Title
Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life.

Permalink
https://escholarship.org/uc/item/0299z0z4

Journal
Cell, 172(6)

ISSN
0092-8674

Authors
Castelle, Cindy J
Banfield, Jillian F

Publication Date
2018-03-01

DOI
10.1016/j.cell.2018.02.016

Peer reviewed
Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life

Cindy J. Castelle\textsuperscript{1,2,3} and Jillian F. Banfield\textsuperscript{1,2,3,4,5,6,*}

\textsuperscript{1}Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA
\textsuperscript{2}Innovative Genomics Institute, Berkeley, CA, USA
\textsuperscript{3}Chan Zuckerberg Biohub, San Francisco, CA, USA
\textsuperscript{4}University of Melbourne, Melbourne, VIC, Australia
\textsuperscript{5}Lawrence Berkeley National Laboratory, Berkeley, CA, USA
\textsuperscript{6}Department of Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, CA, USA

*Correspondence: jbanfield@berkeley.edu

The recent recovery of genomes for organisms from phyla with no isolated representative (candidate phyla) via cultivation-independent genomics enabled delineation of major new microbial lineages, namely the bacterial candidate phyla radiation (CPR), DPANN archaea, and Asgard archaea. CPR and DPANN organisms are inferred to be mostly symbionts, and some are episymbionts of other microbial community members. Asgard genomes encode typically eukaryotic systems, and their inclusion in phylogenetic analyses results in placement of eukaryotes as a branch within Archaea. Here, we illustrate how new genomes have changed the structure of the tree of life and altered our understanding of biology, evolution, and metabolic roles in biogeochemical processes.

Introduction

The tree of life is arguably the most important organizing principle in biology and perhaps the most widely understood depiction of the evolutionary process. It explains to us how we are related to other organisms and where we may have come from. The tree has undergone some tremendous revolutions since the first version was sketched by Charles Darwin. A major innovation was the construction of phylogenetic trees using DNA sequence information, which opened the way for classification of microbial life. As implemented by Carl Woese and collaborators, this work enabled the definition of three domains: Bacteria, Archaea, and Eukaryotes (Woese and Fox, 1977, Woese et al., 1990). More recently, the three-domain topology has been questioned, and eukaryotes—our own branch of life—potentially relocated into the archael domain (Spang et al., 2015, Williams et al., 2013). Beyond this, and as described here, cultivation-independent genomic methods that access sequences from laboratory-intractable organisms have added many new lineages to the tree. Their inclusion completely clarifies the extreme minority of life’s diversity that is represented by macroscopic organisms and underscores that our place in biology is dwarfed by bacteria and archaea.

Genome Resolution of Uncultivated Organisms
Cultivation-independent (e.g., 16S rRNA gene-based) survey methods uncovered evidence for vast microbial diversity that was not represented in the set of organisms available in pure culture (e.g., Dojka et al., 2000, Pace, 1997). This motivated the development of new genomic approaches that can provide comprehensive metabolic insights without the requirement for laboratory growth. The first of these is referred to as community genomics or, more commonly, metagenomics. As metagenomics is sometimes considered to include rRNA gene surveys, we adopt the term “genome-resolved metagenomics” to describe the specific approach that can yield direct information about metabolic capacity on an organism-by-organism basis. In genome-resolved metagenomics, DNA is extracted from a whole community or an enrichment and shotgun sequenced, and the short sequences are assembled into larger genome fragments that are ultimately assigned to genome bins (draft genomes; Tyson et al., 2004). Typically, metabolic predictions are only undertaken for reasonably high-quality genomes (>70% complete with low contamination by fragments from other organisms). In a few cases, further assembly curation has generated complete (closed) genomes (e.g., Albertsen et al., 2013, Kantor et al., 2013). Importantly, the approach is not limited to Bacteria and Archaea but can provide draft genomes for Eukaryotes and partial or complete genomes of phage, viruses, and plasmids (e.g., Paez-Espino et al., 2016). The disadvantage of this method is that the sampled cells comprise a natural population. Thus, the genome is, to differing extents, a composite of sequence variants distinguished by single-nucleotide polymorphisms and insertions/deletions. On the other hand, the reads can be mapped back to the genome sequence to provide a snapshot of the form and extent of variation within the population (Simmons et al., 2008).

The second method, single-cell genomics, involves the sequencing of fragments of DNA amplified from a single cell or collection of cells with similar rRNA gene sequences (e.g., Podar et al., 2007). The method has been especially useful for study of eukaryotic cells (Gawad et al., 2016). The advantage of a method in which a single microbial cell is sequenced is that all reported nucleotides derive from one genome so that linkage patterns can be established; however, binning may be required to remove contaminating sequences. A comparative study that targeted both methods to one microbial community revealed that the single-cell method is of much lower throughput than genome-resolved metagenomics, and the genomes are, on average, significantly less complete. However, the sequences from single-cell genomics and metagenomics analyses generally agree well (Probst et al., 2018). Single-cell sequencing can be implemented in a targeted way following screening of rRNA genes (Rinke et al., 2013). Thus, the resulting sequences directly augment the collection of genomes for organisms belonging to lineages that are unsampled or undersampled.

An Updated Topology of the Tree of Life
Hug et al. (2016) presented a new version of the tree of life that captured information for organisms whose draft genomes had been reconstructed without prior cultivation. Although the deep branching order is not well supported, it is possible that the root of the bacterial domain is placed between well-known bacteria (including the commonly known major lineages, such as Proteobacteria, Actinobacteria, Firmicutes, and Cyanobacteria, with well-studied representatives) and the bacteria of a recently described and seemingly monophyletic group referred to as the candidate phyla radiation (CPR) (Brown et al., 2015). Recent detailed phylogenetic analyses have considered the placement of the DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota), a newly recognized superphylum that is described in more detail below, within the archaeal domain (Williams et al., 2017). To root the archaeal tree, Williams et al. applied a new approach that is based on single-gene tree reconciliation with no need for the use of an outgroup. They proposed that the root of the archaeal domain is placed between the all other archaea and the DPANN. Their analyses suggest the monophyly of the DPANN superphylum and the Euryarchaeota phylum. However, due to low sampling of the DPANN superphylum at the time of their analysis, new sequences and approaches are needed to further evaluate this conclusion. Sequencing and bioinformatics methods are evolving rapidly, and new genomes have recently become available. Figure 1 presents an updated view of the three-domain tree that includes new sequences from DPANN archaea.
Figure 1. Phylogenetic Tree

(A) Updated from Hug et al. (2016) by addition of new DPANN archaeal sequences (many but not all phyla are indicated on the tree). Note that with the use of a significant number of bacterial sequences in a three-domain tree topology (Bacteria, Archaea, and Eukaryotes), the DPANN superphylum do not form a monophyletic clade. As previously discussed, the placement of nanosized lineages and the monophyly of the DPANN within the tree of life remain unclear and important open issues that need to be addressed with the development of new methodological approaches.

(B) We reconstructed a phylogenetic tree with only the archaeal sequences (with bacteria as an outgroup) in which the monophyly of the DPANN clade is retrieved but the branching order of the deepest branches cannot be confidently resolved. Bootstrap support values in are indicated by circles on nodes (black for support from 89% and above and gray for support from 50% to 89%).

The trees were calculated using a maximum-likelihood algorithm (RAxML, with PROTCATLG model) based on genome sequences containing 14 ribosomal proteins (45 ribosomal proteins L5, L13, L4, L5, L6, L14, L15, L18, L22, L24, S3, S8, S17, and S19). The concatenated ribosomal protein alignment was constructed as described previously (Hug et al., 2016). The updated tree of life is available with full bootstrap values in Newick format in Data S1.
The extent of diversity within the CPR is a topic of current debate. Analyses based on 16S rRNA or concatenation of ribosomal proteins suggest that the CPR may be comparable in breadth to all other bacteria (Hug et al., 2016). However, it remains uncertain whether this is due to early divergence or rapid evolution (or both). Other estimates of the scale of the CPR predict that they could comprise a maximum of 26% (Parks et al., 2017) or 25% (Schulz et al., 2017) of bacterial diversity. A reanalysis of this question using the full set of available genomes, genes universal to both groups, and improved phylogenetic methods is required.

After the explosive growth in the number of major branches either discovered or genomically resolved via cultivation-independent methods between 2012 and 2016, one may wonder if the appearance of new phylum-level groups will continue indefinitely. An analysis of the rate of discovery of new groups would suggest that this is not the case. Around 8,000 genomes were reconstructed de novo by exploiting the public short-read archive (circa mid-2016). The archive includes samples from a huge array of environment types (Parks et al., 2017). The majority of genomes were for bacteria and archaea from clades already known to exist. A subset was assigned to 44 putative phyla that were defined by cultivation-independent genomic methods over the past 5 years. 3 potentially additional archaeal phyla and 17 possible bacterial phyla were identified (although some of these may have been independently described in the period between download of the archive and publication). In fact, 2017 saw reports of the first genomic sampling of several new groups, many of which are in the previously undersampled archaeal domain (Table S1). Once found, the major phylum-level lineages of bacteria and archaea are often identified across multiple ecosystems (see below). We suspect that the phylum content of the bacterial and archaeal domains will soon approach saturation, although the topology of the tree may change as algorithms to analyze it improve. The scope for discovery of new classes and groups at finer scales of taxonomic resolution is immense.

If we are approaching full delineation of the major branches of the tree of life, one may wonder why it took so long and then happened so quickly. It is our perspective that this is a reflection of relatively little prior focus on subsurface environments (which constitute a huge part of the biosphere but are difficult to sample) and challenges associated with the application of cultivation-independent genomic methods to the most complicated ecosystems (e.g., soil). Many subsurface environments are anaerobic, and therefore may be refuges for organisms from lineages that diverged early from primitive life forms. These environments are also enriched for organisms not commonly associated with humans, animals, or crops (thus viewed as lower-priority research targets). Some subsurface environments are low in nutrients and thus harbor slow-growing organisms with complex interdependencies that ensure retention of resources within the ecosystem (Probst et al., 2018). Overall, these more recently studied environments are
enriched in organisms that have been difficult (or impossible) to cultivate, so their discovery and genomic characterization awaited the development of community-wide cultivation-independent methods.

**New Phyla and Phylum Radiations**

16S rRNA gene surveys performed by Norman Pace’s group uncovered sequences from organisms from a “candidate division” (analogous to a candidate phylum) labeled OP11 that were detected in hot springs of Yellowstone National Park (Hugenholtz et al., 1998). As the diversity of sequences expanded, it was apparent that the OP11 radiation includes multiple potentially phylum-level groups, including the OD1 superphylum (Harris et al., 2004) now referred to as Parcubacteria (Rinke et al., 2013) and the OP11 superphylum now referred to as Microgenomates (Rinke et al., 2013). Metagenomics-derived rRNA and protein sequences grouped these into the CPR, and many new major lineages were added (Brown et al., 2015). As discussed below, bacteria of the CPR consistently have small genomes and cell sizes, and most are predicted to have symbiotic lifestyles. Importantly, some groups within the CPR are not detected in rRNA gene surveys due to primer mismatch, introns, or both (Brown et al., 2015). Currently, we estimate that 73 CPR groups have been identified (Figure 1 and Table S1).

Rinke et al. (2013) proposed the term Patescibacteria for a superphylum comprising Microgenomates (OP11), Parcubacteria (OD1), and Gracilibacteria (GNO2-BD1-5), but the meaning of the term has become confused and even misused as synonymous with the CPR (Parks et al., 2017). The term CPR is simply a description of a huge monophyletic radiation of phyla and superphyla that includes the group that ultimately may be referred to as Patscibacteria and dozens of additional phyla and superphyla. Community-wide consultation should be undertaken before a formal name for this radiation is proposed.

It should be noted that a widely accepted definition of what constitutes a superphylum or even a new phylum is not available. In our experience, new phylum-level branches are distinct, apparently monophyletic, and have 16S rRNA genes that share <75%-80% identity with the most closely related groups. Ultimately, appropriate phylum and superphylum definitions require strong evidence for deep branch placements, and this is not yet available.

Based on the current summary, 62 non-CPR bacterial phylum-level groups have representatives that have been genomically described via cultivation-independent methods (Table S1). Given that around 30 traditionally have been described bacterial phyla (a handful of which were candidate phyla that now have recently isolated representatives), the scope of the bacterial domain exclusive of CPR may have approximately tripled. When CPR and non-CPR groups are taken together, we estimate that cultivation-independent genomic methods have expanded the scale of domain Bacteria by over five times.
There were early indications for the existence of a group of novel archaea potentially analogous to the CPR organisms. These include the co-cultivation of Nanoarchaeum equitans with its host (Huber et al., 2002), genomic description and imaging of what were initially referred to as archaeal Richmond Mine acidophilic nanoorganisms (ARMAN) archaea (Baker et al., 2006, Baker et al., 2010), and the genomic characterization of extremely salt-adapted nanohaloarchaea (Narasingarao et al., 2012). Based on a collection of draft genomes from single cells (or concentrates), a segment of the archaeal domain referred to as DPANN was proposed (Rinke et al., 2013). Since then, new genomes and the definition of at least nine major groups have greatly expanded the DPANN lineages (Castelle et al., 2015).

Archaea now include at least three other major supergroups: the Euryarchaeota, the TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota; proposed name Eocyta; Lake, 2015), and Asgard archaea, all of which comprise potentially phylum- or superphylum-level clades (see below; Figure 1). In addition, 6 potential phylum-level groups have been proposed within the TACK superphylum, and 15 additional groups affiliate approximately with the Euryarchaeota. Although the phylum-level delineation of Archaea is especially complicated (Adam et al., 2017, Spang et al., 2017), it appears that new archaeal candidate phyla (35 groups listed in Table S1) expand the scope of archaeal diversity by a factor similar to that noted for bacteria.

New Discoveries Regarding Potential Roles of Non-DPANN Archaea in Biogeochemical Cycles

Genomic data for uncultivated archaea resulted in key discoveries related to their roles in the carbon, nitrogen, and sulfur geochemical cycles. Excellent recent reviews have been published on the phylogeny and ecological roles of archaea in diverse ecosystems (Adam et al., 2017, Spang et al., 2017). Thus, we focus here on a few examples related to ammonia and methane metabolism to illustrate the overarching principle that genome-resolved metagenomics can inform our understanding of evolution and biogeochemistry.

An important early result of metagenomics was the discovery of homologs of ammonia monooxygenase genes in archaeal genome fragments reconstructed from a Sargasso Sea metagenome (Venter et al., 2004). The existence of an unknown organism responsible for much of the ammonia oxidation in the ocean had been suspected based on in situ measurements that indicated that ammonia oxidation often proceeds at substrate concentrations significantly below the growth threshold of cultured ammonia-oxidizing bacteria. Subsequently, autotrophic ammonia-oxidizing archaea were isolated and their critical role in the global nitrogen cycle demonstrated (Pester et al., 2011, Stahl and de la Torre, 2012). In 2008, they were proposed as the third archaeal phylum and named Thaumarchaeota (Brochier-Armanet et al., 2008).
Sibling to Thaumarchaeota and Aigarchaeota are the Bathyarchaeota, which appear to be key players in the global carbon cycle in terrestrial and marine anoxic sediments. Some Bathyarchaeota possess the archaeal Wood-Ljungdahl pathway (He et al., 2016, Lazar et al., 2016), suggesting a capacity for CO₂ fixation via acetogenesis, a process thought to be unique to bacteria (He et al., 2016). Importantly, some Bathyarchaeota may be methanogens, given key genes that could produce methane from methanol, methyl sulfides, and methylated amines (Evans et al., 2015). Verstraetearchaeota (TACK superphylum) are also found in anoxic environments with high methane fluxes and, like some Bathyarchaeota, are predicted to conserve energy via methylotrophic methanogenesis (Vanwonterghem et al., 2016).

Metagenomics has been critical for the study of methane oxidation. Anaerobic methane-oxidizing euryarchaea-2d (ANME-2d) are often detected in the sulfate-methane transition zone in marine sediments and use a modified and reverse-methanogenesis pathway for growth (Hallam et al., 2004). Certain ANME-2d from granitic groundwater were incubated with ¹³C-labeled methane to demonstrate that methane oxidation is linked to microbial sulfate reduction (Ino et al., 2018). Others are predicted to conduct reverse methanogenesis using nitrate as the terminal electron acceptor, and this has been experimentally demonstrated using bioreactors and ¹³C and ¹⁵N labeling (Haroon et al., 2013). It has also been predicted that some ANME-2d oxidize methane using manganese (Beal et al., 2009) or iron (Hernsdorf et al., 2017) as the electron acceptor. The genomes of anaerobic Methanomassiliicoccales also encode key enzymes involved in methyl-dependent methanogenesis, unlike other members of Thermoplasmata-related lineages (Borrel et al., 2014). WSA2/Arc1, sibling to both the non-methanogenic Hadesarchaeota (Baker et al., 2016) and MSBL1 (Mwirichia et al., 2016), may be methanogens and were recently named “Candidatus Methanofastidiosa.” Based on metabolic predictions from genome sequences from a wastewater treatment bioreactor (Nobu et al., 2016), they lack pathways for CO₂-reducing and aceticlastic methanogenesis, but they may be capable of methane formation via methylated thiol reduction with H₂.

Overall, recent metagenomic studies have substantially expanded the diversity of archaea predicted to be involved in methane metabolism. The inference that Bathyarchaeota and Verstraetearchaeota, members of the TACK superphylum, are methanogens is important because previously, methanogenesis was unknown outside of the Euryarchaeota. New groups such as the Methanomassiliicoccales and Methanofastidiosida also expand the diversity of Euryarchaeotes predicted to be involved in methanogenesis. However, it should be noted that the genes for methanogenesis and methane oxidation via reverse methanogenesis are essentially the same. This observation underlines the importance of experimental (e.g., bioreactor) studies to appropriately define metabolism in situ.

Discoveries Regarding DPANN Archaea
Understanding of the diversity, distribution, and metabolisms of nano-sized archaea that are members of the DPANN superphylum has been substantially advanced via cultivation-independent genomics studies. DPANN archaea are found in various extreme (e.g., hot, acidic, hypersaline) and temperate (e.g., lake and marine sediments) ecosystems, although they have been relatively rarely detected in soil and the open ocean. The first ARMAN archaeal groups were discovered in acid mine drainage biofilms via genomics-based approaches (Baker et al., 2006, Baker et al., 2010). The coverage of these groups and the diversity of the environments in which they occur has expanded (Chen et al., 2017), and the phyla have been renamed as Parvarchaeota and Micrarchaeota (Baker and Dick, 2013, Castelle et al., 2015). Both lineages are found in association with other archaea (see below). Single-cell genomics and metagenomics uncovered and defined additional phylum-level lineages within the DPANN superphylum. These include Diapherotrites and Aenigmarchaeota (Rinke et al., 2013), Nanoarchaeota (which includes the previously described N. equitans; Huber et al., 2002), Nanohaloarchaeota (Narasingarao et al., 2012), Woesearchaeota, and Pacearchaeota (Castelle et al., 2015). Common features unifying DPANN archaea are their small genomes, small cell sizes and limited metabolic repertoires: many lack core biosynthetic pathways for nucleotides, amino acids, and lipids. Most DPANN archaea depend on other microbes to meet their biological requirements. However, some appear to have the genetic potential to be free living with a heterotrophic aerobic and/or fermentative lifestyle.

A potential phylum-level clade represented by the Altiarchaeales (formerly SM1; Probst et al., 2014) appears to branch deeply within the DPANN superphylum based on recent studies (Bird et al., 2016, Spang et al., 2017) and our analyses (Figure 1). However, this placement requires further analysis. Altiarchaeales dominate some cold subsurface anaerobic groundwater habitats (Probst et al., 2014). Unlike most archaea, they have an outer membrane and unique surface-attached grappling hooks known as hami. They form biofilms and appear to grow autotrophically on carbon monoxide, acetate, or formate (via a modified archaeal Wood-Ljungdahl pathway; Probst et al., 2014).

Some New Discoveries Regarding Potential Roles of Non-CPR Bacteria in Biogeochemical Cycles

Many bacterial groups have new genomically described representatives (Table S1), but it is beyond the scope of this Perspective to provide details for all of these. Many of the bacterial genomes used to define these putative candidate phyla were reported by nantharaman et al. (2016, who provided an extensive table in the Supplemental Information that predicts the biogeochemically relevant capacities for each genome. Here, we review a few select examples from the literature to illustrate the types of insights into ecological roles and evolutionary histories that have been obtained.
A previously unknown bacterium was found to dominate aquifer sediments. A complete curated genome for one population was reconstructed from a metagenome, and the lineage was named Zixibacteria (Castelle et al., 2013). Gene-by-gene analysis yielded a detailed metabolic prediction for this organism, an overview of which is presented in Figure 2. Notably, the genome encodes an extensive repertoire of redox enzymes that likely indicate roles in iron and arsenic oxidation/reduction, nitrogen compound transformations, hydrogen metabolism, and fermentation. This mixture of aerobic and anaerobic pathways likely confers metabolic versatility, enabling this bacterium to proliferate under changing conditions close to the water table (Castelle et al., 2013).

A long-standing mystery is related to the apparent detection of Cyanobacteria in human fecal samples, suggesting the unexpected existence of these bacteria in gut microbiomes. Adult fecal samples that
were relatively enriched in these organisms (based on 16S rRNA gene sequencing) were targeted by genome-resolved metagenomics methods, and a complete (closed) genome was reconstructed (Di Rienzi et al., 2013). The study included an additional genome for a distinct group that was recovered from an acetate-amended sediment metagenome (Wrighton et al., 2012). Metabolic predictions for these genomes indicated the absence of photosynthetic machinery, leading to the conclusion that these bacteria have a fermentation-based metabolism. Based on phylogenetic analyses, the bacteria were assigned to the candidate phylum Melainabacteria (Di Rienzi et al., 2013). Genomes from a distinct but related lineage, Sericytochromatia, were subsequently detected in a coal-bed methane well, a laboratory bioreactor biofilm, and acetate-amended sediments (Soo et al., 2017). Soo et al. propose both Melainabacteria and Sericytochromatia as classes of Cyanobacteria, a disagreement that reflects the common challenge of achieving consensus regarding taxonomic designations. Due to the absence of photosynthetic machinery in lineages sibling to Cyanobacteria, Soo et al. confirm and extend the earlier conclusion (Di Rienzi et al., 2013) that photosynthesis evolved after the divergence of Cyanobacteria. Further, their analyses suggest that aerobic respiration arose after the evolution of photosynthesis.

Another distinct bacterial group is the candidate phylum Rokubacteria (Hug et al., 2015), first genomically described via metagenomic analysis of sediments collected at the same site as Zixibacteria and some Melainabacteria. These bacteria had been previously detected in soil 16S rRNA gene surveys. Based on metabolic predictions, it was suggested that the sediment-associated Rokubacteria are acetoclastic heterotrophs that likely use beta-oxidation of fatty acids for energy generation and produce butyrate that would be consumed by other community members. These Rokubacteria were also predicted to contribute to sulfur cycling (via oxidation of thiosulfate to sulfide and its reduction to hydrogen sulfide) and nitrogen cycling (via nitrite oxidation) and can probably oxidize carbon monoxide. Subsequently, it was proposed that Rokubacteria, which are relatively abundant in grassland soil, play a key role in carbon turnover via methanol oxidation (Butterfield et al., 2016). Rokubacteria were recently described as “genomic giants,” detected in the rhizosphere, volcanic mud, oil wells, aquifers, and the deep subsurface (Becraft et al., 2017).

Another important discovery involves bacteria of the candidate phylum Tectobacteria. The first genomically described members belong to the candidate genus Entotheonella and were described using metagenomic and single-cell sequencing methods targeted at microbial communities in the marine sponge Theonella swinhoei (Wilson et al., 2014). Entotheonella are predicted to produce a huge variety of bioactive compounds that may mediate ecological interactions. As sponges are well known as rich sources of diverse natural products, the research likely addressed the question of which organisms are the source of these compounds. The results of this
study underline the potential of genome-resolved metagenomics targeted to candidate phyla groups to uncover biosynthetic pathways for a vast treasure trove of secondary metabolites that could address the pressing need for new antimicrobial compounds and other pharmaceuticals.

Bacteria of the CPR and Their Similarities to DPANN Archaea

Although the first draft genomes for CPR were first reported only in 2012 (Wrighton et al., 2012), there are now thousands of sequences on hand. Based on the first ~800 genomes, Brown et al. (2015) proposed the existence of at least 35 candidate phyla within the CPR. Parallel cryogenic transmission electron microscope images that targeted post-0.2 μm filtrates collected from the same samples verified small cell volumes for these bacteria (see below; Luef et al., 2015). Interestingly, some candidate phyla within the radiation (Gracilibacteria, BD1-5 and Absconditabacteria, SR1) use an alternative genetic code (Campbell et al., 2013, Kantor et al., 2013, Rinke et al., 2013, Wrighton et al., 2012). The repurposing of the UGA stop codon to code for glycine was confirmed for an enrichment that contained Gracilibacteria by metaproteomics (Hanke et al., 2014). Building upon prior work by Wrighton et al., 2012, Kantor et al., 2013, Brown et al., 2015, Nelson and Stegen, 2015, and others, nantharaman et al. (2016 extended the early observation that CPR genomes are small and that most lack numerous biosynthetic pathways (Figure 3). Many are predicted to be unable to produce nucleotides \textit{de novo} and have minimal amino acid and cofactor biosynthetic capacity. No CPR genomes analyzed to date contain the components necessary to synthesize membrane lipids required for the cell envelope, so further research is needed to determine the nature and sources of these components. The CPR bacteria have unusual ribosome compositions, and whole lineages (groups of putative candidate phyla) are missing what were considered to be universal ribosomal proteins (Brown et al., 2015). Parcubacteria have been reported to lack ribosomal small subunit methyltransferase G and ribosome-silencing factor (Nelson and Stegen, 2015).
In a new analysis reported here, we used a set of high-quality draft genomes and some complete genomes to provide a first glance into the metabolic variation within the CPR and DPANN simultaneously. As is evident from the descriptions above, limited metabolic capacities are common across both radiations and gaps in metabolic features are often shared (Figure 3). As noted previously, typically bacterial genes (e.g., bacterial transcription factors) occur in some DPANN archaea (Rinke et al., 2013). The distribution of isoprenoid-related genes is discussed further below.

Episymbiosis: An Unusual Lifestyle Predicted to Be Common for CPR and DPANN

A few CPR and DPANN species are associated with eukaryotic hosts (Gong et al., 2014). However, most are likely symbionts of bacteria or archaea, given their abundance and diversity in samples that have few, if any, eukaryotes. Based on strong enrichment in post-0.2 μm filtrates, it was predicted that many may be episymbionts—i.e., symbionts that associate with the surfaces of host cells rather than being contained within them. Cryogenic transmission electron microscopy (cryo-TEM) data show pili-like structures that extend from CPR cell surfaces (Figure 4A) and, in some cases, contact other microbial cells (Figures 4B and 4C; Luef et al., 2015). These
might provide access to nucleic acids or other metabolites. A few CPR and DPANN organisms have been directly shown to be episymbionts. For ARMAN (either Micrarchaeota or Parvarchaeota), cryoelectron tomographic data revealed penetration of their cell interiors via cytoplasmic extensions from larger cells without a cell wall (Figure 4C; Baker et al., 2010, Comolli and Banfield, 2014). Based on the overall microbial composition, the larger cells were identified as Thermoplasmatales archaea. In other cases, cells are attached via short bridges to Thermoplasmatales cells (Figure 4D)—probably Gplasma (Baker et al., 2010)—which were renamed Cuniculiplasma divulgatum by Golyshina et al., (2016). These points of contact resemble those associated with N. equitans cells attached to host Ig nicoccus hospitalis archaea in a co-culture of these organisms (Huber et al., 2002). Similarly, a co-culture of a parasitic Saccharibacterium (TM7) and its Actinomyces odontolyticus (Actinobacteria) host has been obtained from the human mouth (He et al., 2015). Transmission electron microscope images indicate host-episymbiont interaction via a region of cell-cell contact similar in form to those seen for ARMAN archaea and host Thermoplasmatales.
Isoprenoid Lipid Biosynthesis Pathways Blur a Historical Distinction of Archaea and Bacteria

Figure 4. Cryo-TEM Images of CPR and DPANN Illustrating Close Association with Other Cells
Images of CPR modified from Luef et al. (2015) (A–C) and of DPANN from Baker et al. (2010) and Cornill and Banfield (2014) (D and E).
(A) Pili-like structures cover the surface of a CPR cell.
(B) Pili-like structures apparently connecting to the surface of an adjacent bacterium.
(C) Ultra-small cell with a pilus-like structure that extends to the surface of another large cell that is covered with nanoparticle aggregates.
(D) Interaction between DPANN (ARMAN) archaea and a Thermoplasmatales archaeon. Also shown is the association of viruses with an ARMAN cell (lower left and right).
(E) A Thermoplasmatales archaeon with two DPANN cells undergoing cell division.
Isoprenoids are metabolites that are essential in all living organisms in the three domains of life. Isoprenoids have diverse metabolic functions, including as quinones, chlorophylls, bacteriochlorophylls, rhodopsins, and carotenoids. Archaeal membranes are composed of isoprenoid-based lipids, the isopentenyl pyrophosphate and dimethylallyl diphosphate precursors of which are typically synthesized via the mevalonate (MVA) pathway. Bacteria use a nonhomologous pathway—the methylerythritol phosphate (MEP) pathway—to synthesize the precursors. Previously, only few bacteria (mostly Gram positive) were known to possess the MVA pathway (Pasternak et al., 2013). Interestingly, most of these bacteria predate other bacteria. In new analyses, we revisited the question of the distribution of the MVA pathway in bacteria, making use of hundreds of new genomes from the CPR and DPANN radiation. The results show that some CPR bacteria (mostly Microgenomates and a few Peregrinibacteria) possess the MVA pathway rather than the bacterial MEP pathway (Figure 3). Intriguingly, the MVA pathway of the CPR is of the type found in eukaryotes. The eukaryote MVA pathway includes three enzymes that do not normally occur in the archaeal MVA pathway (phosphomevalonate kinase [PMK], diphosphomevalonate carboxylase [MDC], and isopentenyl diphosphate isomerase [IDI1]).

As expected for Archaea, the MVA pathway occurs in organisms from several DPANN phyla, including Diapherotrites, Micrarchaeota, and Aenigmarchaeota. Similar to the CPR and distinct from other archaea, the DPANN MVA pathway includes the full set of enzymes found in eukaryotes. The MVA pathway is rarely found in Woesearchaeota. Instead, they have the bacterial MEP pathway. This is important because the MEP pathway has not been reported in archaea.

Previously, the presence of the MVA pathway in a few bacteria was explained by potential acquisition via horizontal gene transfer from archaeal or eukaryotic donors (Boucher and Doolittle, 2000, Wilding et al., 2000). Later, in-depth phylogenomic analyses indicated that the MVA pathway may have been ancestral in all three domains of life (Lombard and Moreira, 2011). We conducted phylogenetic analyses using two key enzymes (MVA kinase and PMK) from the MVA pathways to evaluate whether the new CPR and DPANN sequences throw light on the evolutionary origin of isoprenoid biosynthesis (Figure 5). The resulting phylogenetic tree resolves a novel intermediate clade for members of the Microgenomates superphylum that is placed at the base of both families (Figure 5). Of note, the Microgenomates that harbor the unusual new form of MVA kinase also possess the regular MVA kinase (Figure 5). The tree topology rules out recent horizontal gene transfer(s) from other archaea or eukaryotes as the explanation for the MVA pathway in CPR bacteria, with a few important exceptions (transfers from DPANN to Dependentiae [TM6] and from Firmicutes to Woesearchaeota; Figure 5).
The “lipid divide” has been a central feature used to establish the distinction between bacteria and archaea. The newly reported distribution of the MVA and the MEP pathways in both domains reopens the question of their evolutionary origin and alters the prior conclusion that the MEP pathway is restricted to bacteria.

How the CPR and DPANN May Change Our View of Evolution
As noted above, the availability of genomes for a much more comprehensive set of organisms revealed major new features of the tree of life (Figure 1). The question of whether the tree topology involves two major subgroups within bacteria and archaea—the CPR and DPANN, respectively—is unresolved. Are the apparent groupings of CPR and DPANN artifacts of rapid evolution or reflections of early origins and a long history of diversification?

A phenomenon that could drive fast evolution is genome reduction, which is an obvious consideration for CPR and DPANN, given their small cell and genome sizes. Largely symbiotic lifestyles are linked to rapid gene loss and fast accumulation of mutations in well-studied symbioses (e.g., those that involve partnerships of bacteria and eukaryotic hosts), although these may not be appropriate models for associations in which the host is bacterial or archaeal. The early stages of genome reduction in symbionts of eukaryotes are characterized by proliferation of mobile elements, formation of pseudogenes, multiple genomic rearrangements, and deletion of chromosome fragments. To date, such phenomena have not been described as prominent features of CPR and DPANN genomes (Nelson and Stegen, 2015). In more anciently evolved symbionts, such as Buchnera aphidicola, mobile elements and most pseudogenes have been eliminated (van Ham et al., 2003). Thus, if the CPR and DPANN experienced radiation-wide genome reduction, it may have occurred long ago.

It is important to note that the small genomes of bacterial symbionts of insects and those of human-associated pathogens, such as Chlamydia, do not cluster within the CPR. This suggests that the phylogenetic placement of the CPR is not an artifact of genome streamlining. That said, metabolisms predicted for CPR and DPANN vary dramatically, and while some have sufficient metabolic capacities to suggest that they may be free living, others lack numerous biosynthetic capacities. Thus, we predict that genome reduction may be an important phenomenon in some lineages.

The alternative explanation for the existence of CPR and DPANN as distinct major radiations is that both arose from very early-evolving organisms (possibly with small genomes), and the long branches of the CPR and DPANN are due mostly to undersampling (rather than rapid evolution). Although their apparent deep branching phylogenetic placement remains in question, other observations make an early origin of both groups from a common ancestral pool worth considering. Most important, perhaps, is the similarity in the suite of biosynthetic capacities that both groups possess (extensive genes of the information system) and in missing pathways (e.g., lack of the electron transport chain and tricarboxylic acid [TCA] cycles that are widely distributed in other bacteria and archaea). If, as one might predict, extensive use of oxygen-based respiration arose relatively late (following the advent of \( \mathrm{O}_2 \)-generating photosynthesis around 2 billion years ago), it would make sense that groups with ancient metabolic platforms would be anaerobes. The metabolisms of CPR and DPANN are consistent with such a world, as these organisms appear to be almost exclusively anaerobes, lacking a full TCA
cycle and electron transport chain required for aerobic growth. Similarly, almost all are incapable of dissimilatory nitrate reduction, and none are predicted to have metabolisms based on dissimilatory sulfate reduction. Both oxidized nitrogen and sulfur-based compounds would have been at low abundance prior to an oxidized atmosphere. Obviously, it is speculation, but we imagine that nucleotide and amino acid biosynthesis capacities may have been present in ancestral populations but lost during genome reduction as (most) CPR and DPANN adopted lifestyles that depended on organisms with more recently evolved capacities (e.g., aerobic growth, photosynthesis, and other forms or chemoautotrophy). The alternative explanation that cannot be ruled out at this time is that the many similarities in the genomic features of CPR and DPANN may have arisen by convergent evolution, possibly involving late radiation-wide adaption to anaerobic habitats.

Although what we see now in CPR and DPANN genomes must be a faint echo of what once was, the presence of specific genes in widely divergent groups may hint at features of ancestral organisms that were lost from many modern groups. Included in this list are capacities such as modified glycolysis (reliant on pentose phosphate pathway enzymes), hydrogen metabolism (as suggested for the ancestor of Archaea; Williams et al., 2017), and ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO). In fact, one of the most striking cases of widely but sparsely distributed enzymes found in both CPR and DPANN organisms are the form II/III- and III-like RuBisCOs that apparently function in a nucleotide-based pathway that feeds into lower glycolysis and fermentation (Wrighton et al., 2016). This enzyme occurs in some extremely minimal genomes (Dojkabacteria [WS6] and Pacearchaeota), consistent with its central role in the metabolism of these organisms. Thus, RuBisCO that functions in nucleotide metabolism and central carbon metabolism may have been key to the physiology of ancestral CPR and DPANN. A bacterial or archaeal cell is, on average, 20% RNA, and RNA is 40% ribose by weight (i.e., a cell is about 8% ribose; Schönheit et al., 2016). Therefore, it has been suggested that ribose was likely an abundant sugar available on early Earth for fermentation. Hence, type III- and II/III-based RuBisCO pathway of nucleoside monophosphate conversion to 3-phosphoglycerate may be a relic of ancient heterotrophy.

The alternative explanation for the very striking patterns of similarity in the presence and absence of genes is a combination of gene loss and horizontal gene transfer. Horizontal gene transfer certainly has shaped the evolution of CPR and DPANN (Jaffe et al., 2016), but this remains relatively unstudied. So far, where more complex metabolic capacities occur, they are present sporadically, and the enzymes are highly divergent relative to all currently known sequences. Likely, the evolutionary trajectories of CPR and DPANN have been shaped by combination of gene loss, lateral transfer, and convergent evolution. The balance in importance of these processes across the radiations is very much an open question.
Asgard Archaea Reinforce the Idea that Archaea Were Ancestors of Eukaryotic Cells

The Asgard superphylum, comprising Lokiarchaeota (Spang et al., 2015), Thorarchaeota (Seitz et al., 2016), Odinarchaeota (aremba-Niedzwiedzka et al., 201), and Heimdallarchaeota (aremba-Niedzwiedzka et al., 201) represents an important recent discovery from metagenomics. Based on phylogenetic analyses that included sequences from members of the Asgard superphylum, it was proposed that these archaea share a relatively close evolutionary relationship with Eukaryotes (Spang et al., 2015). There have been many competing hypotheses regarding the origin of eukaryotic cells, and the idea that they arose from the fusion of bacterial and archaeal cells has a long history. Jim Lake and colleagues suggested this in the 1980s (the “eocyte” hypothesis; e.g., Lake, 2015, Lake et al., 1984, Williams et al., 2013). A mixed origin for the eukaryotic cell involving both bacteria and archaea was supported by the fact that genes for DNA replication, transcription, and translation systems of eukaryotes resemble archaeal genes, whereas genes involved in energy and carbon metabolism are similar to genes of bacteria. Martin and Müller (1998) proposed that Eukaryotes arose through symbiotic association of an anaerobic, autotrophic, hydrogen-dependent archaeal host with bacterial symbiont. Others suggested that they arose via symbiosis between methanogenic archaea and an ancestral sulfate-reducing myxobacterium (Moreira and Lopez-Garcia, 1998). The claim regarding the phylogenetic placement of Asgard archaea with Eukaryotes (Spang et al., 2015) was an important new twist and considered controversial by some (Da Cunha et al., 2017).

The phylogenetic argument of Spang et al. (2015) is supported by recent and more comprehensive studies (aremba-Niedzwiedzka et al., 201). Important discoveries in the original and new studies relate to the presence of numerous genes and pathways in Asgard archaea that are normally considered to be eukaryotic signatures (eukaryotic signature proteins, ESPs), thus mostly absent in other archaea. These include some components related to membrane remodeling and trafficking, cytoskeletal features, ubiquitin-modifier systems, tubulins, and vesicle-formation processes in eukaryotes (actin, small Ras GTPases, ESCRT complex; Spang et al., 2015, Zaremba-Niedzwiedzka et al., 2017). Although genes encoding some of these proteins had been noted previously in archaeal genomes (Ettema et al., 2011, Hartman and Fedorov, 2002, Lindås et al., 2008, Yutin and Koonin, 2012), the presence of multiple components adds significant weight to claims based on phylogenetic reconstructions. Thus, Asgard archaea likely provide important clues regarding the nature of the precursors of the first eukaryotic cells. Although it must be acknowledged that the biochemical functions of the intriguing eukaryotic features of Asgard archaea remain to be determined, the presence of such a repertoire may imply that if the ancestor of eukaryotes was archaeal, the organisms were far more complex than otherwise predicted.
Metabolic Handoffs, Microbial Communities

A strength of genome-resolved metagenomics methods is that they provide information about the metabolic potential of coexisting community members. Thus, in addition to resolving which organism can do what (Figure 6A), it is possible to examine the distributions of traits over coexisting organisms within an ecosystem (Anantharaman et al., 2016, Dombrowski et al., 2017, Wrighton et al., 2014). The Anantharaman et al. analysis leveraged over 2,500 genomes from a single aquifer environment (representing around ⅓ of all organisms detected there) and revealed evidence for extensive interconnection among the metabolisms of coexisting community members. These connections involve the supply and consumption of compounds of the nitrogen, sulfur, carbon, hydrogen, and other biogeochemical cycles. For example, an organism may be able to conduct only a subset of the redox transformations in a biogeochemical pathway. Thus, they release a partially oxidized (e.g., thiocyanite that can be oxidized to sulfite) or partially reduced (e.g., nitrite that can be reduced to nitrous oxide) intermediate product for use by other organisms. This ecosystem structure is counter to the expectation that, for example, a single bacterium is responsible for all steps in nitrate reduction. The phenomenon, involving cooperative biogeochemical cycling, has been referred to as “metabolic handoffs” (Anantharaman et al., 2016, Hug et al., 2015) and is diagrammed in Figure 6B. Through metabolic handoffs, the ecosystem can adopt a large number of configurations by shifting the mixture of organisms involved in a process as environmental conditions change. Complex interdependencies, and especially obligate symbioses, may explain why some organisms are difficult to grow in pure cultures (as isolates).
Figure 6: Information from genome-resolved analysis of natural microbial communities. (A) Predicted metabolic features for some newly described bacteria and archaea from a range of different sites and studies (for references, please see the text). The simplified cell cartoons highlight how genomics-based analyses of organisms from candidate phyla have revealed their biogeochemically important roles. Organisms are represented in the context of sediment, an environment that hosts many newly described bacteria and archaea, but they occur in other environments. Genomes of bacteriophage (phage), viruses (of archaea), and plasmids can also be reconstructed, enabling direct analysis of the roles of these entities in augmenting host metabolism and lateral gene transfer. (B) Diagram illustrating the concept of coexisting microorganisms working together to complete a single biogeochemical pathway (ammonia, NH3, to N2). This is based on Figure 6 and traits data in Supplementary Table 9 of nantharaman et al. (2016. Examples of coupling of reduction of nitrogen-based compounds to the iron, carbon, hydrogen, and sulfur cycles are included. The phenomena
illustrated are referred to as “metabolic handoffs” because the product of one organism is made available to another organism with the capacity to use it. Importantly, Anantharaman et al. demonstrated that few organisms have the ability to generate energy from all of the steps in most pathways analyzed. Note that essentially all of the organisms studied are only distantly related to isolated organisms. For taxonomic affiliations, see Table S1.

Concluding Thoughts and Some Frontiers for Future Research

The tsunami of new genome sequences for bacteria and archaea has dramatically reshaped our understanding of life’s diversity and expanded knowledge regarding microbial roles in biogeochemical processes that impact atmospheric chemistry, soil fertility, water composition, and human health. The genomes provide context for metaproteomic and metatranscriptomic data, which can identify the pathways that are operational at any time on an organism-by-organism basis (e.g., Frias-Lopez et al., 2008, Ram et al., 2005). The availability of new genomic information for candidate phyla organisms, arguably the majority of all bacteria and archaea, has demanded that we rethink concepts related to microbial community structure and functioning, the nature of symbioses, and the origins of evolutionary innovations. Huge and fundamental questions remain regarding early evolution that can only be addressed by better and more detailed phylogenetic analyses that leverage more comprehensive sets of genomes.

The scale of microbial diversity in Domain Eukarya is likely underestimated. A major advance anticipated in the coming years is the discovery, metabolic exploration, and analysis of the habitat distribution of microbial eukaryotes. A few studies have already employed cultivation-independent genomic methods for de novo reconstruction of genomes of microbial and even macro-eukaryotes from short read datasets (Kantor et al., 2017, Mosier et al., 2016, Sharon et al., 2013). Although approaches are improving rapidly (West et al., 2017), genome-resolved metagenomics methods have not yet found broad application for this purpose. The combination of cultivation-independent genomic information with transcriptomic data may resolve how the metabolism of microbial eukaryotes depends on the types and activities of associated bacteria and archaea. For example, laboratory studies using a single genotype of the human pathogen Candida albicans (Eukaryota, Opisthokonta, Fungi) showed that it underwent a phenotypic switch, with dramatic alteration of its gene expression patterns, when it was grown in the presence of different bacterial isolates (Fox et al., 2014). Such phenomena could be directly tested by genomically defining Candida in real human gut microbial communities via genome-resolved metagenomics and then evaluating its transcriptome (via metatranscriptomics) as a function of the flanking microbial community composition.

Medical research has been substantially impacted by the introduction of cultivation-independent microbiome studies. However, genomically resolved analyses remain in the minority, partially due to the high costs and slow speed relative to rRNA sequence-based analyses. We anticipate that this will
change with the advent of new, real-time, and low-cost sequencing methodologies. Although rRNA gene surveys and existing metagenomic data suggest that it is unlikely that many major new microbial groups (e.g., phyla) will be discovered in the human microbiome, new species-, genus- and higher-taxonomic-level groups certainly will be uncovered. Importantly, genome sequences for even closely related strains can differ significantly, and variation may involve genes that contribute to pathogenicity and antibiotic resistance. Thus, metagenomic methods have the potential to provide new insights into the medical importance and drug susceptibility of human-associated microbes and ultimately could guide medical treatments.

An understudied question relates the medical significance of CPR and DPANN. Overall, they do not appear to be major players in most human microbiomes. Saccharibacteria (TM7) and Absconditabacteria (SR1) have been found in the mouth (Dewhirst et al., 2010). TM7 have been linked to periodontal disease, and one subtype is found primarily at diseased sites (Brinig et al., 2003). Members of this phylum are also associated with dermatitis and cystic fibrosis (He et al., 2015), and they were detected in other mucus-rich sites associated with vaginosis and inflammatory bowel disease (Fredricks et al., 2005, Kuehbacher et al., 2008). Importantly, experiments involving co-culture of a TM7 with its host showed that TM7 could modulate the immune response (He et al., 2015). Intriguingly, bacteria from these groups, along with a variety of DPANN archaea, are relatively common in dolphin mouths (Dudek et al., 2017), suggesting their broader importance in animal health. Woesearchaeota were recently detected in the lungs of humans (Koskinen et al., 2017), and cell-free DNA of certain CPR bacteria was found in the blood of many human subjects (Kowarsky et al., 2017). Both CPR and DPANN are abundant in groundwater and thus may be ingested regularly by humans and animals. It is our view that PCR-independent methods applied to human microbiome samples from diverse body sites, disease states, and populations could provide new information relating to the potential medical significance of CPR and DPANN.

Throughout this Perspective, our focus has been on genomics, i.e., what can be learned from DNA sequence information. Ultimately, however, experiments are needed to address the biological relevance of the predictions informed by sequence analysis. The landmark study involving the TM7 co-culture with its actinobacterial host provides a template for future work. Metagenomics studies provide information about appropriate environmental sample sources that could be used to seed enrichments. Co-abundance patterns for CPR or DPANN and other bacteria or archaea could be used to identify possible symbiont-host associations. Based on host gene content, it should be possible to design enrichment studies to bring other CPR and DPANN into co-cultures. Going beyond the physiological and morphological studies that co-cultures enable, specific organisms in enrichments could be targeted using modern genome-editing methods to
remove or upregulate genes in either the symbiont or host to discover the mechanisms of their interaction.

With sufficient genomic and linked functional data, it will be possible to really begin to understand how natural ecosystems function. Comprehensive “omics” information is already being collected from extensive suites of time series and spatial samples, as well as from laboratory bioreactors and full field-scale manipulation experiments. Ultimately, across ecosystem types, it should be possible to link organisms to processes and to understand the conditions that cause organisms of interest to proliferate. Thus, genomically informed ecosystem (including human microbiome) manipulations should soon come within reach.

Acknowledgments

Support was provided by grants from the Lawrence Berkeley National Laboratory’s Genomes-to-Watershed Scientific Focus Area. The U.S. Department of Energy (DOE), Office of Science, and Office of Biological and Environmental Research funded the work under contract DE-AC02-05CH11231 and the DOE carbon cycling program DOE-SC10010566, the Innovative Genomics Institute at Berkeley and the Chan Zuckerberg Biohub. We thank the editor and two anonymous reviewers for excellent input to this manuscript.

References


