The bright side of microbial dark matter: lessons learned from the uncultivated majority

Lindsey Solden[1], Karen Lloyd[2] and Kelly Wrighton[1]

Microorganisms are the most diverse and abundant life forms on Earth. Yet, in many environments, only 0.1–1% of them have been cultivated greatly hindering our understanding of the microbial world. However, today cultivation is no longer a requirement for gaining access to information from the uncultivated majority. New genomic information from metagenomics and single cell genomics has provided insights into microbial metabolic cooperation and dependence, generating new avenues for cultivation efforts. Here we summarize recent advances from uncultivated phyla and discuss how this knowledge has influenced our understanding of the topology of the tree of life and metabolic diversity.

Addresses
[1] Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA
[2] Department of Microbiology, University of Tennessee, Knoxville, TN 37996, USA

Corresponding authors: Lloyd, Karen (klloyd@utk.edu) and Wrighton, Kelly (wrighton.1@osu.edu)

Introduction

Today at least 89 bacterial and 20 archaeal phyla are recognized by small subunit ribosomal RNA databases, although the true phyla count is certainly higher [1,2,3••,4,5•] and could range up to 1,500 bacterial phyla [6]. Historically, a majority of what we understood about microbial life was based on information gleaned exclusively from cultivated organisms [2]. Thus, physiologic and genomic information was confined to pure cultures and dominated by representation of the Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes within the Bacteria, and to methanogen and halotolerant members of the Euryarchaeota within the Archaea [2].

The first realizations of just how diverse and unexplored microorganisms are came from analyzing microbial small subunit ribosomal RNA (SSU or 16S rRNA) gene sequences directly from environmental samples [7]. These analyses revealed that less than half of the known microbial phyla contained a single cultivated representative. Phyla composed exclusively of uncultured representatives are referred to as Candidate Phyla (CP). Borrowing language from astronomy, microbiologists operationally define these CP as microbial dark matter, because these organisms likely account for a large portion of the Earth's biomass and biodiversity, yet their basic metabolic and ecological properties are not known. This uncultivated majority represents a grand challenge to the scientific community and until we solve the mysteries of the CP, our knowledge of the microbial world around us is profoundly skewed by what we have cultivated in the laboratory [8•].

In the past five years, scientists have addressed missing information from uncultured organisms through advances in genomic sequencing technologies. Microbial genomes can now be directly sequenced from the environment using metagenomics and single cell genomics, however, these technologies contain their own strengths and challenges. Using metagenomics, DNA is sequenced directly from the environment facilitating the study of organisms in the context of their community and chemical conditions. After shotgun sequencing and assembly, this approach results in genomic fragments from different organisms, which can be binned into separate genomes using shared features (abundance, codon usage, tetranucleotide, homology). This method today, can result in complete, closed genomes [9]. However, binning does not always resolve strains, representing a composite of genomic fragments from separate clonal populations [8•]. In contrast to the bulk sequencing of the entire community, single cell genomics involves physically separating a single cell from the environment, lysing the cell, and amplifying and sequencing the genomic DNA. Although this procedure does not suffer from ambiguity about the number of organisms contributing the DNA, the amplification step biases the genomic coverage, often resulting in fragmented, less complete genomes. Metagenomics and single cell genomics are therefore complementary and together have contributed new insights into uncultivated lineages.

Insights from microbial dark matter genomes

Genome-enabled approaches offer metabolic predictions for previously enigmatic CP organisms. In 2012, Jillian Banfield and colleagues used metagenomics to reconstruct

49 draft genomes from at least five bacterial CP that lacked prior genomic information [10]. A year later, the rapid development of metagenomic sequencing and analyses yielded the first complete, closed genomes from these same or closely related bacterial CP lineages [11,12]. In this same year, Tanja Woyke and colleagues used single cell genomics to target microbial dark matter lineages and genomically sampled at least twenty bacterial and archaeal phyla composed exclusively of uncultivated members for the first time [2].

Recently, the genomic sampling of uncultivated organisms has increased dramatically. In 2015, Brown *et al.* expanded the genomic sampling of earlier identified CP to reveal a bacterial radiation that is estimated to account for at least 15% of the known bacterial diversity [3••]. This radiation was initially composed entirely of uncultivated members and was hence referred to as the Candidate Phyla Radiation (CPR). Also in this same year, Castelle *et al.* reconstructed 151 archaeal CP genomes, expanding the sampling of existing CP and resolving two new phyla, each with nearly fifty genomes sampled [1]. While these aforementioned studies initially contributed large numbers of CP genomes and were foundational in the development of this field, many other researchers have also discovered new phyla and contributed to our understanding of the phylogenetic and metabolic diversity within existing CP [8•,13,14••,15••,16–18,19•,20,21].

Given the pace at which CP genomes are being sequenced, it is no surprise that genomic databases need to be continually updated to remain contemporary. To highlight what is commonly accessible in genomic databases, for each archaeal and bacterial phylum that has a well-curated genomic representative we summarized the relative contribution of isolate genomes to total genomes sampled [22] (Figure 1). The sampling from Bacteria (black), bacterial CPR (blue), and Archaea (orange) are distinguished by font colour. The relative contribution of genomes from uncultivated members, denoted in black, were contributed via single cell genomic and metagenomic approaches. According to this database analyses, a third of the bacterial and archaeal phyla genomically sampled in databases lack a single isolated representative.

The data included in Figure 1, while highlighting the contribution of microbial dark matter to genomic space, is a gross underestimate. For instance, CP genomes from the Archaea [1,13,14••,15••,16], and many of the bacterial CPR are missing [3••,5•]. Specifically, the Parcubacteria (OD1) is shown as a phylum having 14 genomes, but due to the extensive genomic sampling last year of over 400 new Parcubacterial genomes, this phylum is now recognized as a bacterial superphylum [3••,5•]. Today several CP phyla (e.g. Yanofskybacteria of the Parcubacteria with 84 draft genomes) contain as much genomic sampling as many well-studied phyla that contain cultivated organisms. Moreover, with increased genomic sampling, taxonomic boundaries and nomenclature are constantly being reassessed, albeit with some controversy. For instance, the Patesibacteria (Figure 1), a proposed superphylum that included three phyla, is now reorganized and renamed to the CPR which currently includes two superphyla and many more phyla. This and other instances outlined in the section below, exemplifies the taxonomic upheaval and points of contention currently ongoing in the field of microbiology. Below we identify three new discoveries illustrating how information from CP genomes has altered our understanding of the microbial world.

i) Breaking branches of the tree of life

Based on classification by Woese and colleagues, ribosomal RNA resolved Earth's biodiversity into a three-domain model where Bacteria, Archaea, and Eukarya each represented monophyletic groups [23,24]. This three-domain model has arguably stood the test of time, until being challenged by recent CP genomic sampling [25,26]. Based on phylogenomic analyses of new and existing archaeal phyla, Williams and Embley proposed a two-domain model consistent with the 'eocyte' hypothesis [26], where eukaryotes originated from within the archaeal radiation [27]. This interpretation is at odds with the three-domain model based on rRNA gene sequences and membrane lipid content, which separates the eukaryotic and archaeal lines prior to the archaeal radiation [28].

Further instability in the three-domain theory emerged upon the discovery of several reconstructed archaeal genomes from a marine sediment metagenome [14••]. These genomes represented members of a single phylum, and were named Lokiarcheaota after the 3,000 m deep marine hydrothermal vent system, Loki's Castle, where the samples originated [29]. Based on concatenated phylogenetic marker proteins, the authors proposed that Lokiarchaeota were the most closely related prokaryotic lineage to the eukaryotes. Further analyses revealed the Lokiarchaeota genomes contained several genes previously thought to define unique aspects of eukaryotic biology [26]. This proposed relationship with eukaryotes (where eukaryotes are sister to or fall within archaeal domain) is not without contention, as inferences from concatenated gene trees using composite metagenomic information are controversial [30]. Additionally, 16S rRNA genes assign the Lokiarchaeota to the known environmental 'Marine Benthic Group B', a monophyletic collection of archaeal rRNA sequences that lack a clear association with eukaryotes.

More recently, Hug *et al.* summarized the impact of single cell and metagenomic sampling on our current understanding of the tree of life [5•]. This study phylogenetically analyzed 3,083 genomes including at least one representative from all genera for which high quality genomes exist. According to these analyses approximately

Figure 1

Current Opinion in Microbiology
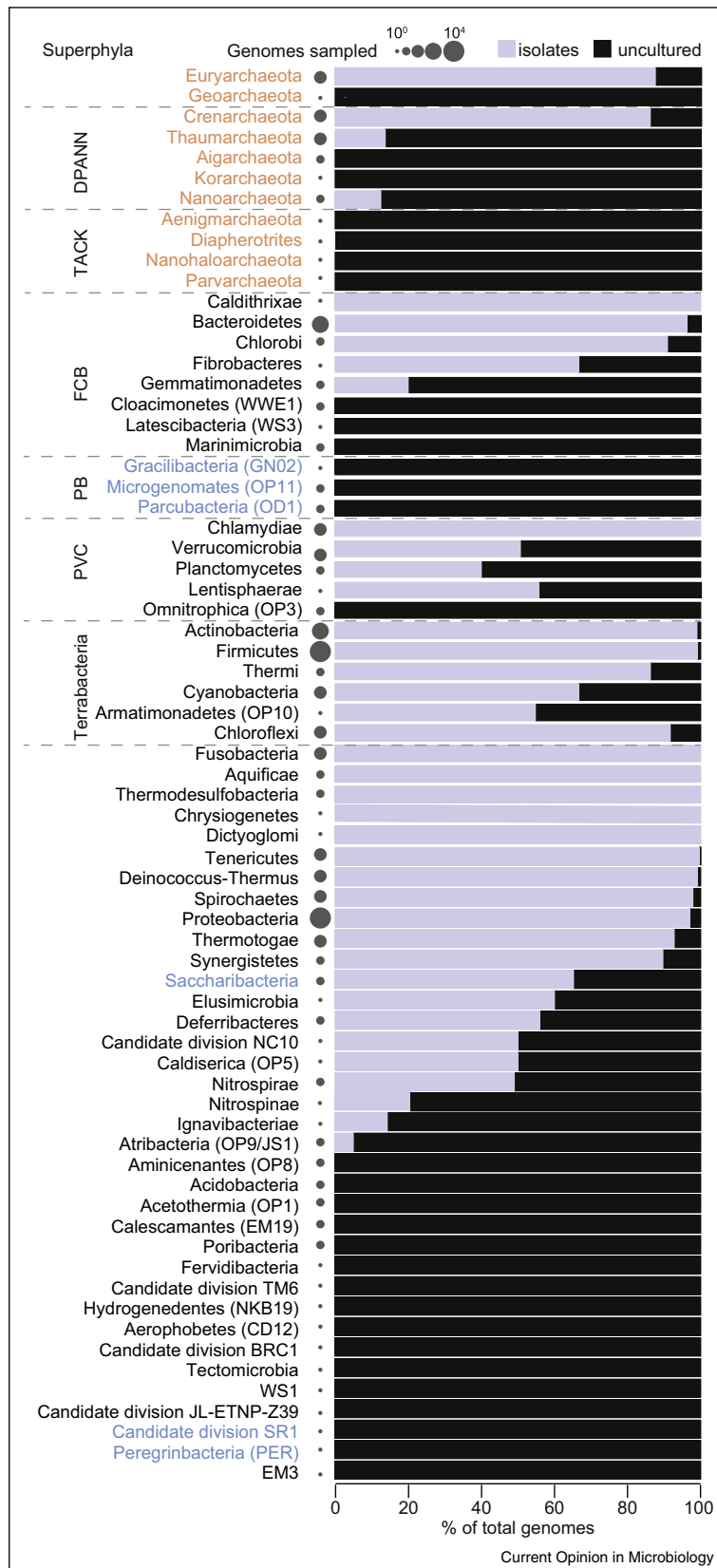
50% of Archaea, 40% of the non-CPR bacterial phyla, and 97% of the CPR bacterial phyla contain uncultivated members, illustrating that bacterial and archaeal phyla without a single cultivated representative clearly comprise the majority of life's current diversity.

Similar to the Lokiarchaoata analyses, this study reiterated the discrepancy between phylogenies constructed from concatenated marker proteins and those from small subunit (SSU) rRNA. For instance, the CPR appear deep branching by a concatenation of 16 ribosomal proteins, but not in the 16S rRNA gene phylogeny [5•]. The position of the CPR as early emerging only on the ribosomal protein tree raises the question as to whether this topology is an artifact from fast evolving symbiont genomes. Additionally, this study reiterates the controversy between the two and three domain models of life. Concatenated ribosomal proteins recapitulate the two domain model, with the Eukarya branching within the TACK superphylum, while this placement is not supported by the SSU rRNA gene phylogeny. Future metagenomic and single cell genomic investigations promise to add missing foliage to the tree of life, and along with the use of additional phylogenetic markers may further resolve the ancestral relationship between Archaea and Eukarya.

ii) Unexpected metabolisms in new phylogenetic places

16S rRNA gene sequence surveys of anoxic marine and terrestrial subsurface environments are often dominated by a group of deeply branching archaea, formerly called the Miscellaneous Crenarchaeotal Group [31], or Marine Group 1.3 [32]. This group has been renamed Bathyarchaeota, based on its discovery in deep-sea environments [33], but has also been found in a wide range of terrestrial environments, including aquifers [1] and coal bed methane wells [15••]. The diversity of 16S rRNA genes within the Bathyarchaeota is befitting of a phylum, with at least 17 subgroups identified, suggesting considerable ecophysiological divergence [34].

The first genome from Bathyarchaeota, obtained from marine sediment, contained the capacity for protein fermentation, a well-studied metabolism in Bacteria never before seen in non-extremophilic Archaea [17]. Twelve additional genomes now available from eight Bathyarchaeota subgroups [18,35] suggest other metabolisms not commonly associated with Archaea, such as homoacetogenesis via a Wood-Ljungdahl pathway [18,35]. Additionally, Gene

Tyson and colleagues recently used metagenomics to reconstruct two near-complete Bathyarchaeota genomes from coal bed fluids and discovered the capacity for methane cycling [15••]. These genomes had complete methanogenic pathways, including a divergent homologue for the key enzyme in methanogenesis and anaerobic methanotrophy. This key enzyme, methyl coenzyme M reductase (Mcr), has never been identified outside the Euryarchaeota phylum in the Archaea [15••,36]. It is likely that other microbial candidate phyla will yield similar metabolic novelty.

### iii) Small genomes and metabolic interdependencies are prevalent

Today ∼35 genomically sampled phyla make up the CPR, with all phyla except one lacking a single cultivated representative [1,5•]. This radiation includes two superphyla Parcubacteria (OD1) and Microgenomates (OP11), that each contain multiple phyla within. Additionally the CPR include many separate phyla including the Peregrinibacteria (PER), Dojkabacteria (WS6), Katanobacteria (WWE3), Berkelbacteria, Saccharibacteria (TM7), Gracilibacteria (BD1-5), Absconditabacteria (SR1), Kazan phyla, as well as three previously unrecognised lineages (CPR 1, 2, 3) (Figure 2) [3••]. Besides adding new branches to bacterial tree of life, CPR genomic analyses have revealed information on cellular information processing, fermentative metabolisms, and host associated lifestyles.

Analyses of CPR genomes have revealed unusual protein synthesis machinery common to other bacterial phyla. For instance, some near-complete CPR genomes are missing ribosomal and biogenesis factors previously considered universal in bacteria [3••]. Another feature reported in many CPR [3••], and in some endosymbionts [37,38], is the presence of self-splicing introns and proteins encoded in the 16S rRNA gene [38,39]. These introns varied in length (up to 3.861 bp), position (one complete Microgenomates genome had 4 different insertions), and were not maintained in transcribed rRNA. Moreover, it was shown that at least 50% of the CPR phyla have divergent 16S rRNA gene sequences and fail to amplify using standard universal bacterial primers (515F and 806R) [3••].

Beyond unusual ribosomes, members of the CPR also use alternative genetic codes. In members of the Absconditabacteria and Gracilibacteria phyla, the UGA stop codon is reassigned as glycine [11,19•,40–42], a finding confirmed

(Figure 1 Legend) Depiction of uncultivated bacterial and archaeal genomes assigned to isolate and uncultivated organisms in each phylum present in the JGI Gold database (Dec. 17, 2015). The percentage of genomes within each phylum with an isolated organism is represented in purple, while the percentage of high quality genomes from uncultivated organisms is in black. Font colour depicts whether the phyla is currently a part of Bacteria (black), bacterial CPR (blue) identified by Hug et al. (2016), or Archaea (orange) [5•]. The circle next to each phylum represents the number of genomes in each phyla (from 100 to 10,000). The designated super phyla proposed originally by Rinke et al. (2013) are indicated on the left (DPANN, Diaperotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota, Nanoarchaeota; TACK, Thaumarchaeota, Aigarchaeota, Crenarchaota, Korarchaeota; PB, Patesibacteria; PVC, Plactomycetes Verrucomicrobia Chlamydia; FCB, Fibrobacteres, Chlorobi, Bacteroidetes) [2].

Figure 2
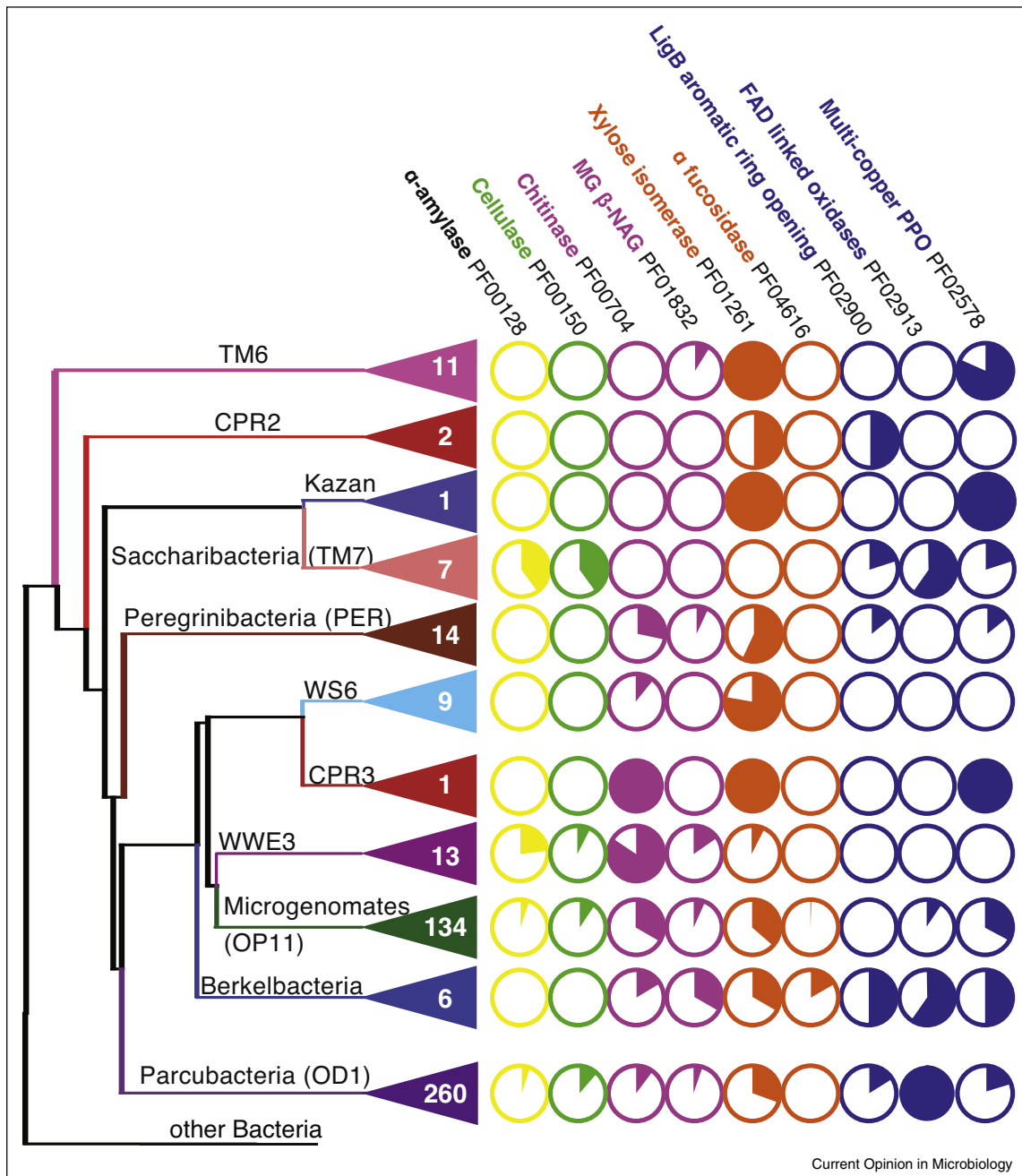
Analyses of glycoside hydrolases from CPR and TM6 genomes. The relative phylogenetic position is based on a 16S rRNA gene maximum likelihood tree previously published [3••]. Carbohydrate active enzymes from CPR and TM6 genomes were identified (http://ggkbase.berkeley.edu/ CPR-complete-draft/organisms). For the most prevalent enzymes across the dataset, the gene annotation and corresponding PFAM number are displayed, with coloring indicating the putative carbon substrate (yellow is amylose, green is cellulose, purple is chitin, orange is hemicellulose, and blue is lignin). The total number of genomes included in this analyses from each phylum is displayed on each node, while the pie chart represents the percentage of genomes in each phylum that have annotated genes within each PFAM category.

by proteomics [41]. In contrast to alternate coding in symbiotic Alphaproteobacteria and mitochondria where UGA encodes tryptophan, these lineages do not appear to use a single tRNA with wobble pairing to accommodate the additional glycine codon, but rather have an additional tRNA$^{Gly}_{UGA}$ [11]. While not currently known, UGA recoding for glycine could be a mechanism for reducing GC content [11] or ensuring genetic perseverance by decreasing horizontal gene transfer rates to maintain genomic innovation [19•].

The inferred metabolic capabilities of CPR genomes sampled to date are limited. Based on recognizable genes these bacteria appear to be obligately fermentative anaerobes without evidence for aerobic or anaerobic respiration. Recent analyses of Parcubacterial genomes from aerobic habitats have identified genes that suggest the capability for using O_2 as a terminal electron acceptor, but the lack of other required electron transport chain proteins leave the functional role for these enzymes uncertain [43]. Furthermore, to examine the capacity for hydrogen generation and consumption we reanalysed the available CPR genomes from Brown *et al.* [3**] and recovered 110 full-length NiFe hydrogenases that formed three new monophyletic groups within the 3b hydrogenases. Consistent with a fermentative metabolism, these hydrogenases are proposed to play critical roles in the disposal of reducing equivalents in CPR organisms. We also analyzed near-complete CPR and TM6 genomes for carbohydrate active enzymes (CAZy) [44] (Figure 2). Our analyses revealed enzymes for xylose, a common hemicellulose backbone sugar, and lignin degradation were the most commonly detected, while cellulose degradation was unevenly distributed phylogenetically. These findings support inferences from a variety of ecosystems that CPR organisms likely play critical roles in anoxic carbon transformations and hydrogen generation [3**,11,43,45,46].

One of the most remarkable features of many CPR genomes is the small genome sizes (often <1 Mb). These genomes are as small or smaller than known free-living bacteria, on par with many obligate endosymbionts [8•]. With the exception of Peregrinibacteria, which have the capacity for nucleotide synthesis [46], this genome reduction is at least in part because of a lack of biosynthesis pathways for nucleotides, lipids, and most amino acids [3**,11, 20,21,42,46]. This auxotrophy suggests that these organisms may be dependent on one or more members of the surrounding community for cellular metabolites.

Based on inferences that small sized genomes often corresponded to small cellular sizes, the same environment where CPR were first genomically identified was resampled using a smaller sized filter (0.1 μm). Groundwater that passed through a 0.2-μm filter contained almost exclusively CPR, and cryogenic transmission microscopy revealed ultra small cells at the lower size limit for life [46,47] (Figure 3). Consistent with metagenomic predictions [10,11], cells lacked a gram-negative cell membrane, had a distinct S-layer, and expressed abundant pili, a feature which may be necessary for interacting with other organisms or the environment via adhesion to extracellular surfaces.

Small genomes on the scale of the CPR have been observed with two distinct microbial lifestyles. These include either (i) streamlined, yet free-living, with a limited metabolic repertoire or (ii) host associated often lacking biosynthetic genes. Compared to the model of SAR11, genome-streamlining evidence (e.g. codon density and pseudogene content) in CPR genomes was not strong [16,48]. Alternatively, many genomic studies have suggested that CPR and other CP with small genomes are host-dependent [3**,5•,11,20,21,49••]. Microscopy and cultivation have been used to validate some of these genome hypotheses [49••,50]. For instance, consistent with separate metagenomic claims [11], microscopic evidence demonstrated that Parcubacterial CPR cells were

Figure 3

Current Opinion in Microbiology
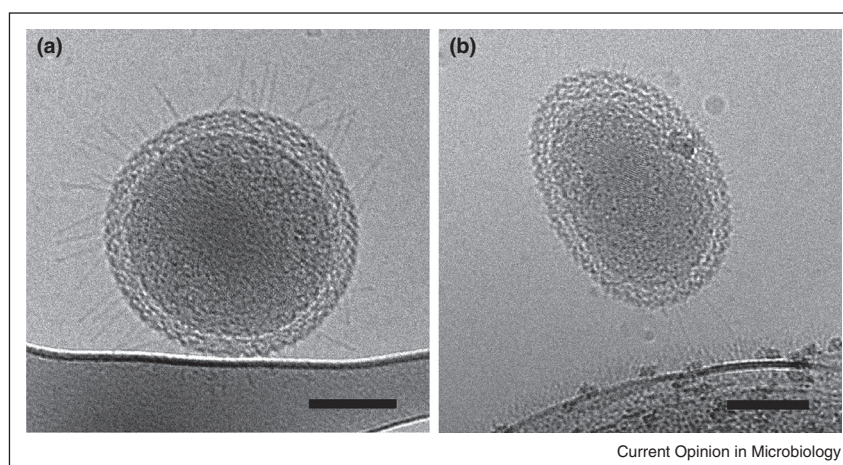
Cryo-TEM images (2D) documenting morphology and size of ultra-small bacterial candidate phyla. **(a)** The cell envelope includes a remarkable and distinct S-layer. Pilli-like structures are clearly discernible: numerous radiating pili-like structures cover the surface of the cell. Notably morphological features are consistent with genomic analyses [3**,10]. **(b)** Polar pili-like structures occur on the cell, apparently connecting it to an adjacent bacterium (only part of the bacterium shown). Previously published in Luef *et al.* [47].

ectosymbiotic with algal endosymbionts localized within a ciliate [50].

Outside the CPR, the bacterial candidate phylum TM6, which is also comprised of small genomes with limited metabolic capacity, was inferred from genomics to be reliant on an amoebal host [27,47]. Additionally, similar trends of small genomes and sparse metabolisms were also reported for many archaeal CP genomes [1]. Together these findings of auxotrophy and host association hint that a cooperative or dependent metabolism may likely be much more widespread across the tree of life.

Illuminating gene functional annotation within CP genomes

The inferred metabolisms discussed above were dependent on drawing similarities between distant gene homologs in the CPR and well-characterized genes from phylogenetically distinct isolated organisms. However, given the phylogenetic divergence and unusual environments some CP genomes were reconstructed from, it is also possible these uncultivated representatives interact with the environment in new ways, containing proteins with currently unknown functions. Since existing methods of genome analyses rely on annotations largely based upon cultivated organisms, gene functionality assignments could be missed due to sequence divergence or misannotation. Recent work, however, has begun to unravel the function of CP proteins and their contribution to the overall physiology of these lineages.

Metagenomic studies recently uncovered a new form (II/III) of RubisCO genes from multiple uncultivated phyla within the bacterial CPR and Archaea [1,10,11,19•,46,51]. Prior to these discoveries, these genes were identified only in cultivated methanogens which all contained a 29 amino acid insertion. Sequences from CPR lineages that were most closely related to methanogen RubisCOs were not known to encode functional proteins based on the considerable sequence divergence to characterized proteins and the presence of a significantly longer insertion (Dojkabacteria), different insertion (Absconditabacteria), or the lack of an insertion (Peregrinibacteria). Using biochemical methods, the CP enzyme from a Peregrinibacterial genome was shown to be catalytically active, physiologically complementing autotrophic CO_2-dependent growth in a RubisCO deletion host strain. Based on the integration of biochemical information and metatranscriptomic data, it was suggested that some bacterial CPR use RubisCO to fix carbon dioxide, not as part of a Calvin-Benson Bassham cycle commonly found in Bacteria, but as part of a nucleoside pathway formerly known only in Archaea [51].

In a second example, single cell genomics revealed the Bathyarchaeota contained a high number of extracellular peptidases, and the metabolic capacity to harness energy from amino acid catabolism [31]. However, many of the peptidases present were only distant matches to known proteins, making these annotations uncertain. Heterologous expression of one of these poorly annotated proteins, named Bathyaminopeptidase, in *E. coli* confirmed peptidase activity, and clearly identified substrate specificity for cysteine residues [52]. This residue preference has not been described outside mammals and was not annotated accurately, demonstrating that CP harbor many new surprises about enzyme functions that would not be predicted by annotation alone.

Bringing CP into the cultivation light: thinking beyond pure, rapidly growing laboratory cultures

One of the most globally important cultivation successes involves the isolation of the first member of the Thaumarchaeota (*Nitrosopumilus maritimus*), demonstrating ammonia oxidation in the Archaea [53]. Here, six month long incubation times and filtered aquarium water provided an environment that closely matched native carbon, growth factors, and environmental conditions. A more recent dark matter-to-isolate story involved the isolation of a member of the OP10 [54], which were originally detected in Yellowstone National Park Obsidian Pools (OP) [55]. Colonies were isolated using a mineral salts media with trace amounts of yeast extract, extended incubation time, and most notably gellen instead of traditionally used agar. Another OP10 organism was isolated from aquatic plants by enriching in a dilute, minimal media with ground plant root as the carbon source [56].

Genomic insight has also guided cultivation efforts. For the cultivation of Saccharibacteria within the CPR, metagenomic predictions suggested the bacteria were capable of anaerobic sugar fermentation and possibly host-dependent [57]. He *et al.* (2015) cultivated Saccharibacteria in the presence of an *Actinomyces* host using a media specially crafted to reflect conditions of the oral cavity and metagenomic insight including anoxic conditions and amendment with human amino acids and high sucrose concentrations [49••,58]. Based on microscopic visualization, where multiple Saccharibacteria cells were attached to the bacterial host, the organism was thought to be an epibiont [49••].

In considering the domestication of other CP in the laboratory, and highlighted by the Saccharibacteria success, one must first acknowledge that many of these organisms are not adapted to live in isolation. Genomic insights suggest that for many CPR and some uncultivated archaeal phyla [1] hydrogen metabolism and auxotrophy are common features. Such a dependence on hydrogen may be a barrier to isolation as these organisms depend on the surrounding community to keep hydrogen partial pressure low. Techniques used to cultivate obligate syntrophs [59] or employing hydrogen catalysts may offer methods for decoupling hydrogen dependency.

Additionally, identifying possible hosts via co-correlation of abundance patterns in environmental datasets and using this biomass to enable the growth of CP lineages may afford new opportunities for exploration in the laboratory.

Additionally, many CP microorganisms are not optimized to grow on laboratory timescales, rather adapted to low flux subsurface environments. Consistent with this, some CPR are inferred to have slow growth rates due to a small number of ribosomes [3**]. In a scenario echoing the success of the Thaumarchaeota [53], nutrient conditions that more accurately reflect environmental carbon substrates and concentrations combined with long incubation times may favor the enrichment of currently uncultivated lineages [60]. It can be difficult to replicate the natural environment stably in the laboratory, thus an alternative approach demonstrated by Epstein and colleagues may enrich subsurface CP organisms in their environment [61]. Here researchers used a diffusion chamber that contains sediment-attached cells and allows access to substrates and growth factors at environmentally relevant concentrations. Growing organisms in stable consortia with other bacteria or hosts, recreating the environmental conditions in the laboratory or field, and combining these approaches with higher throughput microcultivation approaches may increase future physiological insights from uncultivated lineages [62].

Concluding remarks

Until recently, culturability was a pre-requisite for genome sequencing, providing full access to the genetic and physiologic information of individual organisms. The advent of metagenomics and single cell genomics has opened a new window into microbial diversity. Metabolic predictions from these genomes provide a glimpse into the metabolism of CP in the environment. While metagenomics can result in genomic supported hypotheses about the overall community metabolism, it may not always resolve pure genomes, instead leading to mosaic or population resolved genomic bins. It also raises the possibility that chimeric pathways are created within a single reconstructed genome. Single cell genomics ensures the purity of a genomic signal, but often provides less comprehensive sampling, greatly hindering the ability to form metabolic hypotheses. Future work combining the two methods, as well as improvements in each separate method, should help to overcome these issues.

It is clear new sequencing methods have illuminated the identity of organisms and their metabolic capacities, placing them in community and ecosystem contexts. However these genomes and their functions encoded within have also provided new perspectives on microbial evolution, taxonomy, and metabolism. As discussed above, recent genomic sampling has reinvigorated debates between the two and three domain models.

Additionally, CP genomic sampling has enabled a new microbial taxonomic nomenclature system, where an isolated organism is no longer the prerequisite for naming and publication. The presence of a high quality draft genome (>95% complete) and a comprehensive description of the genome sequence, in conjunction with the *Candidatus* designation, has been proposed for naming uncultivated organisms [63]. Lastly, while sequenced based predictions are quickly becoming the standard for assessing biogeochemical roles catalyzed by microbes in the environment, expanding these annotations to more divergent lineages has highlighted the paucity of information that can be gleaned from genomic sequencing alone. The abundance of genes with unknown functions, whether in *E. coli* or in genomes from the deepest branches of the tree of life, highlight how much more there is to learn about microbial biochemistry and metabolism. Future integrated methods, that not only include single cell and metagenomics, but also activity and visual measurements like other meta-omic tools (transcripts and proteomics), isotope probing [64,65], high resolution microscopy [47,66–68], and alternative cultivation regimes will afford more detailed investigations into the vast universe of uncultivated microbes.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM *et al.*: **Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling**. *Curr Biol* 2015, **25**.

2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA *et al.*: **Insights into the phylogeny and coding potential of microbial dark matter**. *Nature* 2013, **499(7459)**:431-437.

3. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A,
•• Wilkins MJ, Wrighton KC, Williams KH, Banfield JF: **Unusual biology across a group comprising more than 15% of domain bacteria**. *Nature* 2015, **523(7559)**:208-211.
This paper sampled 789 bacterial genomes from uncultivated phyla and based on these analyses identified the candidate phyla radiation (CPR).

4. Baker BJ, Lazar CS, Teske AP, Dick GJ: **Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria**. *Microbiome* 2015, **3**:14.

5. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ,
• Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF: **A new view of the tree of life**. *Nat Microbiol* 2016:16048.
This paper contains the most up to date phylogenetic tree containing bacterial and archaeal phyla sampled to date.

6. Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer KH, Whitman WB, Euzeby J, Amann R, Rossello-Mora R: **Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences**. *Nat Rev Microbiol* 2014, **12(9)**:635-645.

7. Baker BJ: **Omic approaches in microbial ecology: charting the unknown**. *Microbe* 2013, **8(9)**:353-359.

8. Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T:
• **Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter"**. *Extremophiles* 2014, **18(5)**:865-875.
This previous review on microbial dark matter describes the sequencing technologies in depth and highlights the advances in this field made for extremophiles.

9. Sharon I, Banfield JF: **Genomes from metagenomics**. *Science* 2013, **342**:1057-1058.

10. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF: **Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla**. *Science* 2012, **337(6102)**:1661-1665.

11. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF: **Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla**. *MBio* 2013, **4(5)**:e00708-e00713.

12. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH: **Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes**. *Nat Biotechnol* 2013, **31(6)**:533-538.

13. Seitz KW, Lazar CS, Hinrichs KU, Teske AP, Baker BJ: **Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction**. *ISME J* 2016 http://dx.doi.org/10.1038/ismej.2015.233.

14. Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K,
•• Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJ: **Complex archaea that bridge the gap between prokaryotes and eukaryotes**. *Nature* 2015, **521(7551)**:173-179.
The discovery of Lokiarchaoeta provides new insight on the relationship between Archaea and Eukaryotes.

15. Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ,
•• Golding SD, Tyson GW: **Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics**. *Science* 2015, **350(6259)**:434-438.
This paper uses metagenomics to show that methane metabolism is more widely dispersed than previously thought.

16. Baker BJ, Saw JH, Lind AE, Lazar CS, Hinrichs KU, Teske AP, Ettema JG: **Genomic inference of the metabolism of a cosmopolitan subsurface Archaea, Hadesarchaea**. *Nat Microbiol* 2016:1.

17. Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD, Stepanauskas R, Richter M, Kleindienst S, Lenk S, Schramm A, Jorgensen BB: **Predominant archaea in marine sediments degrade detrital proteins**. *Nature* 2013, **496(7444)**:215-218.

18. Lazar CS, Baker BJ, Seitz K, Hyde AS, Dick GJ, Hinrichs K, Teske AP: **Genomic evidence for distinct carbon substrate preferences and ecolofical niches of Bathyarchaeota in estuarine sediments**. *Environ Microbiol* 2016, **18**.

19. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P,
• Sczyrba A, Woyke T, Soll D, Podar M: **UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota**. *Proc Natl Acad Sci U S A* 2013, **110(14)**:5540-5545.
This paper describes the unique codon usage of candidate phyla SR1, and the extreme diversity of this candidate phylum within the human microbiome.

20. McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS: **Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum**. *Proc Natl Acad Sci U S A* 2013, **110(26)**:E2390-E2399.

21. Yeoh YK, Sekiguchi Y, Parks DH, Hugenholtz P: **Comparative genomics of candidate phylum TM6 suggests that parasitism is widespread and ancestral in this lineage**. *Mol Biol Evol* 2015, **33**.

22. Reddy TB, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification**. *Nucleic Acids Res* 2015, **43(Database issue)**:D1099-D1106.

23. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci U S A* 1977, **74(11)**:5088-5090.

24. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya**. *Proc Natl Acad Sci U S A* 1990, **87(12)**:4576-4579.

25. Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM: **A congruent phylogenomic signal places eukaryotes within the Archaea**. *Proc Biol Sci* 2012, **279(1749)**:4870-4879.

26. Williams TA, Embley TM: **Archaeal "dark matter" and the origin of eukaryotes**. *Genome Biol Evol* 2014, **6(3)**:474-481.

27. Lake JA, Henderson E, Oakes M, Clark MW: **Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes**. *Proc Natl Acad Sci U S A* 1984, **81(12)**:3786-3790.

28. Villanueva L, Schouten S, Sinninghe Damste JS: **Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the 'lipid divide'**. *Env Micro* 2016 http://dx.doi.org/10.1111/1462-2920.13361.

29. Pedersen RB, Rapp HT, Thorseth IH, Lilley MD, Barriga FJ, Baumberger T, Flesland K, Fonseca R, Fruh-Green GL, Jorgensen SL: **Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge**. *Nat Commun* 2010, **1**:126.

30. Kubatko LS, Degnan JH: **Inconsistency of phylogenetic estimates from concatenated data under coalescence**. *Syst Biol* 2007, **56(1)**:17-24.

31. Takai K, Moser DP, DeFlaun M, Onstott TC, Fredrickson JK: **Archaeal diversity in waters from deep South African gold mines**. *Appl Environ Microbiol* 2001, **67(12)**:5750-5760.

32. Jurgens G, Glockner F, Amann R, Saano A, Montonen L, Likolammi M, Munster U: **Identification of novel Archaea in bacterioplankton of a boreal forest lake by phylogenetic analysis and fluorescent in situ hybridization(1)**. *FEMS Microbiol Ecol* 2000, **34(1)**:45-56.

33. Meng J, Xu J, Qin D, He Y, Xiao X, Wang F: **Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses**. *ISME J* 2014, **8(3)**:650-659.

34. Kubo K, Lloyd KG, Biddle FJ, Amann R, Reske A, Knittel K: **Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments**. *ISME J* 2012, **6(10)**:1949-1965.

35. He Y, Li M, Perumal V, Feng X, Fang J, Xie J, Sievert SM, Wang F: **Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments**. *Nat Microbiol* 2016:16035.

36. Scheller S, Goenrich M, Boecher R, Thauer RK, Jaun B: **The key nickel enzyme of methanogenesis catalyses the anaerobic oxidation of methane**. *Nature* 2010, **465(7298)**:606-608.

37. Elshahed MS, Najar FZ, Aycock M, Qu C, Roe BA, Krumholz LR: **Metagenomic analysis of the microbial community at Zodletone Spring (Oklahoma): insights into the genome of a member of the novel candidate division OD1**. *Appl Environ Microbiol* 2005, **71(11)**:7598-7602.

38. Harris JK, Kelley ST, Pace NR: **New perspective on uncultured bacterial phylogenetic division OP11**. *Appl Environ Microbiol* 2004, **70(2)**:845-849.

39. Raghavan R, Hicks LD, Minnick MF: **Toxic introns and parasitic intein in *Coxiella burnetii*: legacies of a promiscuous past**. *J Bacteriol* 2008, **190(17)**:5934-5943.

40. Dueholm MS, Albertsen M, Stokholm-Bjerregaard M, McIlroy SJ, Karst SM, Nielsen PH: **Complete genome sequence of the bacterium Aalborg_AAW-1, representing a novel family within the candidate phylum SR1**. *Genome Announc* 2015, **3(3)**.

41. Hanke A, Hamann E, Sharma R, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R *et al.*: **Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat**. *Front Microbiol* 2014, **5**:231.

42. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM: **Stop codon reassignments in the wild**. *Science* 2014, **344(6186)**:909-913.

43. Nelson WC, Stegen JC: **The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle**. *Front Microbiol* 2015, **6**:713.

44. Lombard V, Golaconda RH, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic Acids Res* 2013, **42**:D490-D495.

45. Anantharaman K, Brown CT, Burstein D, Castelle CJ, Probst AJ, Thomas BC, Williams KH, Banfield JF: **Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum**. *PeerJ* 2016, **4**:e1607.

46. Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, Thomas BC, Handley KM, Mullin SW, Nicora CD, Singh A *et al.*: **Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer**. *ISME J* 2014, **8(7)**:1452-1463.

47. Luef B, Frischkorn KR, Wrighton KC, Holman HYN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF: **Diverse uncultivated ultra-small bacterial cells in groundwater**. *Nat Commun* 2015:6.

48. Giovannoni SJ, Cameron Thrash J, Temperton B: **Implications of streamlining theory for microbial ecology**. *ISME J* 2014, **8(8)**:1553-1565.

49. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu SY,
•• Dorrestein PC, Esquenazi E, Hunter RC, Cheng G, Nelson KE, Lux R, Shi W: **Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle**. *Proc Natl Acad Sci U S A* 2015, **112(1)**:244-249.
Here, genomic insights and host association are used to design cultivation strategies for domesticating members of the candidate phyla in the laboratory.

50. 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* 2014, **37(1)**:35-41.

51. Wrighton KC, Castelle CJ, Varaljay CA, Satagopan S, Brown CT, Wilkins MJ, Thomas BC, Sharon I, Williams KH, Tabita FR, Banfield JF: **RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultured phyla in bacteria**. *ISME J* 2016, **8**:1452-1463.

52. Michalska K, Steen AD, Chhor G, Endres M, Webber AT, Bird J, Lloyd KG, Joachimiak A: **New aminopeptidase from ''microbial dark matter'' archaeon**. *FASEB J* 2015, **29(9)**:4071-4079.

53. Stott MB, Crowe MA, Mountain BW, Smirnova AV, Hou S, Alam M, Dunfield PF: **Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand**. *Environ Microbiol* 2008, **10(8)**:2030-2041.

54. Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA: **Isolation of an autotrophic ammonia-oxidizing marine archaeon**. *Nature* 2005, **437(7058)**:543-546.

55. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR: **Novel division level bacterial diversity in a Yellowstone hot spring**. *J Bacteriol* 1998, **180(2)**:366-376.

56. Tanaka Y, Tamaki H, Matsuzawa H, Nigaya M, Mori K, Kamagata Y: **Microbial community analysis in the roots of aquatic plants and isolation of novel microbes including an organism of the candidate phylum OP10**. *Microbes Environ* 2012, **27(2)**:149-157.

57. Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR: **Dissecting biological ''dark matter'' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth**. *Proc Natl Acad Sci U S A* 2007, **104(29)**:11889-11894.

58. Tian Y, He X, Torralba M, Yooseph S, Nelson KE, Lux R, McLean JS, Yu G, Shi W: **Using DGGE profiling to develop a novel culture medium suitable for oral microbial communities**. *Mol Oral Microbiol* 2010, **25(5)**:357-367.

59. McInerney MJ, Struchtemeyer CG, Sieber J, Mouttaki H, Stams AJ, Schink B, Rohlin L, Gunsalus RP: **Physiology, ecology, phylogeny, and genomics of microorganisms capable of syntrophic metabolism**. *Ann N Y Acad Sci* 2008, **1125**:58-72.

60. Henson MW, Pitre DM, Weckhorst JL, Lanclos VC, Webber AT, Thrash JC: **Artificial seawater media facilitates cultivating members of the microbial majority from the Gulf of Mexico**. *mSphere* 2016, **1(2)**:e00028-16.

61. Kaeberlein T, Lewis K, Epstein SS: **Isolating ''uncultivable'' microorganisms in pure culture in a simulated natural environment**. *Science* 2002, **296(5570)**:1127-1129.

62. Stewart EJ: **Growing unculturable bacteria**. *J Bacteriol* 2012, **194(16)**:4151-4160.

63. Konstantinidis KT, Rossello-Mora R: **Classifying the uncultivated microbial majority: a place for metagenomic data in the *Candidatus* proposal**. *Syst Appl Microbiol* 2015, **38(4)**:223-230.

64. Martinez-Lavanchy PM, Dohrmann AB, Imfeld G, Trescher K, Tebbe CC, Richnow HH, Nijenhuis I: **Detection of monochlorobenzene metabolizing bacteria under anoxic conditions by DNA-stable isotope probing**. *Biodegradation* 2011, **22(5)**:973-982.

65. Peura S, Eiler A, Bertilsson S, Nykanen H, Tiirola M, Jones RI: **Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1**. *ISME J* 2012, **6(9)**:1640-1652.

66. Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, Verberkmoes NC, Hettich RL, Banfield JF: **Enigmatic, ultrasmall, uncultivated Archaea**. *Proc Natl Acad Sci U S A* 2010, **107(19)**:8806-8811.

67. Orphan VJ, House CH, Hinrichs KU, McKeegan KD, DeLong EF: **Multiple archaeal groups mediate methane oxidation in anoxic cold seep sediments**. *Proc Natl Acad Sci U S A* 2002, **99(11)**:7663-7668.

68. Garcia PA, Ge Z, Moran JL, Buie CR: **Microfluidic screening of electric fields for electroporation**. *Sci Rep* 2016, **6**.