Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations

Cindy J. Castelle [1,2], Christopher T. Brown [1], Karthik Anantharaman [1,5], Alexander J. Probst [1,6], Raven H. Huang[3] and Jillian F. Banfield [1,2,4]*

[1]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA. [2]Chan Zuckerberg Biohub, San Francisco, CA, USA. [3]Department of Biochemistry, University of Illinois, Urbana-Champaign, IL, USA. [4]Department of Environmental Science, Policy, and Management, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Department of Bacteriology, University of WisconsinMadison, Madison, WI, USA. [6]Department of Chemistry, Biofilm Center, Group for Aquatic Microbial Ecology, University of Duisburg-Essen, Essen, Germany.

*e-mail: jbanfield@berkeley.edu

Abstract

Candidate phyla radiation (CPR) bacteria and DPANN (an acronym of the names of the first included phyla) archaea are massive radiations of organisms that are widely distributed across Earth's environments, yet we know little about them. Initial indications are that they are consistently distinct from essentially all other bacteria and archaea owing to their small cell and genome sizes, limited metabolic capacities and often episymbiotic associations with other bacteria and archaea. In this Analysis, we investigate their biology and variations in metabolic capacities by analysis of approximately 1,000 genomes reconstructed from several metagenomics-based studies. We find that they are not monolithic in terms of metabolism but rather harbour a diversity of capacities consistent with a range of lifestyles and degrees of dependence on other organisms. Notably, however, certain CPR and DPANN groups seem to have exceedingly minimal biosynthetic capacities, whereas others could potentially be free living. Understanding of these microorganisms is important from the perspective of evolutionary studies and because their interactions with other organisms are likely to shape natural microbiome function.

Introduction

Recently recognized radiations in both domains Bacteria and Archaea, the candidate phyla radiation (CPR) and DPANN (an acronym of the names of the first included phyla, '*Candidatus* Diapherotrites', '*Candidatus* Parvarchaeota', '*Candidatus* Aenigmarchaeota', Nanoarchaeota and '*Candidatus* Nanohaloarchaeota'), respectively, are a remarkable aspect of the tree of life[1,2,3,4,5,6,7]. The CPR seems to be a monophyletic radiation that represents a substantial fraction of the bacterial domain (Fig. 1a; Supplementary Table 1), with at least 74 potentially phylum-level lineages (Supplementary Table 1). The DPANN radiation, which includes the

previously known *Nanoarchaeum equitans*[8] and the ARMAN archaea (‘*Candidatus* Micrarchaeota’ and ‘*Ca*. Parvarchaeota’[9,10]), was recently proposed to form a monophyletic candidate superphylum composed of five phyla (‘*Ca*. Diapherotrites’, ‘*Ca*. Parvarchaeota’, ‘*Ca*. Aenigmarchaeota’, Nanoarchaeota and ‘*Ca*. Nanohaloarchaeota’[1]). Later, the DPANN superphylum was expanded by addition of genomes from seven putative new phylum-level groups[2,5,7,11,12] (Fig. 1b; Supplementary Table 1). Although some studies support the monophyly of the DPANN superphylum[1,2,13], it is important to acknowledge that the monophyly of this group was not recovered by several phylogenomic analyses, so its status as a superphylum remains uncertain[14,15,16].

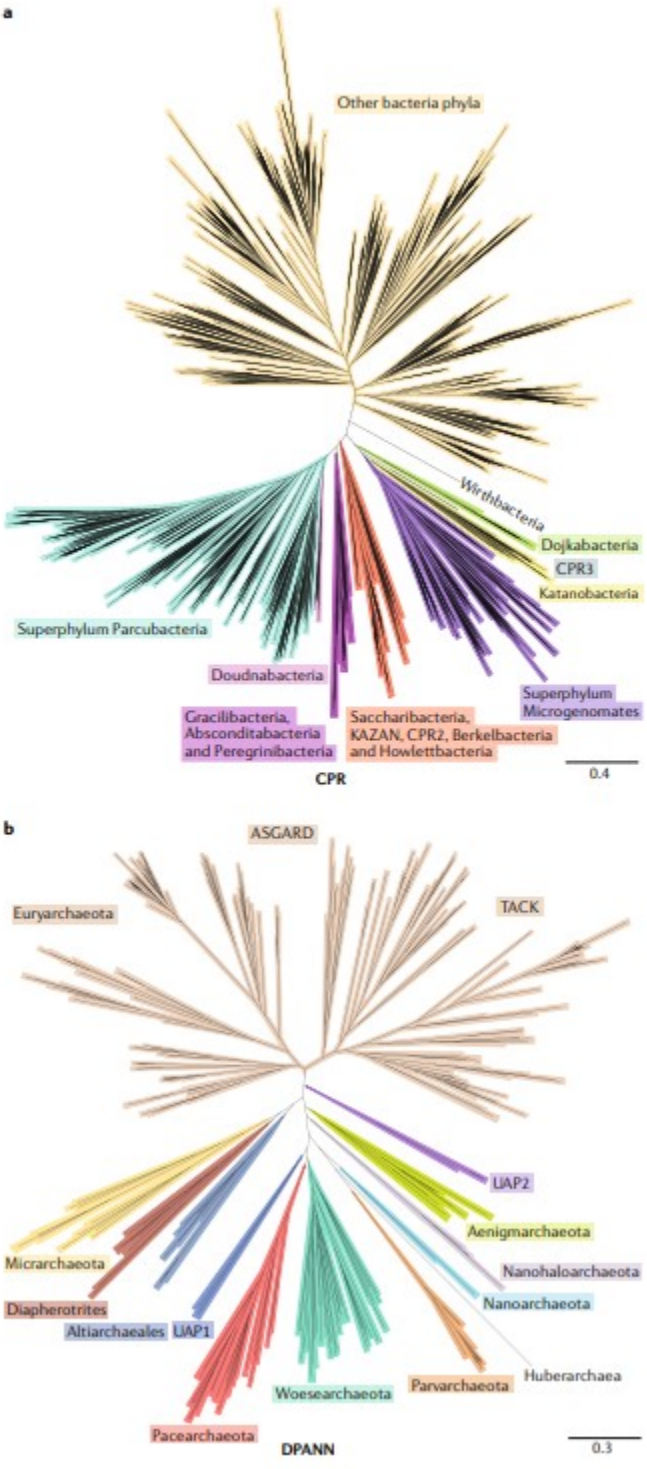

Fig. 1: Phylogenetic trees of Bacteria and Archaea.

The phylogenetic tree of Bacteria is shown in part a, and that of Archaea is shown in part b. Candidate phyla radiation (CPR) and DPANN (‘*Candidatus* Diapherotrites’, ‘*Candidatus*Parvarchaeota’, ‘*Candidatus* Aenigmarchaeota’, Nanoarchaeota and ‘*Candidatus*Nanohaloarchaeota’, in addition to other included lineages) superphyla or phyla are indicated. The maximum likelihood trees (RAxML with PROTCATLG model)

are based on concatenation of 16 ribosomal proteins for Bacteria (ribosomal proteins L2, L3, L4, L5, L6, L14, L15, L18, L22, L24, S3, S8, S17, S19, L16 and S10) and 14 ribosomal proteins for Archaea (as for Bacteria but without L16 and S10), as described previously[4] (Supplementary Methods). The trees in Newick format with full bootstrap values are in Supplementary Data 1and Supplementary Data 2, respectively. The tree in part a is based on the sequences used in ref.4.

We used genome and 16S ribosomal RNA (rRNA) gene sequence analyses to examine the ecological distribution patterns of CPR and DPANN organisms and found them in many environments, including acidic[9,17], alkaline[18] and hypersaline habitats[19,20], freshwater[1,21,22], terrestrial[1,3,7,12,23,24,25] and marine[1,7,26,27,28,29] ecosystems, and animal[7,30,31,32,33,34] microbiomes (Supplementary Table 2). Overall, CPR and DPANN organisms are mostly detected in oxygen-limited or anaerobic environments. Although many CPR organisms seem to be dominant in untreated groundwater, the members of the Parcubacteria candidate superphylum can be selectively enriched in treated water[35]regardless of treatment and disinfectant type[36] and thus can persist in drinking water[37]. Both CPR and DPANN sometimes occur within the human microbiome (Supplementary Table 2). For example, '*Candidatus*Saccharibacteria' (TM7)[38,39] and '*Candidatus* Absconditabacteria' (SR1)[38](both CPR) have been found in the mouth. '*Ca*. Saccharibacteria' and '*Candidatus* Parcubacteria' (OD1), also CPR, occur in the gut[40] and may be implicated in bowel disease[41]. The cell-free DNA of '*Ca*. Parcubacteria' has been detected in blood[42], and cells of DPANN archaea have been detected in lung fluids[34]. Various CPR and DPANN organisms were recently reported in the dolphin mouth[43], including '*Candidatus* Gracilibacteria' (BD1-5), but CPR and DPANN seem to be at relatively low abundance in the surface oceans. When considering the global distribution of CPR and DPANN organisms, it is important to note that primer mismatches in 16S rRNA gene surveys can hinder detection of some phyla within these radiations[3,9,10,44,45] (Supplementary Figure 1and Supplementary Table 3).

Published genome analyses indicate that the CPR and DPANN organisms[1,2,3,7,8,9,11,45,46,47,48,49,50] have small genomes, small cell sizes and notable gaps in core metabolic potential, consistent with a symbiotic lifestyle[2,3,6,46,48,49]. This fascinating pattern, involving massive groups of organisms in both the bacterial and archaeal domains, is not yet well understood. In this Analysis, we extend prior studies, leveraging published genomes[2,3,5,6,7,11,12,19] (Supplementary Table 4) to look broadly across the radiations to investigate the extent of similarity in their metabolic potential, focusing in particular on metabolic gaps and unexpected biological features that are unusual outside of these groups. This undertaking is timely because now there are a sufficient number of DPANN and CPR genomes on which to base such an analysis. Use of almost 1,000 near-complete metagenome-assembled genomes reduced the probability that major findings were incorrect owing to missing information. For each category of metabolism discussed below, we stress that the deductions are

based on analysis of predicted proteins, as experimental validation of metagenome-derived sequences has been accomplished in only a few cases to date. Throughout, the possibility remains that apparently missing functions are present but divergent or that the function occurs via an unknown mechanism. However, when functions are performed by easily recognized, well-characterized proteins or occur in related organisms, we conclude that they are likely absent if the genes are not identified. Among many intriguing biological findings, we discovered introns within the tRNA genes of CPR. We identified phylogenetically distinct subgroups within the CPR and DPANN that, remarkably, lack essentially all biosynthetic pathways and ATP synthase and often have diversity-generating mechanisms that may be required to maintain host associations. Other groups may be capable of independent growth. Overall, the CPR and DPANN display similar patterns of biosynthetic gaps as well as diversity in metabolic platforms.

Small genomes and small cell sizes

A common feature of CPR and DPANN is their small genome size (Fig. 2), an observation that motivates the question of whether they are cellular organisms. CPR and DPANN genomes encode genetic systems for cell division (for example, via FtsZ-based mechanisms), indicating that they are cellular life forms as opposed to DNA resident inside of host cells. These systems are not found in some symbionts with very reduced genomes[51]. The cells are small on the basis of the strong enrichment of CPR and DPANN by groundwater filtration[2,3,52] and cryo-transmission electron microscopy images of enrichments of CPR cells that passed through a 0.2 μm filter[39,52]. Small cell size has also been established for specific Nanoarchaeota[53,54,55,56], Nanohaloarchaeota[20] and ARMAN archaea[10,57] and is characteristic of another group of CPR bacteria when grown under some conditions[39,58]. Although these tiny organisms were previously detected in many environmental studies on the basis of 16S rRNA gene amplification of cells retained on 0.2 μm filters, some seem to be specifically enriched on 0.1 μm filters. For instance, '*Candidatus* Microgenomates' (OP11), '*Ca*. Parcubacteria' that are missing ribosomal protein L1 (OD1-L1)[3] and '*Candidatus* Katanobacteria' (WWE3) are more likely to be detected in small-cell filtrates, whereas the '*Candidatus* Peregrinibacteria' (PER) and other '*Ca*. Parcubacteria' are more often found on the 0.2 μm filter, which is probably indicative of larger cell sizes. Measurements of replication rates[59] and images showing cell division[57] indicate that the cells are metabolically active, at least under some conditions, and thus that small cell size is not simply a consequence of starvation.

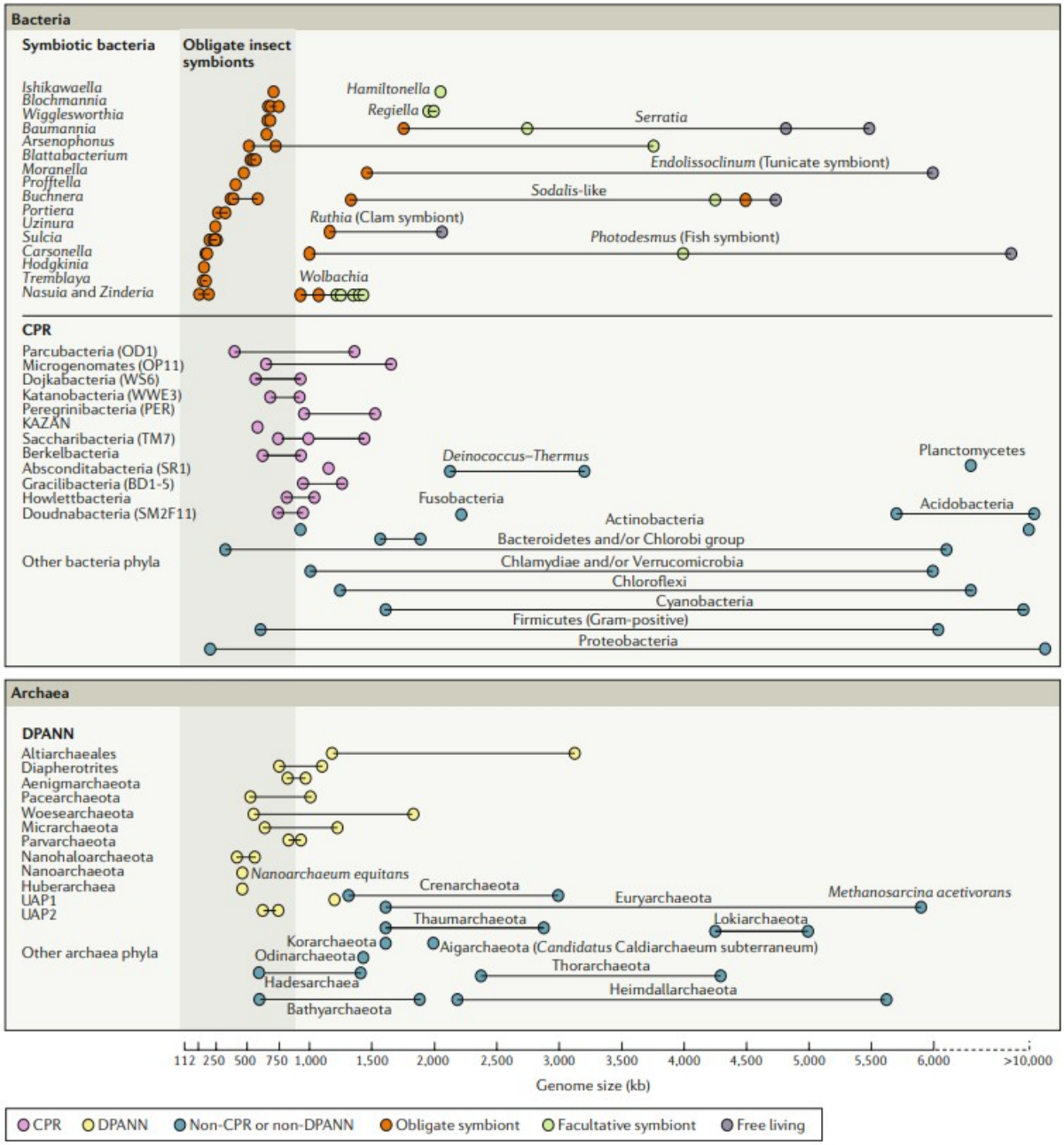

Fig. 2: The size ranges for CPR and DPANN genomes compared with size ranges for the genomes of known bacterial symbionts as well as other bacteria and archaea.

Adapted from ref.61, Annual Reviews.

The top panel shows data for well-studied bacteria that are obligate symbionts (orange dots), facultative symbionts (green dots) and free living (grey dots). The middle panel provides information for candidate phyla radiation (CPR; purple dots), and the bottom panel provides genome size information for DPANN (‘*Candidatus* Diapherotrites’, ‘*Candidatus*Parvarchaeota’, ‘*Candidatus* Aenigmarchaeota’, Nanoarchaeota and ‘*Candidatus*Nanohaloarchaeota’, in addition to other included lineages; yellow dots). The middle and bottom panels also show the size ranges for other bacteria and archaea (blue dots). CPR and DPANN genome sizes overlap with those of obligate symbionts.

An important question regarding the CPR and DPANN is whether small genomes are an ancestral characteristic or whether they are due to recent genome reduction. Genes for chromosome maintenance (for example, for DNA repair and recombination) and related functions[60] are often absent in the genomes of obligate intracellular symbionts undergoing genome reduction[60,61]. Thus, we investigated the presence of *mutM, mutY, mutL* and *mutS* genes involved in base excision repair and DNA mismatch repair pathways in CPR genomes (Supplementary Table 5). Overall, 56% of the CPR genomes have one or more of these genes. The absence of these genes in CPR genomes correlates with smaller genome size (Supplementary Figure 2), as previously reported for organisms with small genome size[60]. However, there is a wide range in genome size, even for those completely lacking *Mut* genes.

Homologous recombination is another capacity that often distinguishes symbionts that have undergone extensive genome reduction from other organisms. CPR and DPANN genomes typically encode genes for homologous recombination (for example, Holliday junction resolvases in CPR and DPANN, RecA in CPR and RadA in DPANN) (Supplementary Table 5). Although the CPR organisms lack the RecBCD enzyme (a helicase–nuclease that initiates the repair of double-stranded DNA breaks by loading RecA on the double-stranded DNA), they have the recFOR system (primarily responsible for recombination initiated at single-stranded DNA gaps, but it can also act at double-strand breaks). One peculiarity is that all '*Ca*. Parcubacteria' lack *recF* (Supplementary Table 5). However, in *Bacillus subtilis*, loss of RecF does not abolish function[62]. Thus, we suspect that '*Ca*. Parcubacteria', as well as most other CPR and DPANN organisms, are capable of homologous recombination and possibly carry out this process via an unusually compact system.

Genome reduction may be ongoing in some lineages. However, given that many CPR organisms have the capacity for homologous recombination, base excision repair and mismatch repair (as well as the apparent lack of abundant pseudo-genes), we conclude that bacteria from many groups are probably not currently undergoing genome reduction owing to unrepaired DNA damage. Rather, they are maintaining the integrity of their genomes. Besides, if adaptive genome streamlining was actively occurring in CPR and DPANN genomes, some larger genomes should have been retrieved. Thus, small genome size may be a reflection of an ancestral state. Regardless of whether it is due to rapid evolution or long evolutionary history, the CPR and DPANN lineages, in combination, account for an extensive amount of bacterial and archaeal diversity[4,7] (Fig. 1).

Anaerobic and symbiotic lifestyles

Most CPR and DPANN organisms lack a respiratory chain, including NADH dehydrogenase and complexes II–IV (Cx II–Cx IV) of the oxidative phosphorylation chain. Also notable is the lack of a complete tricarboxylic

acid cycle (TCA) cycle, although some organisms have a subset of enzymes of this cycle, likely for biosynthetic purposes (Fig. 3; Supplementary Table 5). On the basis of predictions of the metabolic capacities for most (but not all, see below) CPR and DPANN organisms, it is inferred that they are anaerobes.

Fig. 3: Profile of presence or absence of certain metabolic or biosynthetic capacities.

The metabolic or biosynthetic capacities (columns) are shown for 373 selected candidate phyla radiation (CPR) and DPANN (‘*Candidatus* Diapherotrites’, ‘*Candidatus* Parvarchaeota’, ‘*Candidatus* Aenigmarchaeota’, Nanoarchaeota and ‘*Candidatus* Nanohaloarchaeota’, in addition to other included lineages) genomes (rows). The sets of capacities tend to be consistent within specific lineages despite the derivation of the organisms within each group from diverse ecosystem types. Metabolism is sparse in most groups, but ‘*Candidatus*Peregrinibacteria’ consistently have substantially more biosynthetic capacity. Notably, subgroups within ‘*Ca*. Peregrinibacteria’ are distinguished by their choice of biosynthetic pathways, for example, the mevalonate (MVA) versus the 2-C-methylerythritol 4-phosphate (MEP) pathway for synthesis of isoprenoids. Conversely, other groups have especially minimal capacity (for example, ‘*Candidatus* Dojkabacteria’ (WS6) and ‘*Candidatus*Pacearchaeota’); intriguingly, organisms from both ‘*Ca*. Dojkabacteria’ and ‘*Ca*. Pacearchaeota’ also often have ribulose-1,5-bisphosphate (RuBP) carboxylase–oxygenase (Rubisco). Details of the ecosystem type and metabolic capacities are provided in Supplementary Table 4 and Supplementary Table 5, respectively. Another view of the metabolic and biological features in CPR and DPANN is provided in Supplementary Figure 3. FBPase, fructose 1,6-bisphosphatase; PDH, pyruvate dehydrogenase; PEP, phosphoenolpyruvate; PFK, phosphofructokinase; PFORα, pyruvate:2-oxoacid-ferredoxin oxidoreductase subunit alpha; PPP, pentose phosphate pathway; TCA, tricarboxylic acid.

To date, there are no reports of the capacity for complete gluconeogenesis in either the CPR or the DPANN radiation on the basis of the absence of glucose-6-phosphatase (which hydrolyses glucose-6-phosphate to glucose). Overall, a small subset of CPR and DPANN organisms possess a gluconeogenesis pathway up to the level of fructose-6-phosphate and glucose-6-phosphate, which in all likelihood is sufficient for a link to the non-oxidative pentose phosphate pathway (PPP) but not to glycogen or glucose synthesis.

Although many enzymes encoded in CPR and DPANN genomes require cofactors, pathways for their biosynthesis were rarely identified (Supplementary Table 5). Generally, the groups with the most extensive metabolic repertoires were found to have the most cofactor pathways (for example, ‘*Ca*. Peregrinibacteria’ and some ‘*Ca*. Diapherotrites’, ‘*Ca*. Micrarchaeota’ and ‘*Candidatus* Woesearchaeota’). For example, some ‘*Ca*. Peregrinibacteria’, ‘*Ca*. Micrarchaeota’ and ‘*Ca*. Diapherotrites’ can make NAD and NADP from L-aspartate and can produce riboflavin, flavin mononucleotide (FMN) and FAD, and can probably synthesize the folate required for one carbon reactions, as well as CoA. Evidence for cobalamin synthesis, possibly from riboflavin, via three enzymes, was found in one ‘*Ca*. Peregrinibacteria’ genome. However, virtually all CPR and DPANN organisms must acquire cobalamin (needed for ribonucleotide reductase) and other cofactors from external sources. As previously noted, many CPR and DPANN organisms (with the exception of many members of the ‘*Ca*. Peregrinibacteria’, ‘*Ca*. Diapherotrites’, ‘*Ca*. Micrarchaeota’ and some ‘*Ca*. Woesearchaeota’) (Supplementary Table 5) seem to lack complete pathways for the biosynthesis of amino acids[2,3,6,46,47]. Similarly, many (but not all) CPR and DPANN organisms lack the ability to de novo synthesize nucleotides (Fig. 3; Supplementary Table 5).

Perhaps the most surprising finding is that all CPR genomes studied to date possess substantially incomplete pathways for fatty acid biosynthesis (Supplementary Table 5). Lipids required to construct the cell membranes are likely to be derived from other organisms or scavenged from dead cells. Our analyses indicate that most CPR have complete or essentially complete pathways for peptidoglycan synthesis; yet surprisingly, '*Ca*. Katanobacteria' and '*Ca*. Dojkabacteria' (WS6) cannot make this cell wall compound (Fig. 3; Supplementary Table 5). The inability to de novo synthesize a cell envelope (despite the abovementioned evidence that they are cellular life forms) is one of the strongest indicators that many CPR and DPANN organisms fundamentally depend on other organisms for basic resources. The extent to which they rely on other organisms is predicted to vary dramatically on the basis of the inventory of metabolic capacities found in each group.

Filling metabolic gaps

Many questions remain regarding energy generation and redox cycling within CPR and DPANN cells. The majority of genomes encode ATP synthase, yet only a few '*Ca*. Parcubacteria' and one '*Ca*. Aenigmarchaeota' can export protons from the cell via type 4 membrane-bound hydrogenases[2,6]. Otherwise, it is unclear how the majority of CPR and DPANN organisms generate the proton motive force (PMF) needed if the ATP synthase complex is to produce ATP. From previous studies, it seems that some CPR and DPANN live as episymbionts, for example, as cells attached to the surfaces of other bacteria and archaea[10,39,52,57]. In such cases, an intriguing possibility for the source of PMF is to scavenge protons from their host cells by maintaining very close proximity to proton-exporting complexes of the host cells. Alternatively, the ATP synthases might extrude protons to drive antiporters, consuming ATP generated by substrate-level phosphorylation.

CPR and DPANN organisms must salvage amino acids from environmental sources for protein construction and as potential carbon and energy sources. As such, many contain numerous proteases, as well as several transporters, whose substrates are unknown and whose numbers vary greatly (from very few to 75 per genome) within the same lineage and across phyla (Supplementary Table 5). Their genomes often harbour multiple transaminases that convert certain amino acids (including aspartate, glutamate, valine and alanine) to 2-oxoacids (ketoglutarate or pyruvate) (Supplementary Table 5). Similarly, many (but not all) CPR and DPANN must scavenge nucleotides. However, they have nucleases and the genes necessary to repurpose scavenged nucleotides into DNA and RNA.

Most CPR[63] and DPANN are devoid of CRISPR–Cas adaptive immune systems and instead rely on a larger than expected set of restriction systems for nonspecific defence[63]. Although risky, lack of phage or virus avoidance by cell surface receptor proteins, enabling injection of phage

DNA, could represent a source of nucleotides as long as phages are intercepted. If the cells are attached to host cells, they could function as decoys, protecting their hosts from phages that might otherwise infect them[64]. Consistent with a possible function in DNA uptake, many CPR genomes contain at least one copy of ComEC, the DNA specific pore-forming protein required for competence[64], and the DNA protection protein DprA (Supplementary Figure 3; Supplementary Table 5).

CPR and DPANN organisms may have important roles in carbon cycling[2,48]. The inputs into their central carbon currencies and energy generation are often complex carbon compounds such as those acquired from degraded plant or general microbial biomass[48]. Members of both groups can degrade cellulose (often extracellularly) and starch to monomeric carbon, which can be oxidized via glycolysis (Supplementary Table 5). The most common glycoside hydrolase for both CPR and DPANN is α-amylase, which is, in some cases, predicted to be extracellular. This may also indicate widespread consumption of plant-derived and bacterially derived carbon storage compounds such as starch and glycogen.

Most of the genomes of both CPR and DPANN organisms lack phosphofructokinase (PFK) and complete the glycolytic pathway by using a metabolic shunt. Specifically, fructose-6-phosphate is converted into glyceraldehyde-3-phosphate via the non-oxidative PPP. Interestingly, many '*Ca*. Gracilibacteria', '*Ca*. Dojkabacteria' and '*Ca*. Pacearchaeota' have almost no capacity for upper glycolysis, and the PPP is essentially incomplete as well (Supplementary Table 5). The metabolic platform of these cells remains unclear.

It is striking that only the reversible glycolysis enzymes are somewhat consistently present in CPR and DPANN genomes. In metabolic pathways, enzymes catalysing essentially irreversible reactions are potential sites of control. In fact, PFK is lacking in many genomes, yet it is the most important control element for glycolysis[65]. The rate of conversion of glucose into pyruvate must be regulated to meet two major cellular needs: the production of ATP that is generated by the degradation of glucose and the provision of building blocks for biosynthetic reactions. The absence of one, two or three of the regulatory enzymes (for example, glucokinase, PFK and pyruvate kinase) of this pathway poses the question of how the flux through the glycolytic pathway is adjusted in response to conditions both inside and outside the cell.

The central carbon degradation pathways of CPR and DPANN very often lead to pyruvate and acetyl-CoA, raising the question of the fate of these compounds. Many DPANN use the pyruvate dehydrogenase (PDH) complex to decarboxylate pyruvate and form acetyl-CoA. This is interesting, as this complex is usually found in aerobic organisms[40], mostly in Eukarya and Bacteria (Supplementary Table 5). Only few CPR organisms have PDH. Instead, they generally convert pyruvate to acetyl-CoA using pyruvate:2-

oxoacid-ferredoxin oxidoreductase (Supplementary Table 5). Many CPR and DPANN likely use the acetyl-CoA to produce small-chain fatty acids to balance carbon and electron flow. For example, many '*Ca*. Parcubacteria', '*Ca*. Dojkabacteria' and '*Ca*. Microgenomates' are predicted to use ADP-acetyl-CoA synthetase (ADP-Acs) for acetate generation. The ADP-Acs pathway is common in Archaea but rare in Bacteria[42]. '*Ca*. Peregrinibacteria' use a pathway involving acetate kinase (Ack) and phosphotransacetylase (Pta) to produce acetate (the Ack–Pta pathway, which is mainly found in Bacteria). Other than acetate, many CPR and DPANN organisms are predicted to produce lactate, formate and/or ethanol (Supplementary Table 5). Carbon compounds produced via fermentation pathways are probably excreted from CPR and DPANN cells. In fact, this may be a key role of CPR and DPANN in their ecosystems, as acetate, lactate, ethanol and formate produced by fermentation could support growth of aerobic or anaerobic respiratory organisms.

Certain CPR bacteria are abundant in acetate-amended groundwater[3,22,23]. Some '*Ca*. Microgenomates' have genomes that encode AMP-acetyl-CoA synthetase, which may confer the ability to use (as well as produce) acetate. The Ack–Pta pathway found in '*Ca*. Peregrinibacteria' may be completely reversible and may function optimally at high concentrations of acetate. This mechanism may lead to the proliferation of '*Ca*. Peregrinibacteria' following acetate stimulation.

Some CPR ('*Ca*. Parcubacteria', '*Ca*. Microgenomates' and '*Ca*. Katanobacteria') and DPANN ('*Ca*. Micrarchaeota', '*Ca*. Parvarchaeota', '*Ca*. Woesearchaeota' and '*Ca*. Aenigmarchaeota') organisms may generate H_2 via fermentation to dispose of excess reductant from glycolysis, and some may have the ability to use H_2. A subset of CPR and DPANN possess membrane-bound and cytoplasmic NiFe hydrogenases (Fig. 4a), and a few have iron-only hydrogenases (Supplementary Table 5). Phylogenetic analyses of the CPR and DPANN NiFe hydrogenase catalytic subunits revealed that most are type 3b cytoplasmic hydrogenases most closely related to those of fermentative, sulfur-reducing Thermococcales archaea[66] (Fig. 4a). These bidirectional hydrogenases catalyse the reversible oxidation of H_2. When consuming H_2, these enzymes could produce the reduced form of NADPH for anabolic metabolism. When reduced sulfur compounds such as polysulfide are available, the type 3b hydrogenases could use them as terminal electron acceptors during sugar fermentation, producing H_2S (sulfhydrogenases)[2,6].

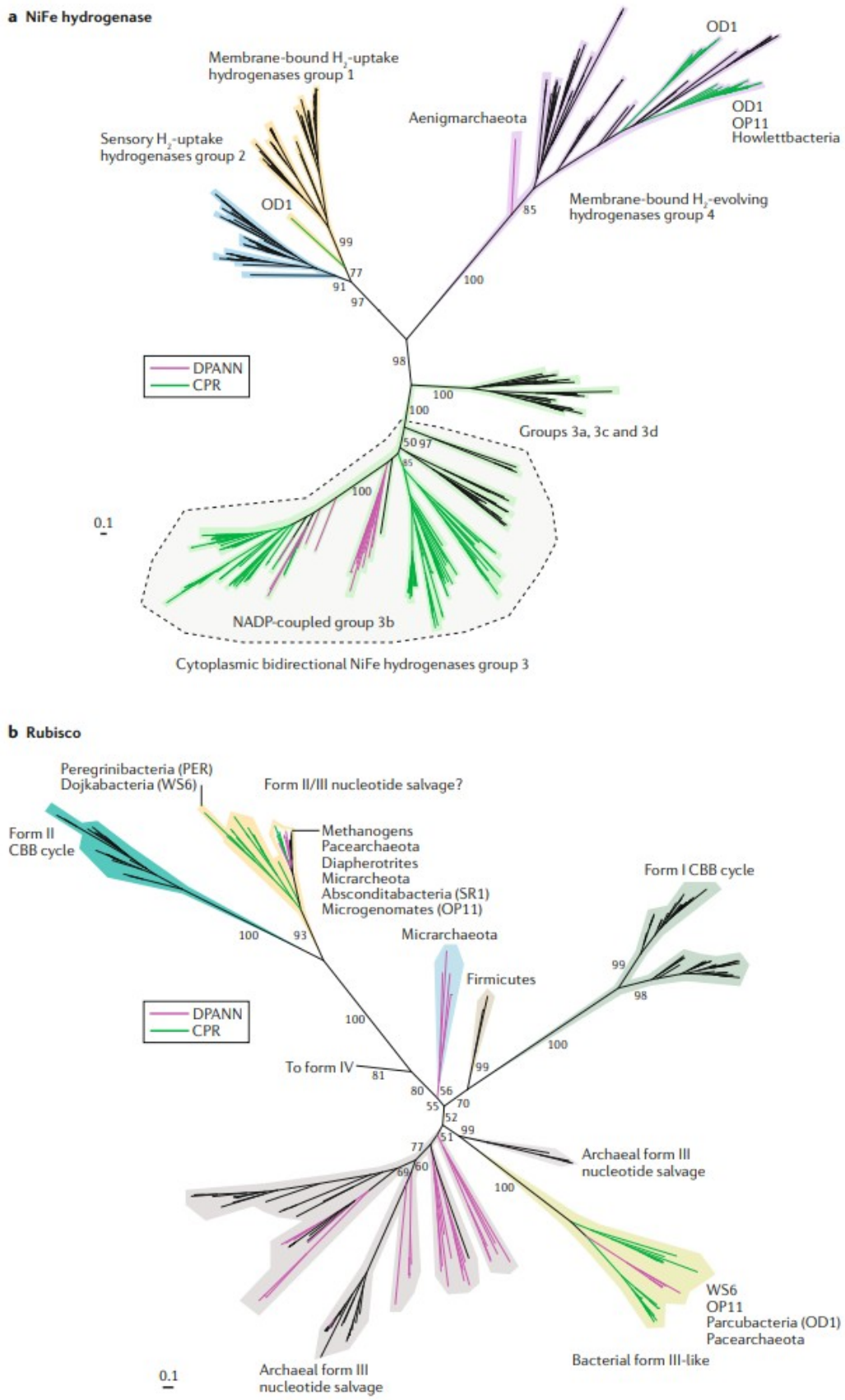
a NiFe hydrogenase
Membrane-bound H_2-uptake hydrogenases group 1
Sensory H_2-uptake hydrogenases group 2
OD1
Aenigmarchaeota
OD1
OD1
OP11
Howlettbacteria
Membrane-bound H_2-evolving hydrogenases group 4
99
77
91
97
85
100
98
100
100
50
97
85
100
DPANN
CPR
Groups 3a, 3c and 3d
0.1
NADP-coupled group 3b
Cytoplasmic bidirectional NiFe hydrogenases group 3
b Rubisco
Peregrinibacteria (PER)
Dojkabacteria (WS6)
Form II/III nucleotide salvage?
Form II
CBB cycle
Methanogens
Pacearchaeota
Diapherotrites
Micrarcheota
Absconditabacteria (SR1)
Microgenomates (OP11)
Form I CBB cycle
100
93
Micrarchaeota
Firmicutes
99
98
DPANN
CPR
100
100
To form IV
81
99
80
56
55
70
52
99
51
77
60
69
Archaeal form III
nucleotide salvage
100
WS6
OP11
Parcubacteria (OD1)
Pacearchaeota
Bacterial form III-like
0.1
Archaeal form III
nucleotide salvage

Fig. 4: Maximum likelihood phylogenetic trees constructed for the catalytic subunits of NiFe hydrogenases and Rubisco.

a | The NiFe hydrogenases tree includes the following different groups: 1, 2, 3a, 3b, 3c, 3d and 4. Most candidate phyla radiation (CPR) and DPANN (‘*Candidatus* Diapherotrites’, ‘*Candidatus* Parvarchaeota’, ‘*Candidatus* Aenigmarchaeota’, Nanoarchaeota and ‘*Candidatus*Nanohaloarchaeota’, in addition to other included lineages) organisms possess the cytoplasmic group 3b hydrogenase, which functions as a bidirectional enzyme enabling them to use H_2 as a source of reducing power, or either protons or elemental sulfur to dispose of reducing equivalents generated from fermentation. One ‘*Ca*. Aenigmarchaeota’ and several ‘*Candidatus* Parcubacteria’ have a membrane-bound group 4 hydrogenase that can produce H_2 and is likely coupled to proton translocation. **b** | The ribulose-1,5-bisphosphate (RuBP) carboxylase–oxygenase (Rubisco) tree includes the following distinct forms: form I, form II, archaeal form III, form IV and recently defined form II/III and form III-like (which currently includes only CPR and DPANN sequences). In most CPR organisms that have a Rubisco gene, it is either form II/III or form III-like (some ‘*Candidatus* Dojkabacteria’ have both forms). The Rubisco in DPANN archaea is most commonly form III, although some have form II/III or form III-like. These three distinct forms of Rubisco likely function in a nucleoside pathway that feeds into lower glycolysis (similar to the pathway proposed in Archaea[68]). Both phylogenetic trees were generated using RAxML with the PROTCAT JTT model (proteins were aligned using MAFFT[100] and manually curated). Bootstrap support values above 50 are indicated on both trees and are based on 100 re-samplings. The NiFe hydrogenases and Rubisco trees are available with full bootstrap values in Newick format in Supplementary Data 4 and Supplementary Data 5, respectively. CBB, Calvin–Benson–Bassham.

Despite being the most extensively studied enzyme to date, multiple new ribulose-1,5-bisphosphate carboxylase–oxygenase (Rubisco) forms have been identified through the analysis of CPR and DPANN genomes. Genes encoding form II/III Rubisco, originally thought to occur in only methanogenic archaea, were discovered in the genomes of CPR bacteria[6,47,49,67] (‘*Ca*. Peregrinibacteria’, ‘*Ca*. Dojkabacteria’ and ‘*Ca*. Absconditabacteria’) and DPANN ‘*Ca*. Pacearchaeota’[2]. Additionally, divergent form III (bacterial form III-like[67]) Rubisco has been reported in CPR genomes. The sequences define a distinct and strongly supported monophyletic group that branches deeply from, yet is most similar to, the archaeal form III Rubisco[67]. New analyses of Rubiscos from recently published DPANN genomes[11,12] reveal that the form III-like Rubisco is also found in some DPANN organisms (Fig. 4b). We also expanded the form II/III Rubisco clade by the addition of new DPANN sequences from the ‘*Ca*. Diapherotrites’ and ‘*Ca*. Micrarchaeota’ phyla (Fig. 4b). These Rubiscos are not believed to function within the Calvin–Benson–Bassham (CBB) pathway but rather in a pathway involving nucleotides or nucleosides[68]. Rubisco genes were recovered in ‘*Ca*. Dojkabacteria’ and ‘*Ca*. Pacearchaeota’ with the smallest median genome size across the CPR and DPANN radiation, respectively (Supplementary Table 5). As they lack upper glycolysis and the entire PPP, they must rely on other members of the community for external ribose to feed into the nucleoside pathway. In fact, some ‘*Ca*. Dojkabacteria’ possess two Rubiscos, which are of different types[69]. The prominence of these genes suggests a central importance of Rubisco in the metabolism of these organisms.

Metabolic capacities

We investigated metabolic capacities across the CPR and DPANN organisms from various ecosystems to determine whether there is evidence for a range of degrees of dependence on other organisms. This analysis is important because it may throw light on the evolutionary processes that led to the current metabolic patterns. We found that, although all seem to have a fermentative-based lifestyle, biosynthetic capacities vary greatly across both groups (Fig. 3). Overall, the analysis of nearly 1,000 CPR and DPANN genomes reconstructed from different studies and environments[2,3,5,6,7,11,12,19] highlights a consistency in their metabolic content, apart from the few exceptions described below (Figs 3,5,6; Supplementary Table 5).

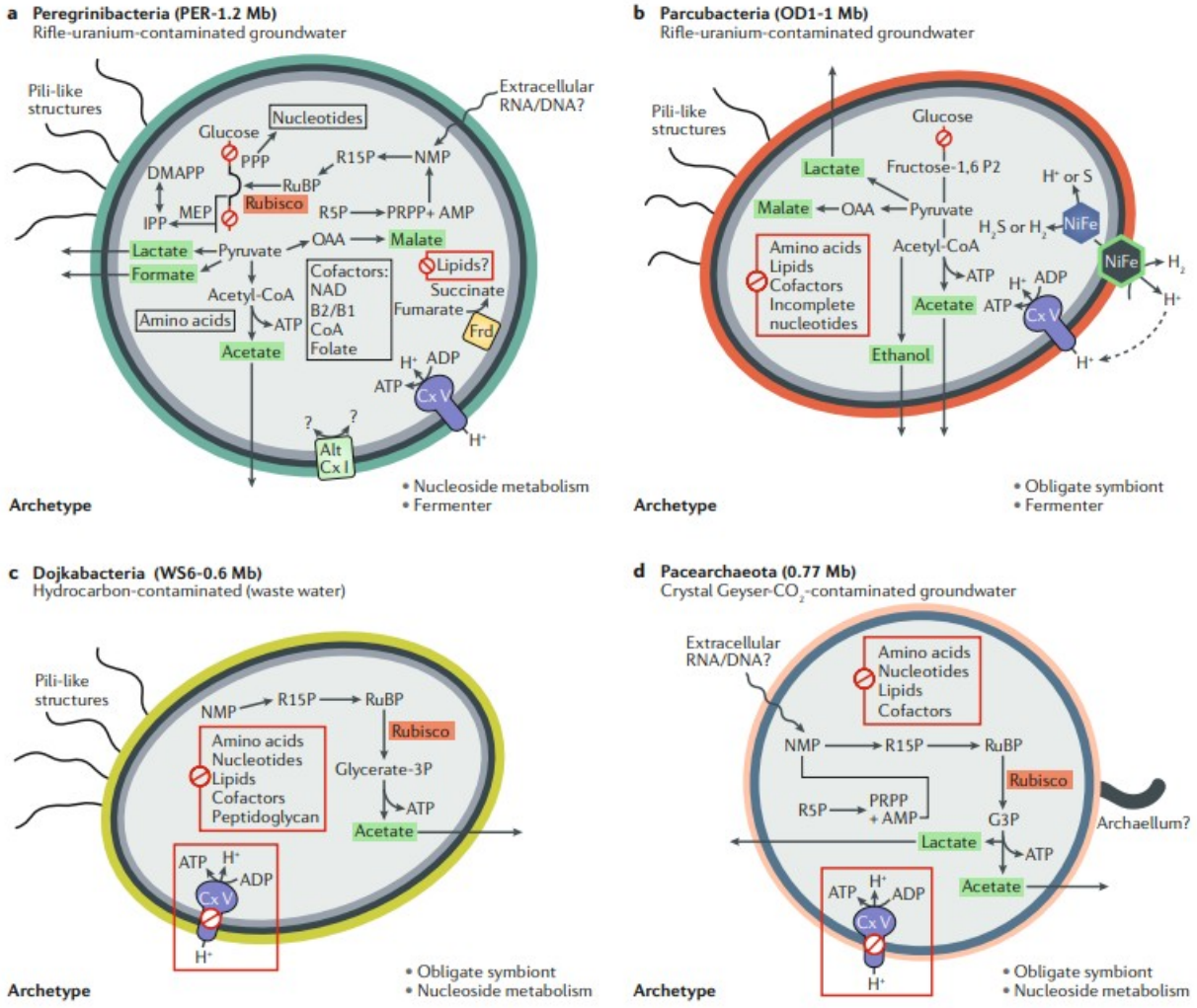

Fig. 5: Central metabolism of some CPR and DPANN, examples of typical configurations.

a | A schematic of capacities of a typical '*Candidatus* Peregrinibacteria' (National Center for Biotechnology Information (NCBI) accession number: LCFW01000001) is shown. Unlike most other candidate phyla radiation (CPR) organisms, these bacteria are predicted to synthesize nucleotides and amino acids. They may be essentially free living but seem to require lipids from an external source. **b** | A typical metabolic prediction for a '*Candidatus* Parcubacteria' (NCBI accession number: LBRD01000001)

is shown. **c** | A schematic for '*Candidatus*Dojkabacteria' with typical minimal capacities but with ribulose-1,5-bisphosphate (RuBP) carboxylase–oxygenase (Rubisco), which is probably central to the nucleotide salvage pathway (UBA4813; NCBI accession number: DHFX00000000), is shown. **d** | A schematic representing the typical predicted minimal capacities in '*Candidatus* Pacearchaeota' is shown, but Rubisco is included and is probably central to the nucleotide salvage pathway (NCBI accession number: MNVW01000002). Alt Cx I, NADH dehydrogenase type II; Cx V, complex V (ATP synthase); DPANN, '*Candidatus* Diapherotrites', '*Candidatus* Parvarchaeota', '*Candidatus* Aenigmarchaeota', Nanoarchaeota and 'Candidatus *Nanohaloarchaeota*', in addition to other included lineages; Frd, fumarate reductase; G3P, glycerate 3-phosphate; MEP, 2-C-methylerythritol 4-phosphate; NiFe, NiFe hydrogenase; NMP, nucleoside 5′-monophosphate; OAA, oxaloacetate; PPP, pentose phosphate pathway; PRPP, phosphoribosyl pyrophosphate; R15P, ribose-1,5-bisphosphate; R5P, ribose-5-phosphate.

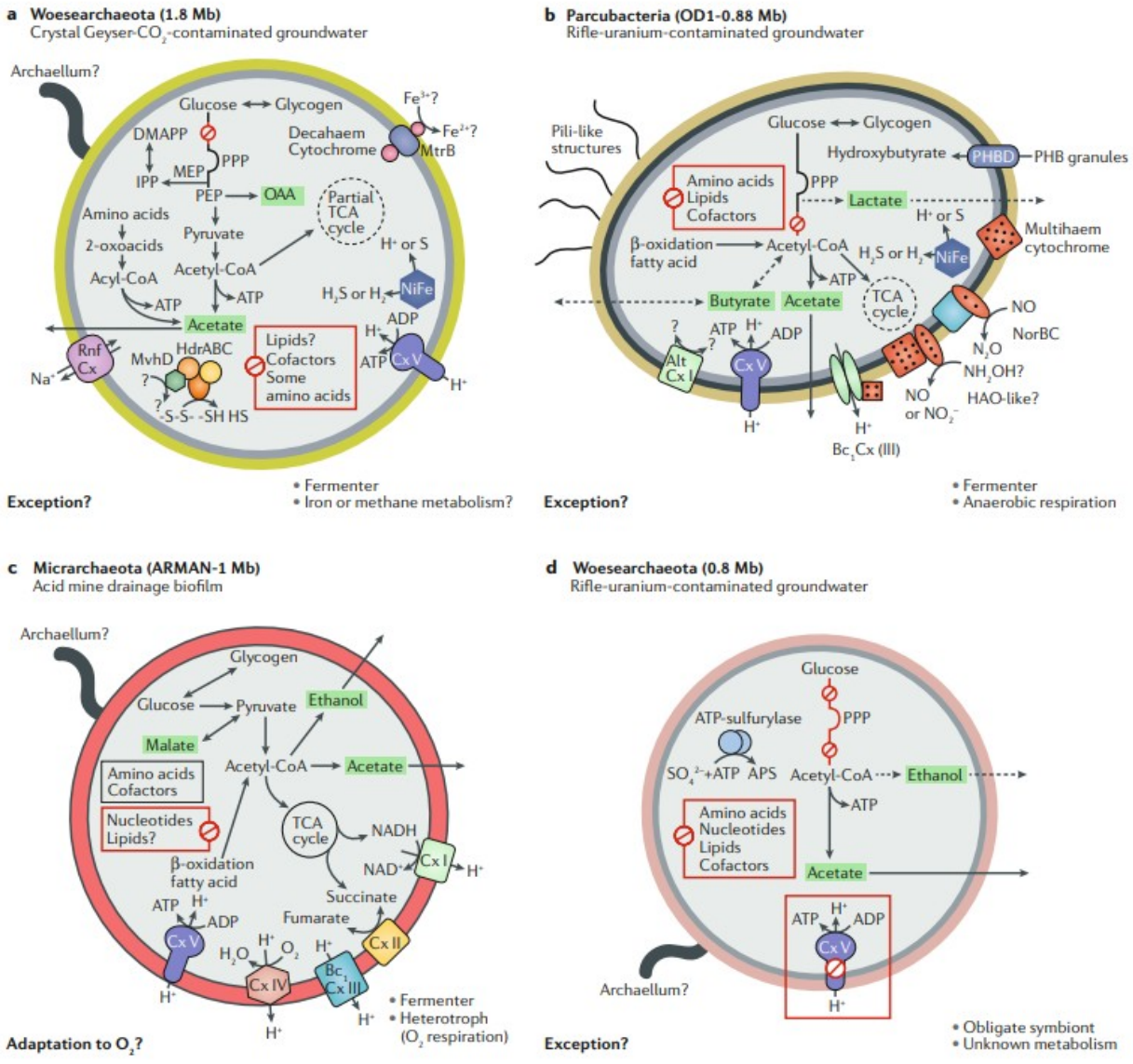

Fig. 6: Central metabolism of some CPR and DPANN, examples of exceptions.

Part **b** adapted from ref.71, Springer Nature Limited. Part **d** adapted with permission from ref.2, Elsevier.

a | Unusually, one '*Candidatus* Woesearchaeota' genome encodes a complex that includes extracellular and cytoplasmic cytochromes, potentially involved in iron metabolism, and heterodisulfide reductase and Rnf complexes (Rnf Cx) (National Center for Biotechnology Information (NCBI) accession number: PCVE01000001). **b** | One bacterium ('*Candidatus*Parcunitrobacter nitroensis') from the Parcubacteria superphylum is predicted to have numerous unexpected capacities, encoded by highly divergent enzymes, and may be involved in nitrogen compound metabolism (NCBI accession number: LBUF00000000). **c** | The capacities predicted to be typical of '*Candidatus* Micrarchaeota' from acid mine drainage (NCBI accession number: MOEG01000001) are shown; these capacities are identical in '*Candidatus* Parvarchaeota' from the same environment). The presence of cytochrome oxidase may be an adaptation to life in an aerobic or microaerophilic environment. **d** | An example of the unusually minimal metabolism of '*Ca*. Woesearchaeota', lacking upper glycolysis and ATP synthase (NCBI accession number: CP010426), is shown. Alt Cx I, NADH dehydrogenase type II; Bc_1 Cx III, Bc_1 complex III; CPR, candidate phyla radiation; Cx I, complex I (NADH dehydrogenase); Cx II, complex II (succinate dehydrogenase); Cx IV, complex IV (terminal oxidase); Cx V, complex V (ATP synthase); DPANN, '*Candidatus*Diapherotrites', '*Ca*. Parvarchaeota', '*Candidatus* Aenigmarchaeota', Nanoarchaeota and '*Candidatus* Nanohaloarchaeota', in addition to other included lineages; HAO, hydroxylamine oxidoreductase; HdrABC, heterodisulfide reductase ABC; MtrB, decahaem-associated membrane protein required for Fe(III) reduction; MvhD, F420-non-reducing hydrogenase iron-sulfur subunit D; NiFe, NiFe hydrogenase; NorBC, nitric oxide reductase; OAA, oxaloacetate; PEP, phosphoenolpyruvate; PHB, polyhydroxybutyrate; PHBD, polyhydroxybutyrate depolymerase; PPP, pentose phosphate pathway; TCA, tricarboxylic acid.

Organisms from the '*Ca*. Peregrinibacteria' phylum have moderately extensive core biosynthetic capacities. All seem to synthesize nucleotides, certain amino acids and cofactors, but they cannot synthesize required fatty acids (Supplementary Table 5). However, most organisms from this phylum invest greatly in biosynthesis of cell envelope components, including peptidoglycan. Their genomes encode large extracellular proteins, some of which are cysteine rich and may function in cell attachment[70]. They typically have the capacity to synthesize isoprenoids via the archaeal mevalonate pathway[70] or the bacterial 1-deoxy-D-xylulose 5-phosphate (DOXP)–2-C-methylerythritol 4-phosphate (MEP) pathway (Figs 3,5a; Supplementary Table 5), but the type and localization of the products remain uncertain. They are non-respiring anaerobes predicted to ferment, likely producing acetate, lactate and formate from pyruvate (no formate dehydrogenases and only one hydrogenase were identified). '*Ca*. Peregrinibacteria' genomes lack detectable mechanisms for recovering additional energy via membrane potential coupled to ATP synthase, but many may use Rubisco for making ATP by shunting nucleotides through central carbon metabolism (Fig. 5a).

The metabolisms of members of the Parcubacteria superphylum can vary greatly, from minimal to more complex (Supplementary Table 5). One archetype for this large group is the consistent lack of complete core biosynthetic capacities (nucleotides, lipids, fatty acids and many amino acids) (Supplementary Table 5). Many can produce acetate, ethanol, lactate and hydrogen as fermentation end products (Fig. 5b). In a few members, hydrogen could be generated via NiFe type 4 membrane-bound hydrogenase and cytoplasmic type 3b hydrogenase (Fig. 5b). In this case,

the dual hydrogenase system may function in intracellular hydrogen cycling, where H_2 and PMF are produced by the membrane-bound hydrogenase, and H_2 is shuttled to the cytoplasmic hydrogenase[6] (Fig. 5b).

One exception within the Parcubacteria superphylum is '*Candidatus*Parcunitrobacter nitroensis' (ref.71), which has an extensive metabolic repertoire (Fig. 6b). This parcubacterium has an essentially complete electron transport chain, both fermentative and respiratory capacities and nitrogen and fatty acid metabolism. Importantly, the sequences of all enzymes involved in nitrogen compound-based respiration (nitrite reductase putative hydroxylamine oxidoreductase and a nitric oxide reductase) are highly divergent from sequences found in other organisms, suggesting that these capacities were not recently acquired from non-CPR organisms through lateral gene transfer (LGT). However, this bacterium lacks the capacities to produce lipids, nucleotides and multiple amino acids. Thus, it seems to depend on its surroundings to meet these requirements[71]. Additionally, several '*Ca*. Parcubacteria'[72]have genomes that encode a copper nitrite reductase (*nirK*) and/or an NADPH nitrite reductase (*nirB*), which form nitric oxide and ammonium from nitrite, respectively (Supplementary Table 5).However, the lack of other respiratory complexes and/or enzymes in these genomes suggests a putative role in detoxification of nitrite rather than anaerobic respiration or denitrification. Finally, analyses of 12 '*Ca*. Parcubacteria' single-cell-amplified genomes from marine sediments revealed the capacity for respiration, and 1 is predicted to have the ability for nitrate reduction[73]. Overall, metabolic versatility is heterogeneous in the Parcubacteria superphylum, possibly reflecting adaptation to different environment types.

The DPANN lineages that are predicted to have far more extensive biosynthetic capacities than others, and therefore could be free living, are the '*Ca*. Diapherotrites', '*Ca*. Micrarchaeota' (Fig. 6c) and '*Ca*. Parvarchaeota' (Supplementary Table 5). Two '*Ca*. Micrarchaeota' and one '*Ca*. Parvarchaeota' have essentially complete electron transport chains. Some DPANN genomes ('*Ca*. Micrarchaeota' and/or '*Ca*. Parvarchaeota' and '*Ca*. Woesearchaeota') encode near-complete TCA cycles and genes for the catabolism of fatty acids via β-oxidation (Figs 3,6c). For example, '*Ca*. Micrarchaeota' ARMAN-1 and ARMAN-2 ('*Candidatus* Micrarchaeum acidiphilum') genomes also encode some membrane-bound subunits of the NADH dehydrogenase (systematically lacking the NADH-binding module), succinate dehydrogenase, cytochrome bd-oxidase and ATP synthase enzymes involved in oxidative phosphorylation (Fig. 6c). One '*Ca*. Woesearchaeota' genome encodes the components of a potential Mtr respiratory pathway involved in iron metabolism[74,75,76] (for example, multihemes, cytochromes and an associated membrane protein) (Fig. 6a). This genome also encodes a cytoplasmic MvhD–HdrABC complex (F420-non-reducing hydrogenase iron-sulfur subunit D and heterodisulfide

reductase ABC), which has been shown to reduce the disulfide of coenzyme M and coenzyme B (CoMS–SCoB) in the final step in all methanogenic pathways[77,78]. However, the lack of other key enzymes in methane metabolism prevents assignment of a role in methanogenesis or methane oxidation. This archaeon may be capable of synthetizing the precursors of isoprenoids via the bacterial MEP pathway, as previously noted in other '*Ca*. Woesearchaeota' genomes[79]. By contrast, another '*Ca*. Woesearchaeota', '*Candidatus* Woesearchaeota archaeon AR20', represented by the first complete woesearchaeotal genome[2], has extremely reduced biosynthetic and metabolic capacities and appears to be an obligatory symbiont (Fig. 6d).

In contrast to CPR and DPANN with relatively complex metabolisms, '*Ca*. Katanobacteria', '*Candidatus* KAZAN' and '*Ca*. Dojkabacteria' (Fig. 5c) within CPR and '*Ca*. Pacearchaeota' (Fig. 5d) within DPANN consistently lack ATP synthase and pyrophosphatases (proton-pumping proteins complexes) (Fig. 3; Supplementary Table 5). These organisms also lack genes for the synthesis of nucleotides, most or all amino acids, lipids, peptidoglycan (except for '*Ca*. KAZAN') and cofactors. Overall, these genomes have the most reduced biosynthetic and catabolic capacities of the organisms studied here, pointing to an unusually strong dependence on other organisms for almost all non-information system requirements.

We conclude that biosynthetic capacities across both the CPR and the DPANN radiation vary greatly; yet in cases where capacities are most limited, the pattern persists across major groups. This may indicate that ancient episodes of genome reduction initiated lineage divergence. Similarly, more complete capacities that suggest that the organisms can be free living tend to occur in distinct phylogenetic groups. However, the exceptional case of the single Parcubacteria ('*Ca*. Parcunitrobacter nitroensis') with extensive but highly divergent metabolism defies explanation, as its gene set cannot be attributed to recent LGTs.

Information systems

Given their tiny genome and cell sizes and minimal metabolism, it is often asked whether these organisms are more like viruses than cells. Their somewhat easily recognizable ribosomes and their ability to transform energy and carbon compounds clearly distinguish them as living organisms. However, the vast majority of CPR bacteria seem to have unusual ribosome compositions. All are missing a ribosomal protein often lacking in symbionts (ribosomal protein L30 (rpL30)), and some specific lineages are missing ribosomal proteins and biogenesis factors considered universal in Bacteria (rpL9, rpL1 and GTPase Der)[3]. These characteristics imply different ribosome structures and mechanisms, highlighting the divergence of these organisms relative to other bacteria.

From the point of view of transcription, most CPR genomes encode multiple σ70-family sigma factors from groups 1–4 (strangely, one '*Ca*.

Peregrinibacteria' genome encodes 17 sigma factors) (Supplementary Table 5). Sigma factors recognize the promoter DNA sequence and alter the structure of the RNA polymerase to initiate transcription[80]. Group 1 (RpoD), which is ubiquitous across CPR, is essential for cell viability. Groups 2–4 (RpoS, RpoE and ECF) are non-essential and control various adaptive responses, such as morphological development and stress management[81], and are common in CPR. This observation suggests that CPR cells are responsive to changing environments. The σ54 factor, which regulates processes related to nitrogen metabolism[82], has not yet been found in CPR. As expected, DPANN archaea mostly use archaeal transcription factors to regulate gene expression. Archaea were thought to lack sigma factors, but recent single-cell genome analyses of some DPANN archaea revealed three proteins in two '*Ca*. Diapherotrites' and one '*Ca*. Woesearchaeota' with σ70 domains, which belong to the non-essential groups 3 and 4 (ref.1).

With respect to translation, both archaeal-type and bacterial-type tryptophan and tyrosyl tRNA synthetases, along with their cognate archaeal as well as bacterial tRNAs, are encoded in CPR genomes[83]. In addition to the case of '*Candidatus* Beckwithbacteria' (within the superphylum Microgenomates)[83], we identified both archaeal-type and bacterial-type tryptophan tRNA synthetases and tRNAs in close proximity in the same genome of a '*Candidatus* Shapirobacteria' species (within the superphylum Microgenomates) (Supplementary Figure 4). Further, we identified a bacterial tryptophan tRNA synthetase along with two archaeal variants in one '*Ca*. Pacearchaeota' (Supplementary Figure 4). The reasons, if any, for such dual systems are unclear, but we speculate that they were acquired via inter-domain LGT.

Numerous and complex modifications of tRNA are important in microbial translation. As expected, CPR genomes encode methionyl-tRNA formyltransferase, which initiates translation in Bacteria. However, one might predict that non-essential tRNA modification enzymes would be absent in small genomes. However, we find that CPR genomes encode various non-essential tRNA modification enzymes, including some that could modulate the fidelity and efficiency of translation efficiency. They do not seem to encode proteins used by most bacteria to modify tRNAs to increase the efficiency and accuracy of protein translation (MnmE and MnmG[84]). Instead, many CPR genomes have a single subunit, elongator protein 3 (Elp3), a protein with a radical sterile α-motif (SAM) domain that is common in Archaea[85] and Eukarya[86] (part of the eukaryotic multisubunit elongator complex), that performs the same function. Within the CPR, Elp3 is not homogeneously distributed across lineages. It is common in '*Ca*. Parcubacteria' and occurs in certain '*Ca*. Microgenomates', '*Ca*. Peregrinibacteria', '*Ca*. Dojkabacteria', '*Ca*. Katanobacteria' and '*Ca*. Gracilibacteria' (Fig. 7). Although Elp3 was noted in some bacteria and archaea previously[85,87], it has not been previously reported in CPR and DPANN genomes. Interestingly, in non-CPR bacteria, Elp3 is largely

restricted to some Chloroflexi (related to *Dehalococcoides* spp. and novel groups) and certain Actinobacteria[85] (Fig. 7). We conducted a phylogenetic analysis that includes CPR and non-CPR bacterial Elp3 sequences, as well as those from DPANN and other archaea (Fig. 7). Some Chloroflexi sequences group with DPANN '*Ca*. Pacearchaeota' sequences, away from those of CPR and, other Chloroflexi and Actinobacteria. By contrast, the actinobacterial and other Chloroflexi sequences place within the radiation of CPR sequences. These findings suggest inter-domain LGT to Chloroflexi from DPANN archaea and inter-phylum LGT from CPR bacteria to other Chloroflexi and Actinobacteria and from '*Ca*. Woesearchaeota' to '*Ca*. Diapherotrites' (Fig. 7). The CPR Elp3 sequences are highly divergent from those in Archaea, and the branch length in the CPR is comparable to that associated with all Archaea, suggesting an evolutionary history on the same scale as that of the entire archaeal domain. This finding could either indicate comparably rapid evolution of Elp3 within the CPR or an origin for this gene in a common ancestor of both Archaea and the CPR.

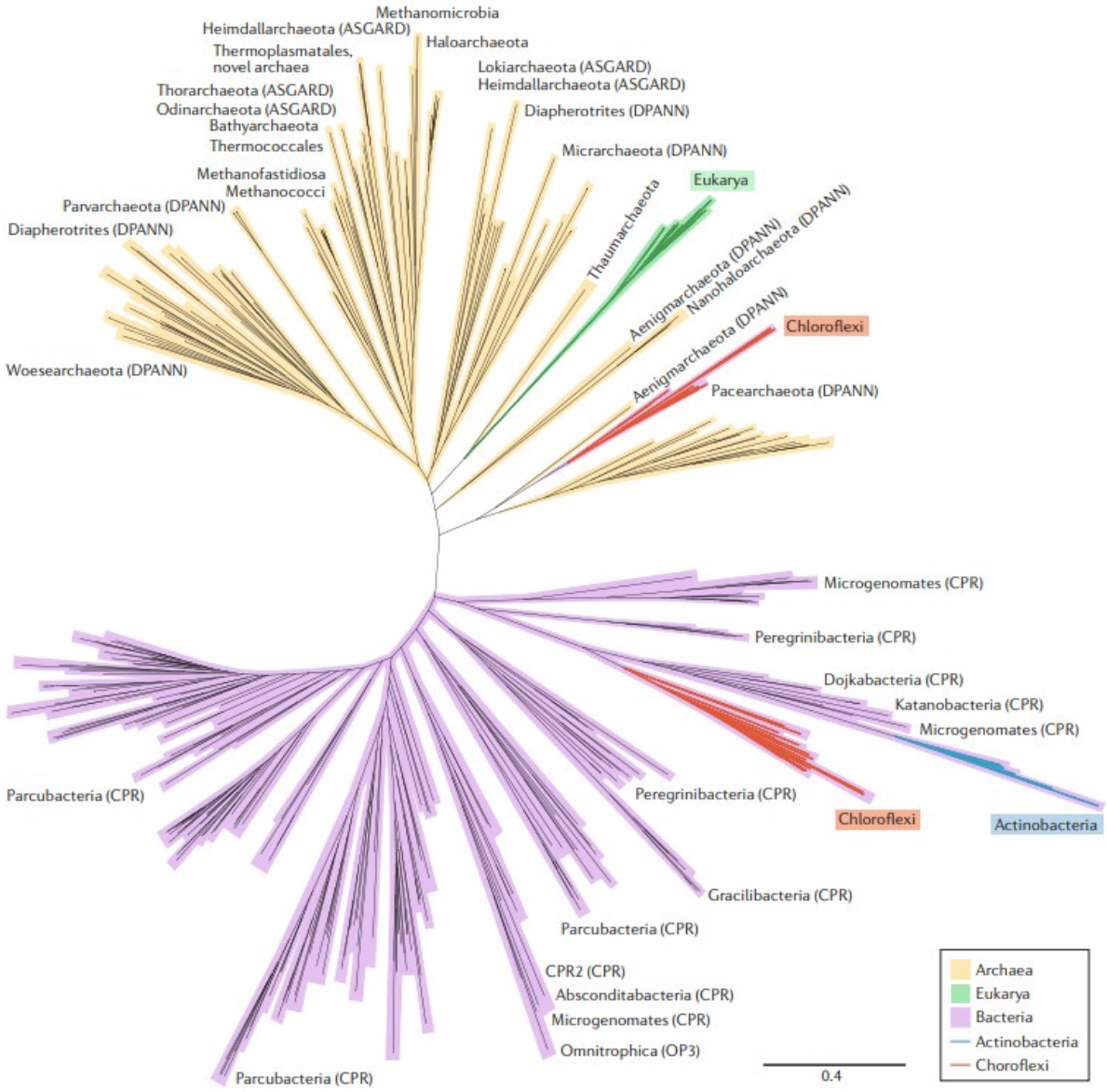

Fig. 7: Maximum likelihood phylogenetic tree of the catalytic subunit Elp3, which is found in genomes of some CPR bacteria and DPANN archaea.

Elongator protein 3 (Elp3) is the homologue of the catalytic subunit of the six-subunit eukaryotic elongator complex that modifies the wobble uridine (U34) at C5 of tRNAs. Almost all archaea and many candidate phyla radiation (CPR) bacteria possess Elp3 homologues but lack genes encoding the other five elongator subunits. A few non-CPR bacterial groups have Elp3, which was possibly acquired via inter-domain and inter-phylum lateral gene transfers. Specifically, Actinobacteria and some Chloroflexi may have received this enzyme from CPR, whereas other Chloroflexi may have acquired the sequences from DPANN (‘*Ca*. Diapherotrites’, ‘*Ca*. Parvarchaeota’, ‘*Ca*. Aenigmarchaeota’, Nanoarchaeota and ‘*Ca*. Nanohaloarchaeota’, in addition to other included lineages) archaea. The phylogenetic tree was generated using RAxML with the PROTCAT JTT model (the protein alignment was generated using MAFFT[100] and was manually curated) (Supplementary Methods). The Elp3 tree is available with full bootstrap values in Newick format in Supplementary Data 6.

Organism–organism interactions

Given indications that many CPR and DPANN organisms are closely dependent on other community members, we sought indications of the mechanisms by which organism–organism interactions might occur. For CPR, many genes are involved in the production of type IV pili, which may have a role in DNA uptake[64] or secretion of compounds and interactions with surrounding cells[52]. The surfaces of some CPR cells are extensively decorated by pili[52].

CPR and DPANN genomes typically encode numerous glycosyltransferases[2,47], indicating that they devote a substantial amount of energy to the production and reconfiguration of saccharides, polysaccharides or glycoproteins. These compounds may be involved in attachment and regulation of the local environment surrounding the cell surface.

Many CPR and DPANN genomes encode proteins containing one or more of the following domains: Ig-like, concanavalins (lectins), pectin lyases, fibronectin type III, β-propeller, von Willebrand factor and polycystic kidney disease (PKD) (Supplementary Table 5). Most are predicted to be on the cell surface or extracellular[2,47]. These proteins are often very large, with up to 10,000 amino acids, and there may be many in a single genome. These large proteins are sometimes encoded in clusters, and a few are cysteine rich[70]. Overall, given their small genome sizes, the observations point to the importance of surface attachment for the survival of CPR and DPANN cells.

One potentially important role for cell surface proteins may be in conferring host specificity. Diversity-generating retroelements (DGRs), a recently discovered family of genetic elements that modify DNA sequences[88] and the proteins they encode, occur in many CPR and DPANN genomes[89,90]. The DGRs target some proteins involved in surface attachment, defence and regulation[90]. Thus, they could enable adaptation to maintain binding specificity in the face of host cell surface protein evolution. DGRs are especially prominent in '*Ca*. Pacearchaeota', which probably adopt an obligate symbiotic lifestyle, given their highly reduced biosynthetic and metabolic capacities (Supplementary Table 5).

Introns in protein-coding genes and RNAs

Previously, the presence of introns in many CPR 16S rRNA and 23S rRNA genes was reported[3]. Such introns have only occasionally been reported in Bacteria[91,92] and Archaea[9,93]. The 16S rRNA introns often occur in many positions within the gene and can total >5 kb in length. Notably, CPR and DPANN organisms typically have only one set of rRNA genes, and so these intron-bearing genes must remain functional. Metatranscriptomic data showed that the introns are excised. Some are predicted to be self-splicing introns, and some encode homing endonucleases (Laglidag 1–Laglidag 3) (Supplementary Figure 5 and Supplementary Data 3). Interestingly, we also found protein-coding genes in the 16S rRNA genes of DPANN from the '*Ca*.

Aenigmarchaeota', '*Ca*. Micrarchaeota', '*Ca*. Pacearchaeota' and '*Ca*. Woesearchaeota' phyla (Supplementary Figure 5, Supplementary Table 3 and Supplementary Methods). As for the CPR, the insertions are located in both variable and conserved regions of the 16S rRNA gene (Supplementary Figure 5). The presence of introns may reduce the frequency with which these CPR and DPANN organisms are detected in 16S rRNA genes surveys[3].

In our analysis, we discovered introns in tRNAs in some CPR organisms. We loosely grouped 24 studied cases into three types. First, classic self-splicing group I introns occur in various tRNAs of '*Ca*. Microgenomates', '*Ca*. Parcubacteria' and '*Ca*. Peregrinibacteria' (Fig. 8). The group I CPR tRNA introns are inserted in the typical position, one nucleotide away from the anticodon. Second, there are intron sequences with similar lengths to those of group I introns (15 examples) (Fig. 8; Supplementary Text and Supplementary Methods), but they were not recognized as group I introns by Rfam[94]. However, as they seem to share overall secondary structure with group I introns, we classify them as divergent group I. Notably, all occur in threonine tRNA with the GGT anticodon, and all terminate with guanine, as expected for group I introns. Interestingly, they are predicted to excise directly adjacent to the anticodon rather than one base pair away from it (however, excision directly after a thymine, the second T in the GTT anticodon, is common for group I introns). Divergent group I introns were found in CPR genomes from six different phyla in samples from geographically divergent locations and environment types (Supplementary Text). We compared the phylogeny of the threonine tRNAs (GGT) with that of the excised intron sequences and found them to be discordant, consistent with lateral transfer of the intron among widely divergent CPR bacteria. Finally, there are various small introns with defined secondary structure, some of which may be Y RNAs (Fig. 8). Excision from non-canonical locations was required to generate a plausible tRNA in a few cases.

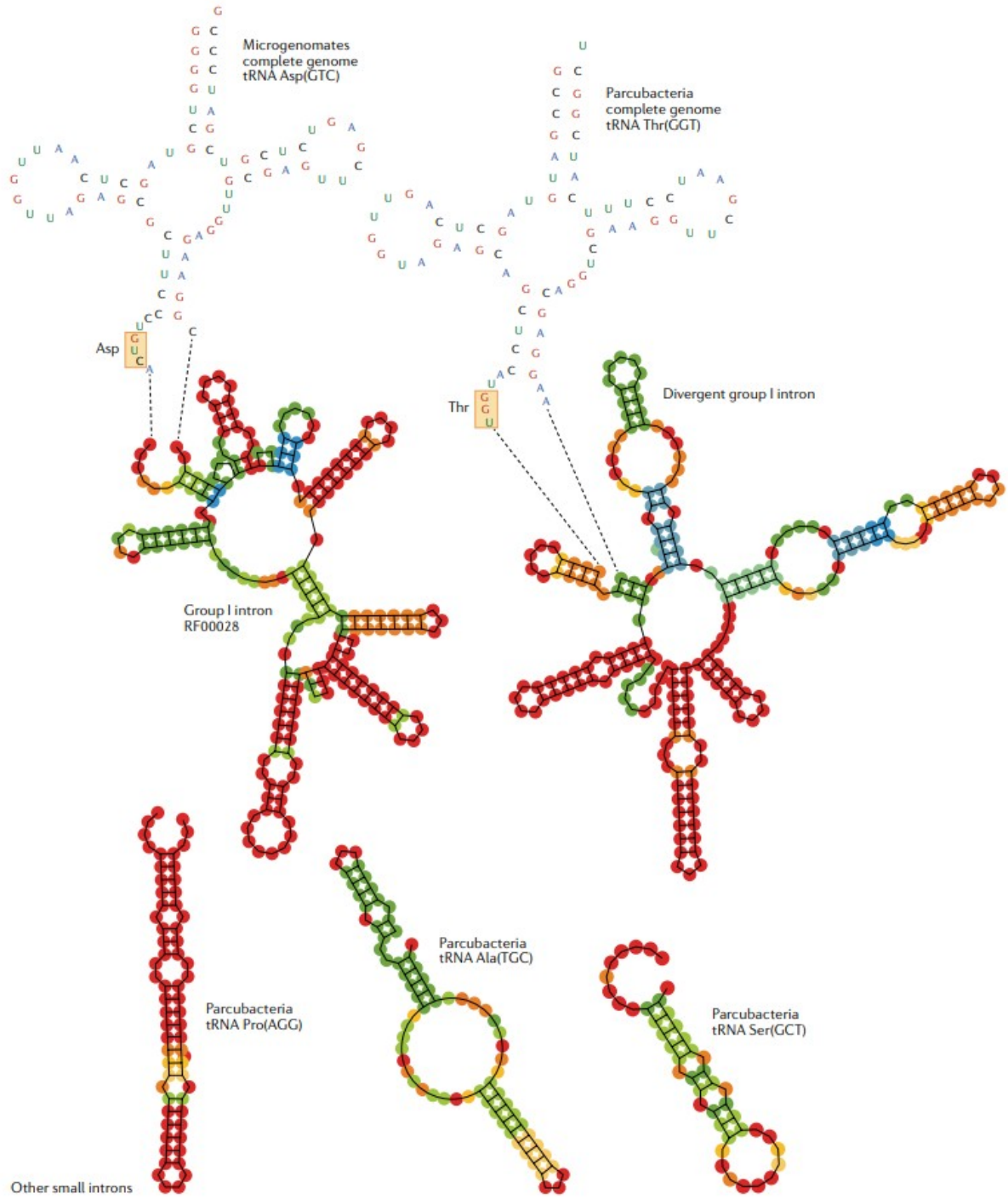

Fig. 8: Examples of tRNA introns identified in some CPR bacteria genomes.

The anticodons in Asp tRNA and Thr tRNA are indicated. Dashed lines link the excision sites to the intron with predicted secondary structure. Shown are a classic group I intron (from a '*Ca*. Microgenomates' genome), a newly reported divergent group I intron that always occurs in Thr tRNA (GGT) from a '*Ca*. Parcubacteria' genome and other small introns with predicted secondary structure. Intron structures are coloured by base pairing probabilities; for unpaired regions, the colour denotes the probability

of being unpaired. More detailed analysis is available in the Supplementary Text. CPR, candidate phyla radiation.

Some CPR and many DPANN genomes encode RtcB, a 3′-phosphate RNA ligase that is often involved in processing of tRNAs that contain introns[95]. RtcB occurs in some CPR (only Parcubacteria) and in most DPANN genomes. Interestingly, the protein lengths vary greatly. An alignment of the RtcB sequences revealed introns in at least four locations (Supplementary Figure 6). The phylogenetic tree for these intron-bearing RtcB sequences (Supplementary Figure 6) suggests lateral transfer of the intein across taxonomically divergent lineages. Another intein was found in one '*Candidatus* Jacksonbacteria' (Parcubacteria); the sequence is short and lacks an encoded endonuclease. The explanation for inteins within key metabolic genes is unclear, but intein excision is crucial for host gene function[96]. In combination, the findings for rRNA, tRNA and other genes indicate that intron proliferation in CPR genomes is a prominent feature. However, in comparison with eukaryotic genomes that tend to have many introns and low coding density, the coding density in CPR and DPANN is similar to that found in other bacteria and archaea (Supplementary Table 4).

Conclusions

In closing this Analysis, it is perhaps appropriate to address the question of why DPANN and CPR organisms should be important targets for future research. From the perspective of evolution, this is easily addressed: these lineages, in combination, may account for approximately half of all the archaeal and bacterial diversity of the planet. They are now recognized as members of natural microbial communities across a remarkable range of environment types (Supplementary Table 2), and yet their ecosystem importance remains uncertain. The vast majority of organisms from both radiations is predicted to live in some type of symbiosis or, at minimum, with dependence on other organisms for many basic metabolic requirements. In a few cases, CPR and DPANN organisms have eukaryotic hosts (protists, the dolphin, and human mouths and lungs[34,39,43,58,97]). In the deeper subsurface where Eukaryotes are rare or absent, the hosts are presumably bacteria and archaea. In many environments, the CPR and DPANN are likely to be episymbionts, as shown for *N. equitans* attached to *Ignicoccus hospitalis*[53], *Nanopusillus acidilobi* and its host *Acidilobus* sp. 7A[56], DPANN associated with Thermoplasmatales archaea[10,44,57] and '*Ca*. Saccharibacteria' on Actinobacteria[39]. In deep aquifers, communities featuring the association of symbionts and hosts may be well suited to stable, low-nutrient environments, where pore space volumes are limited and dispersal to potential host organisms is favoured if the cells are small. Under such conditions, CPR and DPANN organisms may contribute community-essential roles associated with recycling of microbial biomass from dead cells back to nutrients that are useful to other community members[11].

It is very likely that the association of CPR and DPANN organisms with other community members will have a great impact on the activities and metabolic capacities of the organisms that they depend on. Effects could arise from consumption of by-products that may alter the energetics of certain reactions, production of useful substrates such as hydrogen and small-chain fatty acids, competition for resources, provision of services such as phage defense[63] and pathogenic interactions. Owing to their ubiquity in the environment and through their community interactions, CPR and DPANN microorganisms are likely to affect human-relevant activities such as agriculture, water treatments and animal and human health as well as global biogeochemical cycles.

By exploration of the CPR and DPANN, researchers have uncovered a vast swath of new biology. The findings have motivated new renderings of the topology of the tree of life, revised our understanding of how metabolic capacities are distributed, expanded the range of symbiotic lifestyles and motivated new consideration of evolutionary and co-evolutionary processes. Much remains to be learned. This is especially apparent when it is realized that approximately half of the genes in the genomes of these organisms have no known function. These genes and pathways are likely to represent a substantial opportunity for future scientific discovery. If, as predicted, CPR and DPANN depend on other organisms, many of their unknown genes may be involved in organism–organism interactions, phenomena that are under-studied given the history of microbial research on organisms growing in pure cultures. Molecules involved in cell–cell interactions could have applied value, perhaps as antimicrobials or other pharmaceuticals. The biotechnological value of CPR genes has already been considered in the case of CRISPR–Cas systems, which in CPR are mostly novel. Many of the first-described Cpf1 family proteins[98] and the recently described CasY[99] were discovered in CPR genomes. Likely because they evolved in small genomes, both Cpf1 and CasY are moderately compact type II systems. Consequently, they are of considerable interest for genome editing[99]. It is certain that many new and interesting discoveries will follow as we learn more about CPR bacteria and DPANN archaea.

References

1. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437 (2013).

2. Castelle, C. J. et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. 25, 690–701 (2015).

3. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208–211 (2015).

4. Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).

5. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. 7, 13219 (2016).

6. Wrighton, K. C. et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337, 1661–1665 (2012).

7. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2, 1533–1542 (2017).

8. Huber, H. et al. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417, 63–67 (2002).

9. Baker, B. J. et al. Lineages of acidophilic archaea revealed by community genomic analysis. Science 314, 1933–1935 (2006).

10. Baker, B. J. et al. Enigmatic, ultrasmall, uncultivated Archaea. Proc. Natl Acad. Sci. USA 107, 8806–8811 (2010).

11. Probst, A. J. et al. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nat. Microbiol. 3, 328–336 (2018).

12. Probst, A. J. et al. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. Environ. Microbiol. 19, 459–474 (2017).

13. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl Acad. Sci. USA 114, E4602–E4611 (2017).

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

15. Aouad, M. et al. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. Mol. Phylogenet. Evol. 127, 46–54 (2018).

16. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. Genome Biol. Evol. 7, 191–204 (2014).

17. Hua, Z.-S. et al. Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. ISME J. 9, 1280–1294 (2015).

18. Suzuki, S. et al. Microbial diversity in The Cedars, an ultrabasic, ultrareducing, and low salinity serpentinizing ecosystem. Proc. Natl Acad. Sci. USA 110, 15336–15341 (2013).

19. Narasingarao, P. et al. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. 6, 81–93 (2012).

20. Andrade, K. et al. Metagenomic and lipid analyses reveal a diel cycle in a hypersaline microbial ecosystem. ISME J. 9, 2697–2711 (2015).

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

22. Linz, A. M. et al. Bacterial community composition and dynamics spanning five years in freshwater bog lakes. mSphere 2, e00169–00117 (2017).

23. Merkley, E. D. et al. Changes in protein expression across laboratory and field experiments in Geobacter bemidjiensis. J. Proteome Res. 14, 1361–1375 (2015).

24. Ludington, W. B. et al. Assessing biosynthetic potential of agricultural groundwater through metagenomic sequencing: a diverse anammox community dominates nitrate-rich groundwater. PLOS One 12, e0174930 (2017).

25. Lin, X., Kennedy, D., Fredrickson, J., Bjornstad, B. & Konopka, A. Vertical stratification of subsurface microbial community composition across geological formations at the Hanford Site. Environ. Microbiol. 14, 414–425 (2012).

26. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data 5, 170203 (2018).

27. Wright, J. J., Konwar, K. M. & Hallam, S. J. Microbial ecology of expanding oxygen minimum zones. Nat. Rev. Microbiol. 10, 381–394 (2012).

28. Li, M. et al. Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. Nat. Commun. 6, 8933 (2015).

29. Dombrowski, N., Seitz, K. W., Teske, A. P. & Baker, B. J. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. Microbiome 5, 106 (2017).

30. Schauer, C., Thompson, C. L. & Brune, A. The bacterial community in the gut of the Cockroach Shelfordella lateralis reflects the close evolutionary relatedness of cockroaches and termites. Appl. Environ. Microbiol. 78, 2758–2767 (2012).

31. Biedermann, L. et al. Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. PLOS One 8, e59260 (2013).

32. Camanocha, A. & Dewhirst, F. E. Host-associated bacterial taxa from Chlorobi, Chloroflexi, GN02, Synergistetes, SR1, TM7, and WPS-2 Phyla/candidate divisions. J. Oral Microbiol. 6, 25468 (2014).

33. Thomas, T. et al. Diversity, structure and convergent evolution of the global sponge microbiome. Nat. Commun. 7, 11870 (2016).

34. Koskinen, K. et al. First insights into the diverse human archaeome: specific detection of archaea in the gastrointestinal tract, lung, and nose and on skin. MBio 8, e00824–00817 (2017).

35. Bruno, A. et al. Exploring the under-investigated “microbial dark matter” of drinking water treatment plants. Sci. Rep. 7, 44350 (2017).

36. Bautista-de los Santos, Q. M. et al. Emerging investigators series: microbial communities in full-scale drinking water distribution systems — a meta-analysis. Environ. Sci. Water Res. Technol. 2, 631–644 (2016).

37. Pinto, A. J., Schroeder, J., Lunn, M., Sloan, W. & Raskin, L. Spatial-temporal survey and occupancyabundance modeling to predict bacterial community dynamics in the drinking water microbiome. MBio 5, e01135–01114 (2014).

38. Dewhirst, F. E. et al. The human oral microbiome. J. Bacteriol. 192, 5002–5017 (2010).

39. He, X. et al. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc. Natl Acad. Sci. USA 112, 244–249 (2015).

40. Ling, Z. et al. Altered fecal microbiota composition associated with food allergy in infants. Appl. Environ. Microbiol. 80, 2546–2554 (2014).

41. Kuehbacher, T. et al. Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease. J. Med. Microbiol. 57, 1569–1576 (2008).

42. Kowarsky, M. et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. Proc. Natl Acad. Sci. USA 114, 9623–9628 (2017).

43. Dudek, N. K. et al. Novel microbial diversity and functional potential in the marine mammal oral microbiome. Curr. Biol. 27, 3752–3762 (2017).

44. Golyshina, O. V. et al. ‘ARMAN’ archaea depend on association with euryarchaeal host in culture and in situ. Nat. Commun. 8, 60 (2017).

45. Youssef, N. H. et al. Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum ‘Diapherotrites’. ISME J. 9, 447–460 (2014).

46. Nelson, W. C. & Stegen, J. C. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. Front. Microbiol. 6, 713 (2015).

47. Kantor, R. S. et al. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. MBio 4, e00708–e00713 (2013).

48. Wrighton, K. C. et al. Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. ISME J. 8, 1452–1463 (2014).

49. Campbell, J. H. et al. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc. Natl Acad. Sci. USA 110, 5540–5545 (2013).

50. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. 31, 533–538 (2013).

51. Erickson, H. P. & Osawa, M. Cell division without FtsZ — a variety of redundant mechanisms. Mol. Microbiol. 78, 267–270 (2010).

52. Luef, B. et al. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat. Commun. 6, 6372 (2015).

53. Huber, H., Hohn, M. J., Rachel, R. & Fuchs, T. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. 417, 63–67 (2002).

54. Junglas, B. et al. Ignicoccus hospitalis and Nanoarchaeum equitans: ultrastructure, cell-cell interaction, and 3D reconstruction from serial sections of freeze-substituted cells and by electron cryotomography. Arch. Microbiol. 190, 395–408 (2008).

55. Burghardt, T. et al. The interaction of Nanoarchaeum equitans with Ignicoccus hospitalis: proteins in the contact site between two cells. Biochem. Soc. Trans. 37, 127–132 (2009).

56. Wurch, L. et al. ARTICLE Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. Nat. Commun. 7, 12115 (2016).

57. Comolli, L. R. & Banfield, J. F. Inter-species interconnections in acid mine drainage microbial communities. Front. Microbiol. 5, 367 (2014).

58. Soro, V. et al. Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria. Appl. Environ. Microbiol. 80, 6480–6489 (2014).

59. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. Nat. Biotechnol. 34, 1256–1263 (2016).

60. Moran, N. A. & Wernegreen, J. J. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. 15, 321–326 (2000).

61. Moran, N. A. & Bennett, G. M. The tiniest tiny genomes. Annu. Rev. Microbiol. 68, 195–215 (2014).

62. Lenhart, J. S. et al. RecO and RecR are necessary for RecA loading in response to DNA damage and replication fork stress. J. Bacteriol. 196, 2851–2860 (2014).

63. Burstein, D. et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. 7, 10613 (2016).

64. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. Nat. Rev. Microbiol. 2, 241–249 (2004).

65. Brasen, C., Esser, D., Rauch, B. & Siebers, B. Carbohydrate metabolism in Archaea: current insights into unusual enzymes and pathways and their regulation. Microbiol. Mol. Biol. Rev. 78, 89–175 (2014).

66. Silva, P. J. et al. Enzymes of hydrogen metabolism in Pyrococcus furiosus. Eur. J. Biochem. 267, 6541–6551 (2000).

67. Wrighton, K. C. et al. RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. ISME J. 10, 2702–2714 (2016).

68. Aono, R. et al. Enzymatic characterization of AMP phosphorylase and ribose-1,5-bisphosphate isomerase functioning in an archaeal AMP metabolic pathway. J. Bacteriol. 194, 6847–6855 (2012).

69. Hernsdorf, A. W. et al. Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. ISME J. 11, 1915–1929 (2017).

70. Anantharaman, K. et al. Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. PeerJ 4, e1607 (2016).

71. Castelle, C. J., Brown, C. T., Thomas, B. C., Williams, K. H. & Banfield, J. F. Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the Candidate Phyla Radiation. Sci. Rep. 7, 40101 (2017).

72. Danczak, R. E. et al. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome 5, 112 (2017).

73. León-Zayas, R. et al. The metabolic potential of the single cell genomes obtained from the Challenger Deep, Mariana Trench within the Candidate Superphylum Parcubacteria (OD1). Environ. Microbiol. 19, 2769–2784 (2017).

74. Coursolle, D. & Gralnick, J. A. Reconstruction of extracellular respiratory pathways for iron(III) reduction in Shewanella Oneidensis strain MR-1. Front. Microbiol. 3, 56 (2012).

75. Liu, J. et al. Identification and characterization of MtoA: a Decaheme c-Type cytochrome of the neutrophilic Fe(II)-oxidizing bacterium Sideroxydans lithotrophicus ES-1. Front. Microbiol. 3, 37 (2012).

76. Jiao, Y. & Newman, D. K. The pio operon is essential for phototrophic Fe(II) oxidation in Rhodopseudomonas palustris TIE-1. J. Bacteriol. 189, 1765–1773 (2007).

77. Yan, Z, Wang, M. & Ferry, J. G. A ferredoxin-and F 420 H 2-dependent, electron-bifurcating, heterodisulfide reductase with homologs in the domains Bacteria and Archaea. 8, 2285–2301 (2017).

78. Yan, Z. & Ferry, J. G. Electron bifurcation and confurcation in methanogenesis and reverse methanogenesis. Front. Microbiol. 9, 1322 (2018).

79. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the Tree of Life. Cell 172, 1181–1197 (2018).

80. Gross, C. A. et al. The functional and regulatory roles of sigma factors in transcription. Cold Spring Harb. Symp. Quant. Biol. 63, 141–155 (1998).

81. Paget, M. Bacterial sigma factors and anti-sigma factors: structure, function and distribution. Biomolecules 5, 1245–1265 (2015).

82. Merrick, M. J. In a class of its own—the RNA polymerase sigma factor sigma 54 (sigma N). Mol. Microbiol. 10, 903–909 (1993).

83. Mukai, T., Reynolds, N., Crnkovic´, A. & Söll, D. Bioinformatic analysis reveals archaeal tRNATyr and tRNATrp identities in bacteria. Life 7, 8 (2017).

84. Armengod, M.-E. et al. Enzymology of tRNA modification in the bacterial MnmEG pathway. Biochimie 94, 1510–1520 (2012).

85. Selvadurai, K., Wang, P., Seimetz, J. & Huang, R. H. Archaeal Elp3 catalyzes tRNA wobble uridine modification at C5 via a radical mechanism. Nat. Chem. Biol. 10, 810–812 (2014).

86. Krogan, N. J. & Greenblatt, J. F. Characterization of a six-subunit holo-elongator complex required for the regulated expression of a group of genes in Saccharomyces cerevisiae. Mol. Cell. Biol. 21, 8203–8212 (2001).

87. Glatt, S. et al. Structural basis for tRNA modification by Elp3 from Dehalococcoides mccartyi. Nat. Struct. Mol. Biol. 23, 794–802 (2016).

88. Guo, H., Arambula, D., Ghosh, P. & Miller, J. F. Diversity-generating retroelements in phage and bacterial genomes. Microbiol. Spectr. https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014 (2014).

89. Paul, B. G. et al. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. Nat. Commun. 6, 6585 (2015).

90. Paul, B. G. et al. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat. Microbiol. 2, 17045 (2017).

91. Salman, V., Amann, R., Shub, D. A. & Schulz-Vogt, H. N. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. Proc. Natl Acad. Sci. USA 109, 4203–4208 (2012).

92. Baker, B. J., Hugenholtz, P., Dawson, S. C. & Banfield, J. F. Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. Appl. Environ. Microbiol. 69, 5512–5518 (2003).

93. Jay, Z. J. & Inskeep, W. P. The distribution, diversity, and importance of 16S rRNA gene introns in the order Thermoproteales. Biol. Direct 10, 35 (2015).

94. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. Nucleic Acids Res. 31, 439–441 (2003).

95. Tanaka, N., Meineke, B. & Shuman, S. RtcB, a novel RNA ligase, can catalyze tRNA splicing and HAC1 mRNA splicing in vivo. J. Biol. Chem. 286, 30253–30257 (2011).

96. Shah, N. H. & Muir, T. W. Inteins: nature’s gift to protein chemists. Chem. Sci. 5, 446–461 (2014).

97. Gong, J., Qing, Y., Guo, X. & Warren, A. ‘Candidatus Sonnebornia yantaiensis’, a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora. Oligohymenophorea). Syst. Appl. Microbiol. 37, 35–41 (2014).

98. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163, 759–771 (2015).

99. Burstein, D. et al. New CRISPR–Cas systems from uncultivated microbes. Nature 542, 237–241 (2016).

100. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).