (**Table 1** and **Figure 2**). Most of the protein families from those 14 modules have very poor or no functional annotations (**Supplementary Table 3**). Importantly, we identified no modules that are common to all DPANN lineages but absent from other Archaea, although the module 39 is widespread in DPANN and only present in a few non-DPANN genomes (**Figure 2** and **Supplementary Table 2**). Evolutionary reconstructions of each gene family show that families from lineages-specific modules most likely originated around the time of lineage divergence (**Supplementary Figure 2**), and thus are largely consistent with the presence/absence distributions of the families from the modules described in **Table 1**.

Interestingly, one module (Module 98) is assigned to the Pacearchaeota and Woesearchaeota and the 51 families of the module 98 are widespread in these groups whereas mostly absent in other lineages (see section "Materials and Methods" and **Supplementary Table 3**). While most of the protein families in this module have unknown function, one protein family including proteins up to 3,000 aa in length (fam00969) contains a murein-like lytic transglycosylase domain potentially involved in cell defense and/or cell-cell interaction. This was previously noted in just three DPANN genomes (Rinke et al., 2013; Castelle et al., 2015). Murein transglycosylases are near-ubiquitous in Bacteria (but normally absent in Archaea) and specifically bind and degrade murein (peptidoglycan) strands, the main components of the bacterial cell walls. Thus, these large proteins might be a novel class of extracellular enzymes specific to Woesearchaeota and Pacearchaeota and involved in attachment to bacterial cell walls, potentially including those of their hosts.

Overall, the vast majority of the DPANN lineage-specific protein families lack function predictions. For instance, the Pacearchaeota and Aenigmarchaeota each had lineage-specific modules (Modules 222 and 128, respectively) completely lacking functional predictions (**Table 1**). Despite the lack of functional annotations, many of the DPANN lineage-specific protein families are predicted to be either membrane-bound or extracellular (**Table 1** and **Supplementary Table 3**).

Bacterial Systems in DPANN Archaea

To determine the extent to which inter-domain lateral transfer has impacted metabolic capacity in DPANN archaea, we computed a metric ("breadth") that describes the incidence of each DPANN protein family across domain Bacteria (**Figure 3**) (section "Materials and Methods"). Most of those with the highest incidence across Bacteria (as measured by "breadth") are involved in core biological functions that are common to both bacteria and archaea (module 1) (**Supplementary Table 3** and see section "Materials and Methods"). On the other hand, we also observed

TABLE 1 | List of the eleven modules that are assigned to a single DPANN lineage and the module 98 that is taxonomically assigned to both the Woesearchaeota and the Pacearchaeota within the genomes of DPANN archaea.

Modules	Lineage(s)	# Families	SignalP (%)	TMHMM (%)	Hypothetical families (%)	Hits to Bacteria (%)
15,22,94,96,157,235	Woesearchaeota	178	19	48	93	15
98	Woesearchaeota + Pacearchaeota	51	24	53	90	39
216	Diapherotrites	48	25	73	98	2
128	Aenigmarchaeota	37	35	65	100	8
21	Nanohaloarchaeota	36	17	44	97	6
222	Pacearchaeota	28	36	79	100	18
144	Micrarchaeota	26	19	69	96	4

It is a gradient color from 0 (white) to 100 (green).

a number of smaller archaeal protein families that were relatively common in Bacteria (**Figure 3**) and these bacterial-like sequences were more prominent in DPANN compared to other archaea. We reasoned that these uncommon archaeal protein families could either reflect contamination introduced by misbinning or inter-domain lateral gene transfer from Bacteria to Archaea. Thus, we retained them for further analysis. Interestingly, only a few DPANN protein families with unknown or uncharacterized function(s) were detected in bacteria. This could suggest that the specificity of the majority of families of hypothetical proteins are unique to DPANN Archaea; however, we cannot fully exclude the possibility that homology was not detected due to high divergence.

We then further examined small archaeal protein families with a strong signal of bacterial origin ("breadth"), focusing on those with confident functional annotations. Some bacterial systems involved in transcription, protein folding, and cell membrane production are encoded in many DPANN archaeal genomes. For example, in the newly reconstructed genomes we uncovered genes for sigma factors classified as σ-70 (fam00359, **Supplementary Table 3**). These are particularly prevalent in Woesearchaeota (Module 15). Their presence due to mis-binning was ruled out based on the phylogeny of co-encoded and taxonomically informative proteins. We did not identify any possible anti-sigma factors, which are essential for controlling and regulating sigma factors in Bacteria. Thus, if involved in transcription regulation (along with the expected archaeal transcription factors TBP, TFB, TFE that are also found in DPANN), the mechanism controlling them remains unclear.

Other bacterial systems identified in DPANN genomes are the GroEL and Hsp60 group I chaperones involved in protein folding. Group II chaperones consist of the archaeal thermosomes and related eukaryotic cytosolic variants (CCT or TRiC) (Lopez et al., 2015) and generally do not coexist with Group I chaperones in the same cell. Analysis of DPANN genomes revealed that many Woesearchaeota and Pacearchaeota possess both archaeal thermosomes and GroEL/ES (fam00693 and fam02048, respectively). A similar observation was made for some species of mesophilic archaea (methanogens of the order of Methanosarcina) (Figueiredo et al., 2004). The presence of both systems may ensure efficient protein folding of both traditionally archaeal and laterally acquired bacterial proteins.

Another well-known chaperone system in bacteria is the DnaK-DnaJ-GrpE (KJE) system, which funnels its clients toward the native state or ushers misfolded proteins into degradation. Within Domain Archaea, the KJE system only has been reported in a few archaeal species and was likely acquired by horizontal gene transfer (Gribaldo et al., 1999; Petitjean et al., 2012). The KJE system has been considered an essential building block for a minimal bacterial genome and has been reported to be mostly absent from highly reduced endosymbionts or obligate bacterial pathogens (McCutcheon and Moran, 2011). Surprisingly, we find that many DPANN archaea have the complete bacterial KJE system (Module 1: DnaK:fam00458; DnaJ: fam00163 and fam00567; GrpE: fam01146, **Supplementary Table 3**).

Previously, we reported the distribution of archaeal and bacterial isoprenoid synthesis pathways [the archaeal mevalonate (MVA) and the bacterial 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate (MEP/DOXP) pathways] in many DPANN genomes (Castelle and Banfield, 2018). Inclusion of hundreds of new DPANN genomes in this study allowed us to further evaluate the distribution of the MEP pathway. Interestingly, we found this pathway is widespread in a monophyletic clade composed of 52 Woesearchaeota genomes (**Figure 1**) that were reconstructed from a wide range of terrestrial and marine ecosystems. The genes involved in the biosynthesis of bacterial isoprenoids appear to be co-located in Woesearchaeota genomes (**Supplementary Table 4**). In proximity to the MEP pathway genes we identified a gene encoding for an UbiA prenyltransferase (fam00115), which is involved in the synthesis of archaeal ether lipids, and transfers prenyl groups to hydrophobic ring structures such as quinones, hemes, chlorophylls, vitamin E, or shikonin (Villanueva et al., 2014). Some of these archaea have genes involved in bacterial fatty acid synthesis, including the bacterial glycerol-3-Phosphate dehydrogenase, and in some genomes they are co-located with MEP pathway genes (G3P; fam00008; **Supplementary Table 4**).

One of the most common predictions for the metabolic basis for growth of DPANN archaea is fermentation (Castelle et al., 2015, 2018). For instance, many DPANN archaea are capable of producing acetate, primarily via a single enzyme, an acetyl-CoA synthetase that is the most common enzyme for this function found in Archaea. On the other hand, most Bacteria use a pathway involving acetate kinase (Ack) and phosphotransacetylase (Pta), which is the reversible reaction

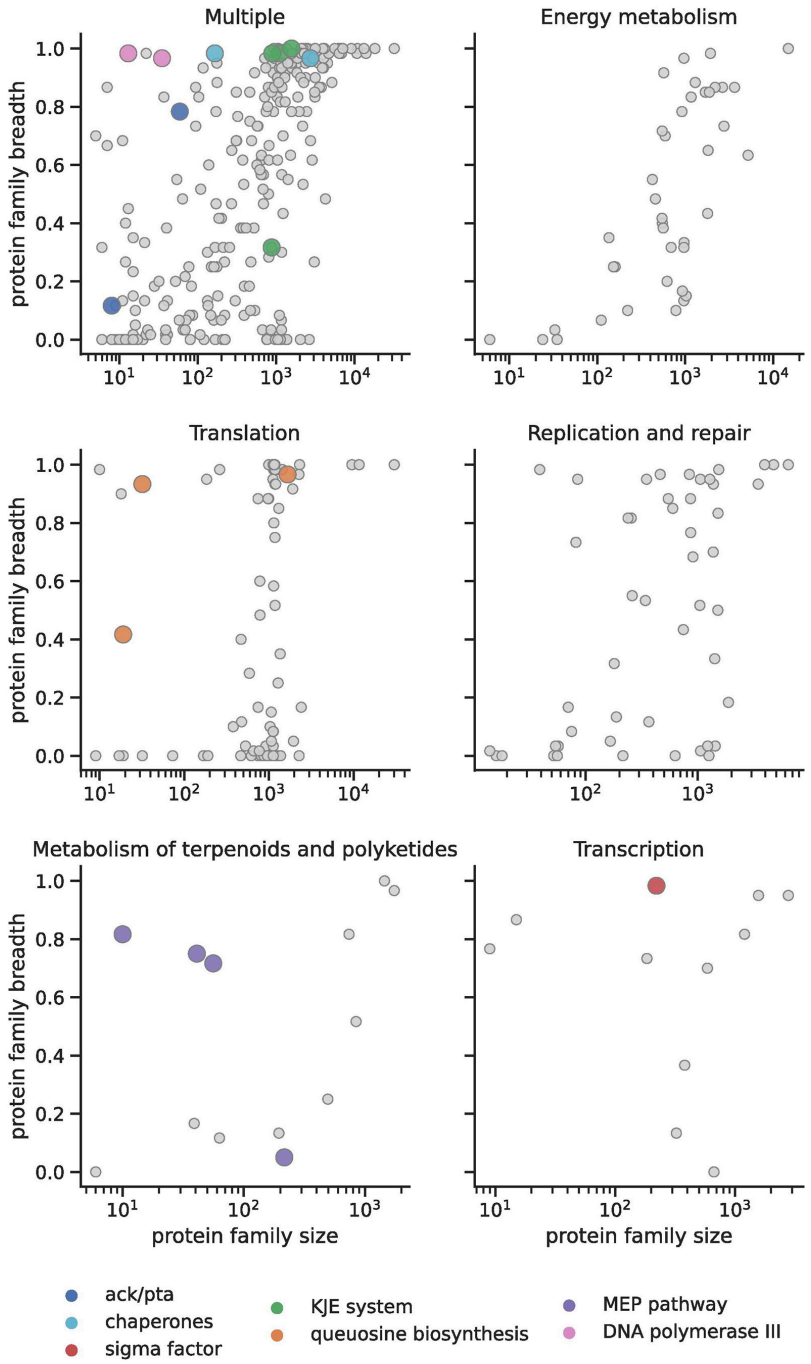

FIGURE 3 | Bacterial protein families found in DPANN genomes. In the top left plot, "multiple" refers to protein families that are found in more than one metabolic pathway. Colored dots indicate those instances in DPANN. For example, translation genes in the DPANN (pink dots) are genes involved in queosine biosynthesis (see key).

allowing Bacteria to oxidize or produce acetate. The only Archaea currently known to use the Ack/Pta pathway are members of the methanogenic Methanosarcina (Fournier and Gogarten, 2008). Here, we report that the Ack/Pta pathway occurs in many Pacearchaeota and Woesearchaeota (fam06325 and fam12605, respectively) (**Supplementary Table 3**). Thus, these DPANN archaea have the capacity to both excrete and assimilate acetate, as occurs in some bacteria.

Transfer RNA (tRNA) is structurally unique among nucleic acids in harboring an astonishing diversity of post-transcriptionally modified nucleosides. Two of the most radically modified nucleosides known to occur in tRNA are

queuosine and archaeosine, both of which are characterized by a 7-deazaguanosine core structure (Itaya, 2003). In spite of the phylogenetic segregation observed for these nucleosides (queuosine is present in Eukarya and Bacteria, while archaeosine is present only in Archaea), their structural similarity suggested a common biosynthetic origin. Surprisingly, we identified the complete and/or partial bacterial queuosine biosynthesis pathway in 34 archaeal genomes (including 12 DPANN genomes). In DPANN, this queuosine synthesis pathway includes co-localized (**Figure 4** and **Supplementary Table 5**) QueA (S-adenosylmethionine:tRNA ribosyltransferase-isomerase; fam24423), QueH-like (epoxyqueuosine reductase; fam24901), and Tgt (queuine tRNA-ribosyltransferase; Fam00366).

Some DPANN genomes encode DNA polymerase III (fam03494), a bacterial polymerase not typically found in Archaea. Three out of the 35 DNA polymerase III sequences found in fam03494 were verified as part of DPANN genomes based on the presence of archaeal ribosomal proteins and other phylogenetically informative genes in close genomic proximity (**Supplementary Table 6**). Interestingly, the genes that encode the DNA polymerase III from the two Pacearchaeota are located next to genes encoding the small and the large subunits of the DNA polymerase II. Several genes near the DNA polymerase III gene encode phage protein suggesting the DNA polymerase genes may have been acquired by horizontal gene transfers. Placing the three DNA polymerase III sequences in phylogenetic context reveals that some sequences from Aenigmarchaeota cluster with sequences from Chloroflexi, possibly reflecting lateral transfer from this group (**Figure 5**). Additionally, the two Pacearchaeota sequences are highly divergent from bacterial sequences, yet the domain structure supports their classification as DNA polymerase III (**Supplementary Table 6**).

CONCLUSION

Much remains to be learned about DPANN archaea, and the question of whether they share a common ancestor remains unclear. We found no set of DPANN-specific protein families, as was observed for the CPR bacteria. However, some modules of protein families are shared by multiple DPANN distinct lineages. For example, Woesearchaeota and Pacearchaeota share distinct protein family modules, and although they are clearly separate phylogenetic lineages and have distinct metabolism capacities. Thus, these lineages, at least, could have shared a common ancestor from which they inherited a set of lineage specific proteins.

Based on the protein family analysis, DPANN archaea are distinct from other archaea. This may be in part due to patterns of gene loss as well as lateral gene transfer. Similar patterns of gene loss on the path to symbioses could have arisen via convergent evolution or could indicate that phylogenetically distinct lineages have shared ancestry. Our findings related to lateral gene transfer extend those of a very recent report that suggested the phylogenetic analyses of DPANN archaea may be confounded by this process (Dombrowski et al., 2020). However, our results suggest that sets of proteins, including those with known and unknown functions, were established at the origins of the major lineages and have been largely preserved within each phylum.

The question of why DPANN archaea have acquired bacterial proteins is intriguing, and may indicate close physical associations between DPANN and bacteria. It remains to be seen whether, in some cases, DPANN have bacterial hosts. The presence in DPANN proteomes of typically bacterial systems, including proteins potentially involved in genome regulation and protein refolding, may explain in part how these hybrid protein inventories are maintained. The findings hint at the existence of mechanisms that enable lifestyles based on close associations that involve organisms from the bacterial and archaeal domains. Furthermore, the presence of bacterial systems even in reduced DPANN genomes is interesting, considering that large-scale bacterial gene import has been suggested to underlie the origin of major archaeal lineages (Nelson-Sathi et al., 2015).

MATERIALS AND METHODS

Genome Collection

Five hundred sixty-nine unpublished genomes were added to the 2,618 genomes of Archaea downloaded from the NCBI genome database in September 2018 (**Supplementary Table 1**).

One hundred thirty-two genomes were obtained from metagenomes of sediment samples. Sediment samples were collected from the Guaymas Basin (27°N0.388, 111°W24.560, Gulf of California, Mexico) during three cruises at a depth of approximately 2,000 m below the water surface. Sediment cores were collected during two Alvin dives, 4,486 and 4,573 in 2008 and 2009. Sites referred to as "Megamat" (genomes starting with "Meg") and "Aceto Balsamico" (genomes starting with "AB" in name), Core sections between 0 and 18 cm from 4,486 and from 0–33 cm 4,573 and were processed for these analyses. Intact sediment cores were subsampled under N_2 gas, and immediately frozen at −80°C on board. The background of sampling sites was described previously (Teske et al., 2016). Samples were processed for DNA isolation from using the MoBio PowerMax soil kit (Qiagen) following the manufacturer's protocol. Illumina library preparation and sequencing were performed using Hiseq 4000 at Michigan State University. Paired-end reads were interleaved using interleav_fasta.py[1] and the interleaved sequences were trimmed using Sickle[2] with the default settings. Metagenomic reads from each subsample were individually assembled using IDBA-UD with the following parameters: –pre_correction –mink 65 –maxk 115 –step 10 –seed_kmer 55 (Peng et al., 2012). Metagenomic binning was performed on contigs with a minimum length of 2,000 bp in individual assemblies using the binning tools MetaBAT (Kang et al., 2015) and CONCOCT (Alneberg et al., 2014), and resulting bins were combined with using DAS Tool (Sieber et al., 2017). CheckM lineage_wf (v1.0.5) (Parks et al., 2015) was used to estimate the percentage of completeness and contamination of bins. Genomes with more

[1]https://github.com/jorvis/biocode/blob/master/fasta/interleave_fasta.py
[2]https://github.com/najoshi/sickle

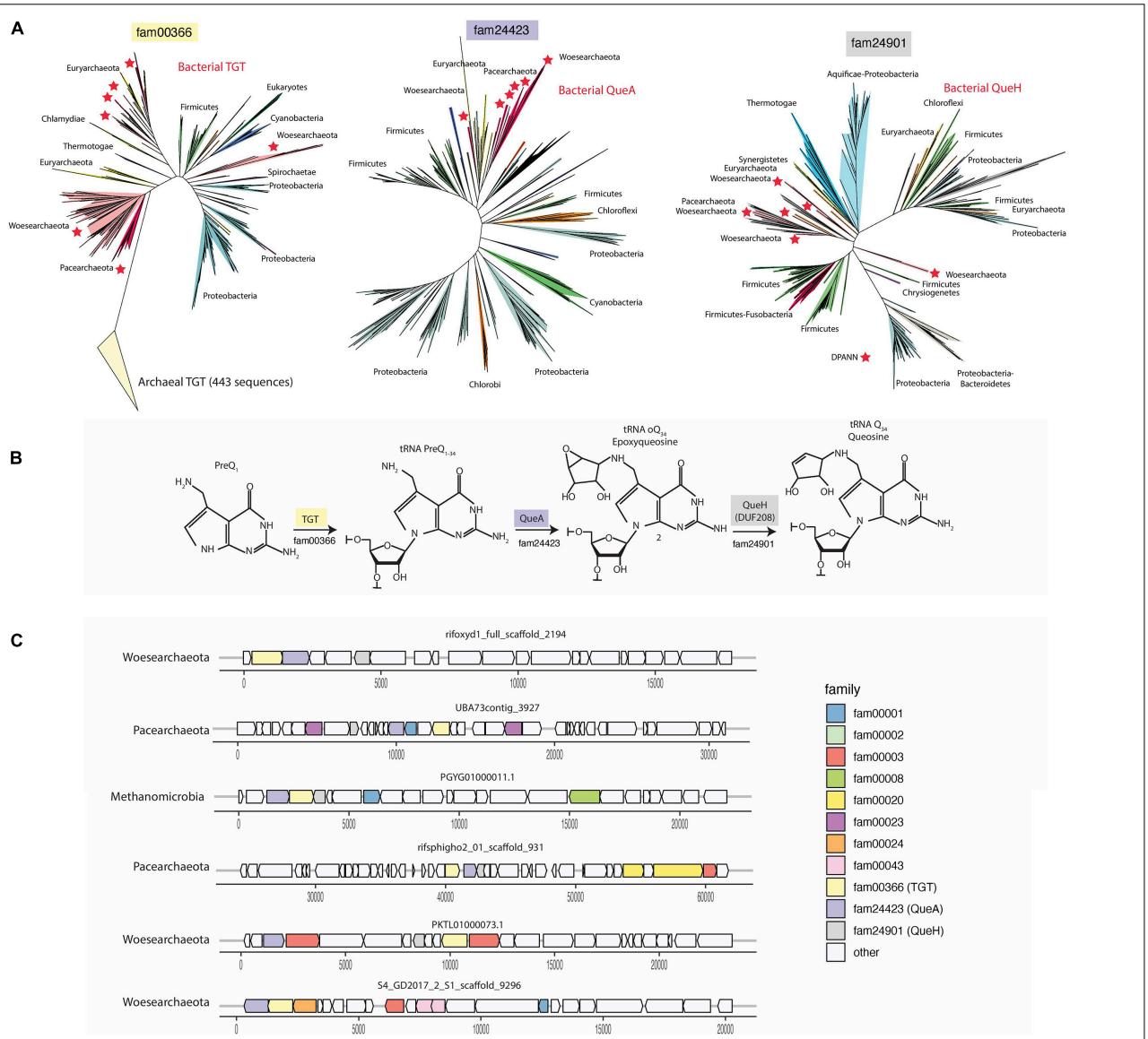

FIGURE 4 | Bacterial queuosine pathway in DPANN genomes and some Euryarchaeota (Methanosarcinales). **(A)** Phylogenetic affiliation and distribution of the identified QueA, QueH and TGT in archaea selected for this study. **(B)** The three last steps of the queuosine biosynthesis pathway in Bacteria. **(C)** Selected examples of genomic context and gene synteny around the genes QueA, QueH (DUF208) and TGT. QueA, S-adenosylmethionine:tRNA ribosyltransferase-isomerase; QueH (DUF208), epoxyqueuosine reductase; TGT, queuine tRNA-ribosyltransferase. A full gene synteny and genomic context of all the genes involved in the queuosine pathway in this study is available in **Supplementary Table 5**.

than 50% completeness and 10% contamination were manually optimized based on GC content, sequence depth and coverage using mmgenome (Karst et al., 2016).

One hundred eighty-eight genomes were obtained from groundwater samples from Genasci Dairy Farm, located in Modesto, CA, United States (He et al., 2021). Over 400 L of groundwater were filtered from monitoring well 5 on Genasci Dairy Farm over a period ranging from March 2017 to June 2018. DNA was extracted from all filters using Qiagen DNeasy PowerMax Soil kits and ∼10 Gbp of 150 bp, paired end Illumina reads were obtained for each filter. Assembly was performed using MEGAHIT (Li et al., 2015) with default parameters, and the scaffolding function from assembler IDBA-UD was used to scaffold contigs. Scaffolds were binned on the basis of GC content, coverage, presence of ribosomal proteins, presence of single copy genes, tetranucleotide frequency, and patterns of coverage across samples. Bins were obtained using manual binning on ggKbase (Wrighton et al., 2012), Maxbin2 (Wu et al., 2016), CONCOCT (Alneberg et al., 2014), Abawaca1, and Abawaca2,[3] with DAS Tool (Sieber et al., 2017) used to choose the best set of bins from all programs. All bins were manually checked to remove incorrectly assigned scaffolds using ggKbase.

[3]https://github.com/CK7/abawaca

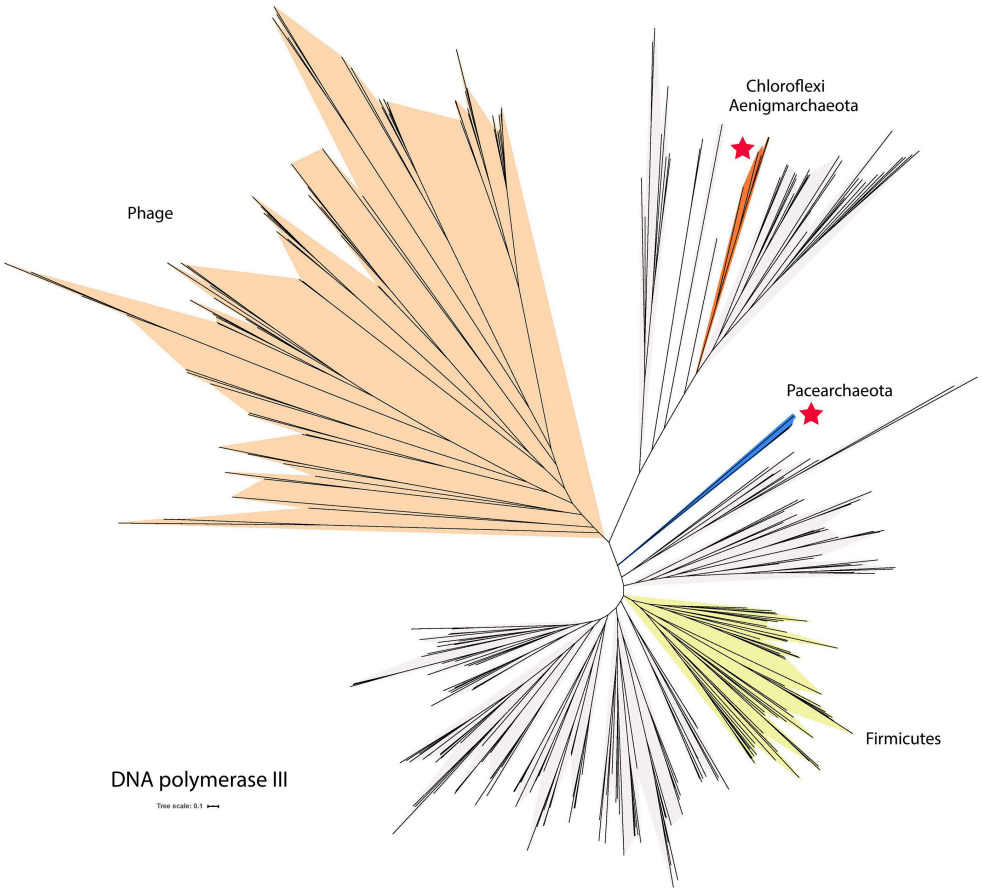

FIGURE 5 | Phylogenetic tree of DNA polymerase III, a bacterial polymerase not typically found in Archaea. The tree includes sequences from phages and some bacteria (light cream shading). Asterisks indicate groups of sequences found in DPANN genomes. The scale bar indicates the mean number of substitutions per site.

Additionally, 168 genomes were obtained from an aquifer adjacent to the Colorado River near the town of Rifle, CO, United States, at the Rifle Integrated Field Research Challenge (IFRC) site (Anantharaman et al., 2016). Sediment samples were collected from the "RBG" field experiment carried out in 2007. Groundwater samples were collected from three different field experiments. All groundwater samples were collected from 5 m below the ground surface by serial filtration onto 1.2, 0.2, and 0.1 μm filters (Supor disc filters; Pall Corporation, Port Washington, NY, United States). DNA was extracted from all frozen filters using the PowerSoil DNA Isolation kit (MoBio Laboratories Inc., Carlsbad, CA, United States) and 150 bp paired end Illumina reads with a targeted insert size of 500 bp were obtained for each filter. Assemblies were performed using IDBA-UD (Peng et al., 2012) with the following parameters: –mink 40, –maxk 100, –step 20, –min_contig 500. All resulting scaffolds were clustered into genome bins using multiple algorithms. First, scaffolds were binned on the basis of % GC content, differential coverage abundance patterns across all samples using Abawaca1, and taxonomic affiliation. Scaffolds that did not associate with any cluster using this method were binned based on tetranucleotide frequency using Emergent Self-Organizing Maps (ESOM; Dick et al., 2009). All genomic bins were manually inspected within ggKbase.

Fifty genomes were obtained from the Crystal Geyser system in UT, United States (Probst et al., 2018). Microbial size filtration from Crystal Geyser fluids was performed using two different sampling systems. One system involved sequential filtration of aquifer fluids on 3-, 0. 8-, 0. 2-, and 0.1-μm filters (polyethersulfone, Pall 561 Corporation, NY, United States). The second system was designed to filter high volumes of water sequentially onto 2. 5-, 0. 65-, 0. 2-, and 0.1-μm filters (ZTECG, Graver Technologies, Glasgow, KY, United States). Metagenomic DNA was extracted from the filters using MoBio PowerMax soil kit. DNA was subjected to 150 bp paired end illumina HiSeq sequencing at the Joint Genome Institute. Assembly of high-quality reads was performed using IDBA_UD with standard parameters and genes of assembled scaffolds (>1 kb). Genome bins were obtained using different binning algorithms: Semi-automated tetranucleotide-frequency based ESOMs, differential coverage ESOMs, Abawaca1, MetaBAT, and Maxbin2. Best genomes from each sample were selected using DASTool. All bins were manually checked to remove incorrectly assigned scaffolds using ggKbase.

Finally, 41 genomes were obtained from the Uncultivated Bacteria and Archaea project (Parks et al., 2017), four genomes from Homestake Mine, SD, United States (Rinke et al., 2013) and three genomes from Lake Tyrrell, Australia (Narasingarao et al., 2012) but were manually curated using ggKbase.

Genome Completeness Assessment and De-Replication

Genome completeness and contamination were estimated based on the presence of single-copy genes (SCGs). Genome completeness was estimated using 38 SCGs (Anantharaman et al., 2016). DPANN genomes with more than 22 SCGs and less than 4 duplicated copies of the SCGs were considered as draft-quality genomes. Genomes were de-replicated using dRep (Olm et al., 2017) (version v2.0.5 with ANI > 95%). The most complete and less contaminated genome per cluster was used in downstream analyses. Completeness and Contamination based on CheckM (version 1.1.2) (Parks et al., 2015) were also performed to allow comparison with our custom made set of 38 SCGs.

Concatenated 14 Ribosomal Proteins Phylogeny

A 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). Homologous protein sequences were aligned using MAFFT (version 7.390) (–auto option) (Katoh and Standley, 2016). The protein alignments were concatenated and manually refined, with a final alignment of 551 genomes and 3,677 positions. Phylogenetic tree was inferred using RAxML (version 8.2.10) (Stamatakis, 2014) [as implemented on the CIPRES web server (Miller et al., 2010)], under the LG + R10 model of evolution, and with the number of bootstraps automatically determined.

Taxonomic Assignment of the Genomes

Taxonomy assignment was performed for each genome using the GTDB-Tk software (v1.3.0) (default parameters) (Chaumeil et al., 2019) on the Genome Taxonomy DataBase (GTDB) database (release 05-RS95) (Parks et al., 2020).

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).

Sequence Similarity Analysis

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

To calculate the "breadth" metric, describing the incidence of archaeal protein families across domain Bacteria, we first found the percentage of representative genomes in each bacterial phylum with an above-threshold hit to each family. For those families with at least one hit among all bacterial reference genomes, we then secondarily calculated the percentage of unique bacterial phyla in which at least one third of representative genomes encoded a sequence hit. Finally, the breadth metric was compared with protein family size, taking into consideration KEGG categories derived from the functional annotation step described above, and used in part to identify systems of potential bacterial origin in DPANN (**Figure 3** in main text).

ALE Analysis

To examine protein family evolution in the DPANN, we performed an automated gene-species tree reconciliation workflow adapted from Sheridan et al. (2020). For each family, sequences were aligned with MAFFT (version 7.390) (–auto option) (Katoh and Standley, 2016). The alignment was further trimmed using Trimal (version 1.4.22) (–gappyout option) (Capella-Gutiérrez et al., 2009) were used to infer maximum-likelihood phylogenetic trees using IQ-TREE with 1000 ultrafast bootstrap replicates (-bnni -m TEST -st AA -bb 1000 -nt AUTO) (version 1.6.6) (Nguyen et al., 2015), using ModelFinder (Kalyaanamoorthy et al., 2017) to select the best model of evolution, and with 1000 ultrafast bootstrap (Hoang et al., 2018). A random sample of 100 bootstrap replicates were then used to probabilistically reconcile each protein family with the inferred species tree using the ALE package (ALE_undated) (Szöllősi et al., 2013). Estimates of missing gene fraction were derived from the genome completeness estimates described above. We

then calculated the total number of originations (horizontal gene transfer from non-DPANN, or *de novo* gene formation) and intra-DPANN horizontal transfers over each branch of the species tree and mapped branch-wise counts for each event to a species-tree cladogram in iTol (Letunic and Bork, 2016).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI BioProject, PRJNA288027 and PRJNA692327.

AUTHOR CONTRIBUTIONS

CC, RM and JB designed the study. KS, XG, and BB collected the samples from the Guaymas Basin and assembled the genomes. CC and RM created the genome dataset. RM performed the protein family, the module detection, and the genome annotations. CC performed the phylogenetic and functional analyses. AJ performed the bacterial analysis. AJ and RM performed the ALE analysis. CC, RM, AJ, and JB wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was supported by the Chan Zuckerberg Biohub to JB and the Innovative Genomics Institute at UC Berkeley. XG acknowledges funding support from the Fundamental Research Funds of Shandong University (Grant no. 2018HW011). This work was also supported by funding to BB from the Simons Foundation (Award 687165).

ACKNOWLEDGMENTS

We thank Dr. Christine He for the permission to use the metagenomic dataset from Genasci dairy farm.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.660052/full#supplementary-material

Supplementary Figure 1 | Taxonomic assessment and distribution of the newly DPANN genomes with metabolic similarities and distribution among DPANN lineages. Maximum-likelihood phylogeny of the DPANN lineages using Altiarchaeota as an outgroup and based on a 14-ribosomal-protein concatenated alignment (3677 amino acids, LG+R10 model). The presence of targeted metabolic traits is indicated with colored boxes. The scale bar indicates the mean number of substitutions per site. Bootstrap values are also indicated at the branches.

Supplementary Figure 2 | Evolution of protein families in the DPANN archaea. Each panel displays the species tree of the 390 genomes used for the protein families analysis in cladogram format. The size and hue of circles represent the cumulative number of originations (orange) - defined as either lateral transfer from outside the lineages examined here, or de novo evolution - and intra-DPANN transfers (blue) predicted to occur on that branch. Evolutionary reconstructions are shown for gene families in all/widespread modules (panel ab) and lineage-specific modules detailed in **Table 1** (panels c-i).

Supplementary Table 1 | List of the 3,197 genomes used in this study. For each genome (column A), its NCBI accession, GGKBASE link, number of scaffolds, genome size and number of CDS are displayed in columns B, C, D, E and F respectively. Genome source is in column G, dRep cluster in column H. Genome completeness and the contamination based on single copy genes are displayed in columns I and J respectively. Column K informs about the concatenated ribosomal proteins. The 1,179 representative genomes are indicated in column L. The phylum and superphylum (DPANN and non-DPANN) taxonomy of the representative genomes are provided in columns M and N. Taxonomy based on the different databases we pulled out the genomes is shown in column O. Taxonomy based on the GTDN system is shown in column P. CheckM gene set, completeness and contamination are provided in columns Q, R and S.

Supplementary Table 2 | Taxonomy distribution of the 17 modules. Module name is indicated in column A whereas the number of families is indicated in column B. Suggested taxonomic distribution is indicated in column C. Column D details the genomes used to define the taxonomic distribution (phylum, number of genomes).

Supplementary Table 3 | Annotation of the 1,143 families detected in the 17 modules. Column A: module number. Column B: family accession. Column C: number of proteins in the family. Column D: median length of the proteins. Column E: ratio of proteins predicted to contain a signal peptide. Column F: median number of predicted transmembrane helix per protein. Column G: domain architecture reported by Pfam. Columns H, I, J, K, L: KEGG annotations. Column M: Cazy annotation. Columns N to AC indicate the ratio of genomes having the given family in the given archaeal phylum. Columns AD to CK indicate the ratio of genomes having the given family in the given bacterial phylum.

Supplementary Table 4 | Genes neighboring the MEP pathway. The fifteen genes downstream and upstream of each fam03888 gene (column I) were identified and annotated using the protein clustering (column F), the PFAM (column H) and the KEGG databases (column G).

Supplementary Table 5 | Genes neighboring the three genes encoding the enzymes of the queuosine biosynthesis pathway. The seven genes downstream and upstream of each QueA (Sadenosylmethionine: tRNA ribosyltransferase-isomerase; fam24423), QueH-like (epoxyqueuosine reductase; fam24901) and Tgt (queuine tRNA-ribosyltransferase; Fam00366) gene (column H) were identified and annotated using the protein clustering (column E), the PFAM (column G), and the KEGG databases (column F).

Supplementary Table 6 | Genes neighboring the three genes encoding the bacterial DNA polymerase III. The 15 genes downstream and upstream of each DNA polymerase III gene (column H) were identified and annotated using the protein clustering (column E), the PFAM (column G) and the KEGG databases (column F). The taxonomy of the five best hits on the GTDB are provided in columns I, J, K, L, and M.

REFERENCES

Adam, P. S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* 11, 2407–2425. doi: 10.1038/ismej.2017.122

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide

predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7:13219.

Aouad, M., Taib, N., Oudart, A., Lecocq, M., Gouy, M., and Brochier-Armanet, C. (2018). Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* 127, 46–54. doi: 10.1016/j.ympev.2018.04.011

Baker, B. J., Comolli, L. R., Dick, G. J., Hauser, L. J., Hyatt, D., Dill, B. D., et al. (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl. Acad. Sci. U. S. A.* 107, 8806–8811. doi: 10.1073/pnas.0914470107

Baker, B. J., Tyson, G. W., Webb, R. I., Flanagan, J., Hugenholtz, P., Allen, E. E., et al. (2006). Lineages of acidophilic archaea revealed by community genomic analysis. *Science* 314, 1933–1935. doi: 10.1126/science.1132690

Bernardes, J. S., Vieira, F. R. J., Zaverucha, G., and Carbone, A. (2016). A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics* 32, 345–353. doi: 10.1093/bioinformatics/btv582

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Castelle, C. J., and Banfield, J. F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* 172, 1181–1197. doi: 10.1016/j.cell.2018.02.016

Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., and Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* 16, 629–645. doi: 10.1038/s41579-018-0076-2

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., et al. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* 25, 690–701. doi: 10.1016/j.cub.2015.01.014

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Comolli, L. R., Baker, B. J., Downing, K. H., Siegerist, C. E., and Banfield, J. F. (2009). Three-dimensional analysis of the structure and ecology of a novel, ultra-small archaeon. *ISME J.* 3, 159–167. doi: 10.1038/ismej.2008.99

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10:R85.

Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P., and Spang, A. (2019). Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* 366, 1–12. doi: 10.1093/femsle/fnz008

Dombrowski, N., Williams, T. A., Sun, J., Woodcroft, B. J., Lee, J.-H., Minh, B. Q., et al. (2020). Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* 11:3939. doi: 10.1038/s41467-020-17408-w

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Figueiredo, L., Klunker, D., Ang, D., Naylor, D. J., Kerner, M. J., Georgopoulos, C., et al. (2004). Functional characterization of an archaeal GroEL/GroES chaperonin system: significance of substrate encapsulation. *J. Biol. Chem.* 279, 1090–1099. doi: 10.1074/jbc.m310914200

Fournier, G. P., and Gogarten, J. P. (2008). Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. *J. Bacteriol.* 190, 1124–1127. doi: 10.1128/jb.01382-07

Gribaldo, S., Lumia, V., Creti, R., Conway de Macario, E., Sanangelantoni, A., and Cammarano, P. (1999). Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. *J. Bacteriol.* 181, 434–443. doi: 10.1128/jb.181.2.434-443.1999

Hamm, J. N., Erdmann, S., Eloe-Fadrosh, E. A., Angeloni, A., Zhong, L., Brownlee, C., et al. (2019). Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. U. S. A.* 116, 14661–14670. doi: 10.1073/pnas.1905179116

He, C., Keren, R., Whittaker, M. L., Farag, I. F., Doudna, J. A., Cate, J. H. D., et al. (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* 6, 354–365. doi: 10.1038/s41564-020-00840-5

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., and Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417, 63–67. doi: 10.1038/417063a

Itaya, T. (2003). Studies on the Syntheses of the Hypermodified Nucleosides of Phenylalanine Transfer Ribonucleic Acids. *ChemInform* 34:200338247. doi: 10.1002/chin.200338247

Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S., and Banfield, J. F. (2020). The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biol.* 18:69. doi: 10.1186/s12915-020-00804-5

Jahn, U., Gallenberger, M., Paper, W., Junglas, B., Eisenreich, W., Stetter, K. O., et al. (2008). Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. *J. Bacteriol.* 190, 1743–1750. doi: 10.1128/jb.01731-07

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462.

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. doi: 10.7717/peerj.1165

Karst, S. M., Kirkegaard, R. H., and Albertsen, M. (2016). mmgenome: a toolbox for reproducible genome extraction from metagenomes. *Biorxiv* doi: 10.1101/059121

Katoh, K., and Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 32, 1933–1942. doi: 10.1093/bioinformatics/btw108

Krause, S., Bremges, A., Münch, P. C., McHardy, A. C., and Gescher, J. (2017). Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Sci. Rep.* 7:3289.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–5.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Lopez, T., Dalton, K., and Frydman, J. (2015). The Mechanism and Function of Group II Chaperonins. *J. Mol. Biol.* 427, 2919–2930. doi: 10.1016/j.jmb.2015.04.013

McCutcheon, J. P., and Moran, N. A. (2011). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26. doi: 10.1038/nrmicro2670

Méheust, R., Burstein, D., Castelle, C. J., and Banfield, J. F. (2019). The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* 10:4173.

Méheust, R., Castelle, C. J., Jaffe, A. L., and Banfield, J. F. (2020). Early acquisition of conserved, lineage-specific proteins currently lacking functional predictions were central to the rise and diversification of archaea. *Biorxiv* doi: 10.1101/2020.07.16.207365

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES Science Gateway for inference of large phylogenetic trees," In *proceedings of the Gateway Computing Environments Workshop (GCE)*. US: IEEE.

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., et al. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* 6, 81–93. doi: 10.1038/ismej.2011.78

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., et al. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517, 77–80. doi: 10.1038/nature13805

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi: 10.1038/ismej.2017.126

Ortiz-Alvarez, R., and Casamayor, E. O. (2016). High occurrence of Pacearchaeota and Woesearchaeota (Archaea superphylum DPANN) in the surface waters of oligotrophic high-altitude lakes. *Environ. Microbiol. Rep.* 8, 210–217. doi: 10.1111/1758-2229.12370

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38, 1079–1086. doi: 10.1038/s41587-020-0501-8

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542. doi: 10.1038/s41564-017-0012-7

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Petitjean, C., Moreira, D., López-García, P., and Brochier-Armanet, C. (2012). Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea. *BMC Evol. Biol.* 12:226. doi: 10.1186/1471-2148-12-226

Probst, A. J., Ladd, B., Jarett, J. K., Geller-McGrath, D. E., Sieber, C. M. K., Emerson, J. B., et al. (2018). Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat. Microbiol.* 3, 328–336. doi: 10.1038/s41564-017-0098-y

Probst, A. J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J. B., Rattei, T., et al. (2014). Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun.* 5:5497.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Rose, P. W., Prliæ, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* 45, D271–D281.

Sheridan, P. O., Raguideau, S., Quince, C., Holden, J., Zhang, L., Consortium, T., et al. (2020). Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. *Nat. Commun.* 11:5494.

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2017). Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. *BioRxiv* doi: 10.1101/107789

Spang, A., Caceres, E. F., and Ettema, T. J. G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* 357:eaaf3883. doi: 10.1126/science.aaf3883

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. *Syst. Biol.* 62, 901–912. doi: 10.1093/sysbio/syt054

Teske, A., de Beer, D., McKay, L. J., Tivey, M. K., Biddle, J. F., Hoer, D., et al. (2016). The Guaymas Basin Hiking Guide to Hydrothermal Mounds, Chimneys, and Microbial Mats: complex Seafloor Expressions of Subsurface Hydrothermal Circulation. *Front. Microbiol.* 7:75. doi: 10.3389/fmicb.2016.00075

Villanueva, L., Sinninghe Damsté, J. S., and Schouten, S. (2014). A re-evaluation of the archaeal membrane lipid biosynthetic pathway. *Nat. Rev. Microbiol.* 12, 438–448. doi: 10.1038/nrmicro3260

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., et al. (2012). Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science* 337, 1661–1665. doi: 10.1126/science.1224041

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Wurch, L., Giannone, R. J., Belisle, B. S., Swift, C., Utturkar, S., Hettich, R. L., et al. (2016). Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* 7:12115.

Conflict of Interest: JB co-founded the company Metagenomi, which seeks to discover genome editing tools to address medical challenges; thus, there is no commercial or financial relationship to this work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.