Title

Diversity within sparsity: insights into the ecology, metabolism, and evolution of the Candidate Phyla Radiation bacteria

Author

Jaffe, Alexander L.

Publication Date

2021

Peer reviewed|Thesis/dissertation

Diversity within sparsity: insights into the ecology, metabolism, and evolution of the Candidate Phyla Radiation bacteria

By

Alexander L. Jaffe

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian F. Banfield, Chair
Professor Mary K. Firestone
Professor Mary E. Power

Fall 2021

Abstract

Diversity within sparsity: insights into the ecology, metabolism, and evolution of the Candidate Phyla Radiation bacteria

by

Alexander L. Jaffe

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Genomes from metagenomes have made enormous contributions to understandings of the phylogenetic diversity of bacteria and archaea from lineages lacking isolated representatives. This is especially true for members of the Candidate Phyla Radiation (CPR), a highly phylogenetically diverse clade of bacteria with ultrasmall cell sizes and predicted epi-symbiotic or parasitic lifestyles. However, much remains to be learned about the environmental distribution of CPR bacteria, relationships with their microbial hosts, and the ecological and evolutionary factors shaping their key metabolic capacities. Here, genome-resolved metagenomics and comparative genomics approaches were used to shed light on the processes governing gene content in CPR bacteria, looking both across and within natural ecosystems including groundwater, freshwater lakes, soil, and animal microbiomes.

The first chapter of this dissertation resolves a robust internal phylogeny for the CPR bacteria, and uses this phylogenetic framework to quantify the distribution of core metabolic capacities (e.g. carbon metabolism) across the radiation. Analysis of gene patchiness indicated that genetic components of glycolysis have been shaped by vertical transmission, loss, and lateral transfer to differing extents. Intriguingly, similarity in core metabolic platform was decoupled from phylogenetic relatedness, suggesting that gene gain and loss of similar genetic components have likely been commonplace across the CPR bacteria. Extensive gene gain and loss were also evident for NiFe hydrogenases, which may be involved in both energy conservation and sulfur metabolism in diverse CPR bacteria.

The second chapter focuses on RuBisCO, which was found to be patchily distributed across the CPR radiation. CPR bacteria encode a wide diversity of archaeal form III RuBisCO, including a novel clade (‘form III-c’) of sequences which likely function in an archaeal pathway for

nucleotide assimilation that incorporates carbon dioxide. Evolutionary analysis suggested that RuBisCO in CPR has likely undergone extensive lateral gene transfer, including episodes of interdomain exchange that impacted the distribution of RuBisCO forms across the tree of life.

The third chapter examines the way overall gene content (i.e., entire proteomes) in CPR bacteria varies across broad environment types. A subset of lineages within the CPR were examined for linkages between phylogeny, habitat of origin, and gene content to reconstruct the path of transition from the environment to human/animal microbiomes. The results suggest that CPR from animal microbiomes have on average smaller proteomes than their environmental counterparts but are simultaneously enriched for a number of functions that may enable use of habitat-specific resources or tolerance of stressors. However, acquisition of these capacities likely did not enable habitat transitions; instead, we infer that transitions were driven by the suitability of available hosts and subsequently reinforced by gene gain.

Similar themes are explored in the final chapter, which concerns CPR bacteria and their surrounding microbial communities in a permanently stratified lake. Gene content in CPR bacteria is not highly differentiated across the lake's compartments, which experience a gradient of oxygen conditions from relatively saturated to entirely anoxic. This stands in contrast to non-CPR organisms, where metabolic capacities covary with depth and in some cases may be impacted by the availability of light and oxygen. CPR bacteria throughout the water column can significantly contribute to overall metabolic potential for carbon fixation through RuBisCO.

Overall, the results presented in this dissertation shed light on the processes, both ecological and evolutionary, that have acted on CPR gene content over time and contributed to their reduced genomes, variable metabolic platforms, and lifestyles in which they are dependent on other bacteria. These observations are of interest because CPR bacteria organisms likely represent a relatively unique evolutionary trajectory within the domain Bacteria, and thus broaden our overall understanding of 'the rules of life.'

Table of Contents

Introduction

Over the last four decades, molecular sequencing approaches have been critical in expanding our view of the tree of life, particularly in the domains Bacteria and Archaea. Such approaches have helped to overcome the inherent limitations and biases of culture-based approaches, and shed light on the vast diversity of microorganisms that are not easily cultivated under laboratory conditions. While early 'culture independent' methods primarily targeted 'universal' marker genes encoded by both bacteria and archaea, development of a newer set of techniques capturing the entire genomic complement of environmental microbes have catapulted the study of novel microbial lineages into a golden age. In particular, one technique, originally termed community genomics (Tyson et al., 2004) and now referred to as 'genome-resolved metagenomics' (Kantor et al., 2015), has been particularly transformative in that it now routinely provides high-quality draft or complete genomes for uncultivated microbes from any environment type, regardless of complexity. Such techniques have not only continued to flesh out the outer branches of the tree of life but, critically, moved past simple surveys of diversity to shed light on the gene content and metabolic capacities of novel lineages with relevance for human health, biogeochemical cycling, the origins of life, and more.

Among the organisms for which the advent of genome-resolved metagenomics has been crucial are the Candidate Phyla Radiation (CPR), a highly diverse clade of bacteria which is almost entirely uncultivated. Early marker gene studies uncovered the presence of several candidate divisions - including the SR1, OP11, and OD1 - which would only later be recognized as a monophyletic clade of many distinct lineages now called the CPR (Brown et al., 2015). Beyond their enormous phylogenetic diversity, which constitute a significant portion of known bacterial diversity (Hug et al., 2016; Parks et al., 2018; Zhu et al., 2019), the CPR bacteria also possess several intriguing biological characteristics that unite them across large phylogenetic distances. First, CPR bacteria are generally ultra-small, attaining cell sizes on the order of 0.2 μm in diameter or smaller (Luef et al., 2015). Second, based on metabolic reconstructions of complete and near complete genomes, CPR bacteria are generally highly metabolically reduced, and commonly lack the genetic capacities for core cellular processes like the synthesis of amino acids, lipids, and nucleotides (Castelle et al., 2018).

The degree of metabolic reduction observable in the CPR bacteria raised the possibility that these organisms are dependent on other members of microbial communities for key resources. More recent work on a subset of CPR lineages confirmed aspects of this hypothesis, indicating that some CPR adopt episymbiotic lifestyles in which they are at least transiently attached to bacterial host cells (He et al., 2015; Utter et al., 2020). These relationships probably range along a spectrum of obligate symbiosis to parasitism (Moreira et al., 2021). While much remains to be learned, the unique mix of genomic traits and lifestyles suggest that CPR bacteria may constitute

a unique form of symbiosis in the bacterial domain, perhaps paralleled only by a set of genomically reduced lineages in the Archaea, called the DPANN (Castelle and Banfield, 2018).

Paradoxically, while CPR proteomes are reduced compared to other bacteria, they are highly heterogeneous, containing one of the largest unique combinations of protein families observed across the bacterial domain (Méheust et al., 2019). Included in these proteomes are vast numbers of hypothetical protein families without known function (Vanni et al., 2021) as well as those with the potential to impact carbon, sulfur, hydrogen, and nitrogen cycles in the environment (Castelle et al., 2018; Danczak et al., 2017; Wrighton et al., 2012). CPR genomes also contain proteins that may be involved in relationships with their microbial hosts, like the type IV pili proteins found in many genomes so far analyzed (Méheust et al., 2019; Shaiber et al., 2020). Theoretically, CPR proteomes should provide clues to both their evolutionary history and lifestyles. Yet, the balance of vertical inheritance, gene loss, and gene acquisition by later transfer that have shaped the overall gene content in CPR bacteria are so far poorly understood. Further, much remains to be learned about how environmental conditions and host associations select for the distinctive combinations of capacities that define CPR lineages. Many of these key questions can be addressed using genomes recovered from uncultivated consortia.

In this thesis, I draw on genome-resolved metagenomics, comparative genomics, and phylogenomics to analyze the distribution of metabolic capacities across the CPR bacteria and the balance of ecological and evolutionary processes shaping them. First, I build a robust internal phylogeny for the CPR which serves as a framework for analysis of metabolic repertoires, specifically genes involved in glycolysis and a family of NiFe hydrogenases that may be involved in sulfur metabolism. I also examine the balance of lateral gene transfer, vertical inheritance, and gene loss that likely shaped the distribution of these capacities in CPR. Second, I quantify the diversity of RuBisCO genes among the CPR bacteria (and DPANN archaea) and describe the history of lateral gene transfer that may have contributed to their spread across the domains Bacteria and Archaea. Third, I examine the association of phylogeny and gene content with habitat transition in a subset of CPR lineages, in particular the colonization of human and animal microbiomes. I also consider the potential role of host associations in habitat transitions. Finally, I reconstruct numerous genomes of CPR bacteria from a permanently stratified lake and ask how gene content changes with depth and oxygen concentrations. These trends are contrasted with those shaping non-CPR organisms in the same ecosystem.

Overall, this dissertation leverages high-quality genomic information to illuminate the extent and drivers of the ‘diversity with sparsity’ that appears to be characteristic of CPR bacteria.

Acknowledgments

A Ph.D. takes a village, and I am extremely grateful to mine. First and foremost, I would like to thank my graduate advisor, Jill Banfield, for her tireless support and advocacy over the years, for giving me the space to grow and to err as a scientist, for modeling scientific integrity, for teaching me clarity in my writing, and for encouraging cultivation of my artistic passions alongside my scientific ones. It has truly been a pleasure and privilege working with her. I would also like to thank Cindy Castelle and Simonetta Gribaldo, for their invaluable mentorship and friendship that have deeply enriched my time both in Berkeley and abroad.

I must also thank my many other scientific mentors along the way, including but not limited to Eoin Brodie, Matt Traxler, Mary Firestone, Mary Power, Daniel Barsky, Peggy Lemaux, Karthik Anantharaman, Paula Matheus Carnevali, Graham Slater, Francesco Santini, Gabriel Gartner, Shane Campbell-Staton, Andrew Berry, Jeff Kim, and Eric Bapteste. I want to especially thank Jonathan Losos, Steve Haddock, and Mike Alfaro for taking a chance on me early in my scientific career and giving me opportunities and guidance that would profoundly shape my path (and for oh so many letters of recommendation).

I would also like to thank the broader Banfield lab and PMB community at Berkeley, especially my close friend (and occasional collaborator) Alexa Nicolas, whose scientific journey has been very much intertwined with my own since day one of graduate school. I am also grateful to Audra Devoto, Kate Lane, Nikolin Oberleitner, Tyler Arbour, Adair Borges, and Raphaël Méheust, who have grown from co-workers into close friends during my time at Berkeley.

Finally, I would like to acknowledge my parents, Andrea and Toby, and my sister Isabelle. Thank you for supporting and investing in my education, and for encouraging my passion for science from childhood through to my Ph.D. graduation.

1. The rise of diversity in metabolic platforms across the Candidate Phyla Radiation

Alexander L. Jaffe, Cindy J. Castelle, Paula B. Matheus Carnevali, Simonetta Gribaldo, and Jillian F. Banfield

Published in *BMC Biology*, 2020.

A unifying feature of the bacterial Candidate Phyla Radiation (CPR) is a limited and highly variable repertoire of biosynthetic capabilities. However, the distribution of metabolic traits across the CPR and the evolutionary processes underlying them are incompletely resolved. Here, we selected ~1,000 genomes of CPR bacteria from diverse environments to construct a robust internal phylogeny that was consistent across two unlinked marker sets. Mapping of glycolysis, the pentose phosphate pathway, and pyruvate metabolism onto the tree showed that some components of these pathways are sparsely distributed and that similarity between metabolic platforms is only partially predicted by phylogenetic relationships. To evaluate the extent to which gene loss and lateral gene transfer have shaped trait distribution, we analyzed the patchiness of gene presence in a phylogenetic context, examined the phylogenetic depth of clades with shared traits, and compared the reference tree topology with those of specific metabolic proteins. While the central glycolytic pathway in CPR is widely conserved and has likely been shaped primarily by vertical transmission, there is evidence for both gene loss and transfer especially in steps that convert glucose into fructose 1,6-bisphosphate and glycerate 3P into pyruvate. Additionally, the distribution of Group 3 and Group 4-related NiFe hydrogenases is patchy and suggests multiple events of ancient gene transfer. We infer that patterns of gene gain and loss in CPR, including acquisition of accessory traits in independent transfer events, could have been driven by shifts in host-derived resources and led to sparse but varied genetic inventories.

***N.B**. All main figures for this manuscript can be found below in section 1.5. All supplementary files (including figures and tables) can be found online with the published manuscript.*

1.1 Introduction

Metagenomics approaches have been extremely fruitful in the discovery of new lineages across the tree of life (Anantharaman et al., 2016; Brown et al., 2015; Parks et al., 2017; Rinke et al., 2013). Genomes recovered from poorly represented or novel groups have helped greatly to elucidate the evolutionary processes contributing both to broad bacterial and archaeal diversity and also to the distribution of metabolic capacities over various lineages (Adam et al., 2017; Castelle et al., 2018; Spang et al., 2015).

The Candidate Phyla Radiation is a large group of bacterial lineages that lack pure isolate cultures and have been primarily defined through genome-resolved metagenomics (Brown et al., 2015; Luef et al., 2015). While estimates vary depending on the methods used (Hug et al., 2016; Parks et al., 2018), CPR bacteria are predicted to constitute a significant portion of bacterial diversity that is distinct and divergent from other groups (Zhu et al., 2019). Additionally, CPR bacteria generally have relatively small genome and cell sizes, extremely reduced genomic repertoires, and often lack the capacity to synthesize lipids, amino acids, and nucleotides (Brown et al., 2015; Kantor et al., 2013; Luef et al., 2015). The CPR may have diverged early from other bacteria and subsequently diversified over long periods of time, or they may have arisen via rapid evolution involving genome streamlining/reduction (Castelle and Banfield, 2018). Arguing against recent diversification from other bacteria are the observations that CPR bacteria do not share genomic features associated with recent genome reduction, have uniformly small genomes, cluster independently from other metabolically reduced symbionts, and possess metabolic platforms consistent with projections for the anaerobic environment of the early Earth (Castelle and Banfield, 2018; Méheust et al., 2019; Schönheit et al., 2016).

Recently, an analysis of entire proteomes showed that genetic capacities encoded by CPR bacteria are combined in an enormous number of different ways, yet those combinations tend to recapitulate inferred phylogenetic relationships between groups (Méheust et al., 2019). These analyses also revealed that some lineages have relatively minimal core gene sets compared to others within the CPR (Castelle et al., 2018; Méheust et al., 2019), suggesting variation in the degree of genome reduction across the radiation. Additionally, previous work has shown that lateral gene transfer probably underlies distributions of specific protein families in CPR bacteria, including RuBisCO (Jaffe et al., 2016, 2019). The observation that organisms from this group also encode genes for nitrogen, hydrogen, and sulfur compound transformations at a low frequency (Castelle et al., 2018; Danczak et al., 2017; Wrighton et al., 2012, 2014, 2016) raises the possibility that these capacities may also have been shaped by lateral transfer. Overall, the extent to which lateral transfer, genomic loss, and vertical transfer have interacted to shape evolution of metabolic repertoires across the CPR is still unknown (Castelle and Banfield, 2018).

Here, we integrate insights from CPR bacterial genomes from diverse environments with a robustly-resolved internal phylogeny to investigate the processes governing the evolution of metabolic pathways in this group. A key aspect of our approach was the development of custom cutoffs for HMM-based metabolic annotation that are sensitive to the divergent nature of proteins from CPR organisms. We investigated central carbon metabolism (glycolysis and the pentose phosphate pathway), hypothesizing that these pathways may be primarily shaped by vertical inheritance, as well as sparsely distributed traits (nitrogen, hydrogen, sulfur metabolism) that we predicted were shaped by lateral transfer. Mapping of metabolic capacities onto the reconstructed reference tree and gene-species tree reconciliations showed that a mixture of vertical inheritance, gene loss, and lateral transfer have differentially shaped the distribution of functionally linked gene sets. Information about the evolution of gene content may help to shed light on evolutionary scenarios that shaped the characteristics of extant CPR bacteria.

1.2 Materials and Methods

Genome collection and construction of phylogenetic marker sets

We compiled a large set of genomes from metagenomes from CPR bacteria from several previous studies of various environments. We also binned an additional set of genomes from metagenomes previously generated from sediment from Rifle, Colorado (Anantharaman et al., 2016), groundwater from Crystal Geyser (Probst et al., 2017, 2018), a cyanobacterial mat from the Eel River network in northern California (Bouma-Gregson et al., 2019), groundwater from a cold sulfide spring in Alum Rock, CA, and human saliva. Binning methods and taxonomic assignment followed Anantharaman et al. (2016). The total set was initially filtered for genomes that had been manually curated by any method to reduce the occurrence of misbinning, yielding a starting set of approximately 3,800 genomes. We next computed contamination and completeness for all genomes using a set of 43 marker genes sensitive to described lineage-specific losses in the CPR (Anantharaman et al., 2016; Brown et al., 2015) using the custom workflow in checkm (Parks et al., 2015). Results were then used to secondarily filter the genome set to those with >= 70% of the 43 marker genes present and <=10 % of marker genes duplicated. The resulting ~2300 genomes were de-replicated at 95% ANI using dRep (-sa 0.95 -comp 70 -con 10) (Olm et al., 2017), yielding a set of 991 non-redundant genomes used for downstream analysis.

We re-predicted for each genome using Prodigal (“single” mode) (Hyatt et al., 2010), adjusting the translation table (-g 25) for CPR lineages (Gracilibacteria and Absconditabacteria) known to utilize an alternative genetic code. Next, we assembled two sets of HMMs, representing the 16

syntenic ribosomal proteins (rp16) and, separately, the two subunits of RNA polymerase (RNAp), from the TIGRFAMs and Pfams databases and ran each against predicted proteins using HMMER v3.1b2 (http://hmmer.org). To maximize extracted phylogenetic information, including partial genes with robust homology to the marker genes, we set custom thresholds for each HMM using trees generated from all significant ($e < 0.05$) hits to a given HMM (aligned using MAFFT, tree inference with FastTreeMP) (Katoh and Standley, 2013; Price et al., 2010). Thresholds were usually set at the highest bitscore attained by proteins outside the clade of interest (Additional file 2, Figure S1), which were verified with BLASTp. HMM results and thresholds were visualized by in bitscore vs. e-value plots (Additional file 2, Figure S1ab). Phylogenetic analysis of HMM hits revealed that many proteins below model-specific thresholds were legitimate, often partial, hits to the targeted HMM (Additional file 2, Figure S1b).

Next, we curated phylogenetic marker sets for both rp16 and RNAp by addressing marker genes present in multiple copies in a given genomic bin. Multi-copy genes can result from remnant contamination after filtering, ambiguous bases in assembly leading to erroneous gene prediction (Parks et al., 2015), or legitimate biological features. We first identified marker genes fragmented by errors in gene prediction by searching for contiguous, above-threshold hits to the same HMM on the same assembled contig. This issue was particularly prevalent for rpoB and rpoB', possibly due to repetitive regions in that gene impacting accurate assembly. For upstream fragments, we removed protein residues after stretches of ambiguous sequence to avoid introducing mis-translated bases into the alignment stage while maximizing phylogenetic information. If additional stretches of ambiguous sequence were present in downstream fragments, we removed them. Finally, we built a corrected, non-redundant marker set for each genome by selecting the 16 ribosomal proteins and, separately, 2 RNA polymerase subunits, that first maximized the number of marker genes on the same stretch of assembled DNA and, secondarily, maximized the combined length of encoded marker genes.

Species tree inference, curation, and analysis

Results for each marker gene in the rp16 and RNAp sets were individually aligned with MAFFT (Katoh and Standley, 2013) and subsequently trimmed for phylogenetically informative regions using BMGE (-m BLOSUM30) (Criscuolo and Gribaldo, 2010). Gene trees for each marker were then constructed using IQTREE's model selection and inference (-m TEST -nt AUTO -st AA) and manually inspected for major incongruencies.

In preparation for creating a concatenated alignment for each marker set, we next extracted corresponding rp16 and RNAp marker sets for a diverse bacterial outgroup consisting of ~170 bacterial genomes from GenBank sampled evenly across characterized taxonomic divisions. We then merged the outgroup dataset with the existing marker gene sets, individually aligning hits for each marker gene and trimming them as described above. We then concatenated individual

protein alignments, retaining only those with both RNAp subunits and at least 8 of 16 syntenic ribosomal proteins. Maximum likelihood trees were inferred for both the concatenated rp16 (1427 AA) and RNAp (1652 AA) sets using ultrafast bootstrap and IQTREE's extended FreeRate model selection (-m MFP -st AA -bb 1500) (Hoang et al., 2018; Kalyaanamoorthy et al., 2017; Nguyen et al., 2015), given the importance of allowing for site pattern heterogeneity in concatenated alignments (Wang et al., 2019).

We next identified phylogenetic outliers in the resolved maximum likelihood topologies by searching for genomes that did not form a monophyletic clade with other organisms of the same taxonomy. These genomes, potentially due to mixed phylogenetic signal or undersampling, were retained only if they were assigned to a previously described novel lineage, or formed a conserved, uncharacterized clade with >1 member in both rp16 and RNAp trees. Genomes that did not fit these criteria were pruned. Concatenated trees were then re-inferred with the modified genome set. Where possible, we manually curated taxonomic assignments for genomes that clearly resolved within monophyletic clades of different taxonomic classification in both the rp16 and RNAp trees. Finally, we assessed broad-scale phylogenetic patterning within the CPR by examining the distribution of ribosomal proteins L1 and L9 employing the same HMM-based approach as described above.

Metabolic annotation, analysis, and gene tree inference

To probe metabolism within CPR bacteria, we assembled a broad set of HMMs from TIGRFAMs (tigrfams.jcvi.org/cgi-bin/Listing.cgi), Pfam (pfam.xfam.org), and a previous publication (Anantharaman et al., 2016) representing metabolisms relevant for biogeochemical cycling and energy production in this clade (Castelle et al., 2018; Kantor et al., 2013; Wrighton et al., 2012) (Additional file 3, Table S2). We interrogated protein sequences from each genome with the HMM set using HMMER and set custom bitscore thresholds as described above to ensure that divergent but functionally valid proteins were retained. Model-specific thresholds were often much higher than maximum bitscores of hits, even in cases where we were able to assign putative function to relatively high scoring clusters through BLAST and phylogenetic analyses. In a few cases (PRPP, PEP synthase, PGI, ROK family), we secondarily annotated HMM protein hits with additional Pfam domains or manually inspected placement within a reference tree to guide setting of accurate manual cutoffs. These additional domain HMMs and all custom thresholds are specific to this dataset and are listed in Additional file 3, Table S2. If a protein had multiple above-threshold hits to a set of HMMs, we selected the HMM with the highest bitscore. We additionally selected the highest-scoring HMM hit within a genome bin for each HMM to generate a final set of metabolic markers for downstream analysis.

We next analyzed distributions of metabolic capacities in two ways: first, we created a presence/absence matrix for all metabolisms with at least one hit among the genome set,

combining profiles for HMMs representing the same function (e.g. PGI, FBA, RuBisCO) into a single merged category. We then filtered the matrix to include only lineages with eight or more genomes and traits that were detected at least three times over all genomes. Finally, we averaged presence/absence across lineages, generating a frequency at which that trait was present among genomes of a particular taxonomy. We then used this information to generate a Bray-Curtis distance matrix using the ecopy package in Python. Finally, we performed a principal coordinates analysis using scikit-bio learn and plotted the resulting axes to examine clustering and variation within and among metabolic platforms of CPR bacteria. Second, we measured phylogenetic conservation and patchiness over the rp16 tree using the consenTRAIT algorithm (Npermutations = 1000, count_singletons = F, min_fraction = 0.90) (Martiny et al., 2013) as implemented in the R package castor and consistency index (CI) as implemented in the R package phangorn and proposed in (Mendler et al., 2019) (sitewise = T). We integrated these two metrics to generate an "evolutionary profile" for each gene.

To assess how patchiness of given metabolisms varied with genome completeness, we sub-sampled the genome set iteratively at increasing thresholds from 70% through 95%, and for each iteration, pruned down the existing rp16 reference tree to include only those genomes. We then re-computed patchiness for each trait as done previously. For the comparative analysis of patchiness among glycolytic enzymes, we gathered all non-CPR bacterial genomes from three major studies of groundwater microbial communities from which the majority of CPR bacterial genomes were assembled (Anantharaman et al., 2016; Probst et al., 2017, 2018). To increase our phylogenetic sampling, we combined this set with additional genomes from metagenomes from a large scale study of multiple environments that used similar methods (Parks et al., 2017) and selected three major lineages with adequate size to use for downstream analysis (Proteobacteria, Firmicutes, and Bacteroidetes). We calculated completeness and contamination for the non-CPR genomes using the same set of 43 markers as before and de-replicated them at 95% ANI with dRep, using the calculated completeness and contamination to again filter at 70% completeness and 10% contamination. We next extracted the rp16 phylogenetic markers using a similar approach (though, for simplicity, HMMs were thresholded using model-specific noise cutoffs) and processed as before. Next, a random subsample of 50 CPR bacterial genomes was taken and their phylogenetic markers were separately concatenated with those of each non-CPR lineage. Alignment, alignment trimming, and tree building was performed as previously for each set of sequences. For each of the three non-CPR groups, HMMs corresponding to glycolytic enzymes were run against predicted proteins and manually re-thresholded. Finally, genome sets for each lineage were sub-sampled at increasing completeness thresholds as for CPR bacteria, and patchiness was computed for each glycolysis enzyme over each tree as above. Results were combined with those obtained for CPR organisms and visualized.

To build reference protein sets for the metabolic genes of interest, we queried proteins from the set of ~170 bacterial reference genomes with same HMMs described above and applied the

model-specific noise cutoff (for Pfam or TIGRFAMs HMMs) or the published cutoff (for custom HMMs). These proteins were then concatenated with the corresponding above-threshold hits from the CPR bacterial genomes and aligned as described above with MAFFT. Additionally, for four HMMs corresponding to glycolytic functions (PF06560, TIGR02128, TIGR00306, TIGR00419) we also queried a set of proteins from ~300 archaeal reference genomes assembled in a similar fashion to the bacterial reference set. Resulting protein hits were concatenated with the bacterial sequences. For all single-gene alignments, columns with 95% or more gaps were trimmed using Geneious. Maximum-likelihood gene trees were then inferred using IQ-TREE with the following parameters: -m TEST -st AA -bb 1500. Trees were rooted on the largest monophyletic group of reference sequences present in the topology; if multiple monophyletic groups of reference sequences were present, trees were rooted at the midpoint.

To generate a gene tree for the NiFe hydrogenases, we assembled a comprehensive reference set of large subunit sequences from several published sources (Constant et al., 2011; Greening et al., 2016; Matheus Carnevali et al., 2019), dereplicated them at 95% amino acid identity using usearch --cluster_fast, and concatenated the resulting centroids with large subunit sequences recovered from CPR bacteria. Sequences were aligned, alignments were trimmed, and the gene tree was inferred as described above for other metabolic genes. The trimmed alignment is available in Additional file 4. We next manually identified sequences within the immediate genomic context of 3b-related catalytic subunits that also scored highly against HMMs for anaerobic sulfite reductase A/B, as described previously for subunits in the Group 3b hydrogenases of Pyrococcus furiosus (Ma et al., 1993; Pedroni et al., 1995) and searched them for conserved domains in phmmer (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer). We identified one iron-sulfur cluster and one NAD binding domain that were conserved among these proximal proteins (Additional file 3, Table S2), and then queried all proteins from CPR bacteria with these HMMs to identify putative 3b-related subunits across the entire genome set. We performed the same search for an additional Pfam domain associated with the 3b-hydrogenase small subunit (Additional file 3, Table S2). For all three HMMs, manual thresholds were set using the paired visualization-phylogenetic approach described above. Finally, presence/absence of putative subunits were mapped onto the resolved tree of large-subunit sequences to examine patterns of association with phylogenetic clades of 3b-related hydrogenase using iTol (Letunic and Bork, 2016).

For the genomic context analysis of 3b-related forms, we gathered protein sequences within a 20 ORF radius (or less, if the scaffold ended) in both directions of the identified large subunits. Each ORF was assigned a genomic position relative to the large subunit (position 0). All recovered proteins were concatenated into a single file and passed through a two-part, de novo protein clustering pipeline recently applied to CPR genomes, in which proteins are first clustered into ‘subfamilies’ and highly similar/overlapping subfamilies are merged using and HMM-HMM comparison approach (--coverage 0.50) (Méheust et al., 2019). Recovered protein families were

compared with subunit HMM results and linked if the majority of proteins within the family had above-threshold hits to a given HMM. An alignment and gene tree for those proteins labelled as the small subunit hydrogenase (fam019) were made as described above.

Finally, counts for genes encoding the recovered families were plotted as a function of their relative position to the focal catalytic subunit of the hydrogenase across all CPR bacterial genomes. This was performed only if there were instances of the genes on the same strand (as predicted by Prodigal) as the large subunit hydrogenase. The relative positions of genes were multiplied by their strand orientation such that a negative position would signify being "upstream" of the focal catalytic subunit, whereas a positive position would signify being "downstream." Positions were also adjusted in several cases where the focal subunit was split into multiple consecutive fragments, possibly due to local assembly errors.

1.3 Results

A robust reference phylogeny for the CPR

We gathered a large set of curated genomes of CPR bacteria from diverse environments, including both previously published and newly assembled sequences (Materials and Methods). Quality filtration of this curated genome set at ≥70% completeness and ≤10% contamination and subsequent de-replication yielded a non-redundant set of 991 genomes for downstream phylogenetic and metabolic analysis (Additional file 1, Table S1). To improve recovery of phylogenetic markers from the collected set of genomes, we combined visualization of HMM bitscores with a phylogenetic approach to set sensitive, custom thresholds for two independent sets of markers composed of 16 syntenic ribosomal proteins (rp16) and the two RNA polymerase subunits (RNAp) (Materials and Methods; Additional file 2, Figure S1). Phylogenies based on these two marker sets were generally congruent for deep relationships within the CPR, with both trees supporting the distinction of CPR from the bacterial outgroup and the monophyly of the Microgenomates and Parcubacteria superphyla, respectively (Figure 1a; Additional file 2, Figure S2). Some clades were also supported by the absence of particular ribosomal proteins - the Microgenomates, along with the Dojkabacteria and Katanobacteria, lacked ribosomal protein L9 (rpL9), while a subset of Parcubacteria lacked the ribosomal protein L1 (rpL1), as observed previously (Brown et al., 2015). Our results also suggested the presence of four generally well-supported (≥95% ultrafast bootstrap in three of four cases), monophyletic subgroups within the Parcubacteria (Figure 1a, Parcubacteria 1-4). Although internal relationships between these subgroups varied slightly between trees (Additional file 2, Figure S2), in both cases Parcubacteria 1 (comprising 9 lineages) was the deepest clade, whereas Parcubacteria 4 (10 lineages**)** was the most shallow (Figure 1a). Ten other lineages of Parcubacteria formed paraphyletic clades outside of these subgroups. We also show that Dojkabacteria (WS6),

Katanobacteria (WWE3), Peregrinibacteria, Kazanbacteria, and Berkelbacteria are among the most deeply-rooting clades outside the established superphyla (Figure 1a).

CPR bacteria encode variable and overlapping metabolic repertoires

We next leveraged the robust reference tree of the CPR to evaluate the distribution and combinations of capacities across the radiation. While CPR bacteria lack some core biosynthetic capacities, they do in fact possess numerous metabolic capacities involved in carbon, hydrogen, and possibly sulfur and nitrogen cycling (Castelle et al., 2018; Danczak et al., 2017; Kantor et al., 2013; Wrighton et al., 2012). We focused on these traits for our subsequent analysis, reasoning that they are most likely to impact the ability of CPR bacteria to derive energy from organic compounds and contribute to biogeochemical transformations in conjunction with their hosts and other community members. To overcome the challenges inherent to metabolic annotation of divergent lineages and minimize the chance of false negatives, we extended our custom HMM thresholding approach to the selected set of biogeochemically-relevant traits (Materials and Methods; Additional file 2, Figure S3; Additional file 3, Table S2) and mapped the resulting binary presence/absence profiles for specific functionalities onto the reconstructed rp16 tree. Looking across the selected traits, we observed a high degree of variation in the overall repertoires of lineages within the CPR, including some with extremely minimal metabolic complements like Dojkabacteria and Gracilibacteria. This is consistent with both observations from genomic studies of these lineages (Kantor et al., 2013) as well as more recent insights examining entire proteomes (Méheust et al., 2019).

An important open question is whether clades of CPR bacteria within broad phylogenetic groupings possess similar combinations of metabolic capacities. To investigate this, we used the distributions of the targeted traits to compute the frequency at which each trait was found within lineages. We then generated a distance matrix from the results and performed a principal coordinate analysis to visualize clustering of lineages based on the similarity of their overall metabolic platforms (Figure 1b, Materials and Methods). We reasoned that genes missing due to genome incompleteness could impact clustering, particularly for small lineages with only several members. Thus, we restricted the analysis to those groups with at least 8 member genomes. The results suggest that member lineages within some broad phylogenetic groupings are metabolically similar (e.g., Parcubacteria 3 and 4) but others clustered more closely with lineages that are distantly related. For example, lineages within the Microgenomates and Parcubacteria 1 were highly dispersed across the axes of variation (Figure 1b), suggesting that member groups encode highly variable combinations of traits.

Functionally linked metabolic genes display different evolutionary profiles

The observation that distributions of traits are variable and potentially decoupled from phylogenetic relatedness raises the possibility that more complex, enzyme-specific patterns might underlie processes contributing to metabolic diversity in the CPR. To address this, we drew upon trait distributions to compute two metrics across the reference tree - the first to quantify the average branch length of clades in which a trait is conserved (phylogenetic depth) and the second to analyze trait patchiness, related to the number of gains/losses of a binary trait over a tree (Materials and Methods) (Mendler et al., 2019). Generally, traits with a high phylogenetic depth correspond to those that are conserved in more deeply-rooting clades, whereas traits with lower depth correspond to those that occur primarily among shallow clades. Similarly, high patchiness is expected when a given trait is more randomly dispersed across a clade, whereas traits with low patchiness scores correspond to those that are highly conserved within groups. These two metrics are therefore complementary and were integrated to create an 'evolutionary profile' for each trait. Among CPR bacteria, we observed that an increase in phylogenetic depth generally correlates with a decrease in patchiness (Figure 2a). High-depth traits also corresponded to larger protein families more frequently observed across the radiation (family size), though several smaller protein families (phosphate acetyltransferase, AMP phosphorylase, RuBisCO) reached relatively high phylogenetic depths because they were conserved in deeply-rooting clades like the Dojkabacteria and Peregrinibacteria. On the other hand, hydrogen/sulfur metabolism, acetate/lactate metabolism, and the oxidative pentose phosphate pathway exhibited relatively high patchiness and low phylogenetic depth, consistent with their sparse but also wide distributions across distantly related groups (Figure 2). Intriguingly, some traits displayed a relatively low phylogenetic depth but were less patchily distributed than would be predicted from the overall trend (e.g., genes involved in aerobic metabolism).

As our analysis drew upon draft genomes (≥ 70% of genome markers present), it is possible that genome incompleteness impacted estimates of presence/absence for metabolic genes. We reasoned that the patchiness metric in particular would be sensitive to this issue, as it is computed (unlike the phylogenetic depth metric) using the number of state transitions of binary characters over the tree. To address this possibility, we undertook a parameter sensitivity analysis that tested the robustness of patchiness scores to genome completeness. We iteratively subsampled the genome set at increasing thresholds of completeness and recomputed trait patchiness over a pruned version of the reference tree (Materials and Methods), observing only modest decreases in patchiness for individual components of the target pathways as genome completeness increased to 95% (Additional file 2, Figure S4a), Specific glycolytic enzymes showed a similar pattern, with the exception of TIM, GADPH, and PGK, which were already essentially universal in CPR bacteria at the lowest completeness threshold (Figure 3b; Additional file 2, Figure S4ab). Taken together, these results suggest that while genome incompleteness

probably impacts calculations of patchiness to a small extent, our observations are mostly due to a biological, not methodological, signal.

Surprisingly, metabolic genes within the same pathway often showed disparate evolutionary profiles - for example, enzymes involved in glycolysis displayed a wide range in depth and patchiness (Figure 2a). Similar patterns were observed for the nucleotide salvage pathway and non-oxidative pentose phosphate pathway (Figure 2a). These observations might suggest that evolutionary histories of the component enzymes of these pathways are decoupled; specifically, that traits with high phylogenetic depth and low patchiness are likely ancient and conserved (low loss), whereas those with lower depth and higher patchiness are more likely to have been impacted by loss and/or horizontal gene transfer. To test these hypotheses, we investigated two cases in more detail - first, glycolysis, as an example of a core pathway with a wide range of phylogenetic depth and patchiness among component enzymes, and, second, NiFe hydrogenases, an accessory trait with high patchiness and low depth (Figure 2a, Figure 3a).

Gene trees for glycolytic enzymes reflect different patterns of gene loss and transfer

We first examined glycolysis, noticing that three enzymes from the central part of the pathway - triose phosphate isomerase (TIM), glyceraldehyde 3-phosphate (GAPDH), and phosphoglycerate kinase (PGK) were found in nearly all CPR bacteria with little to no patchiness (Figure 3a). With the possible exception of ultra-reduced forms like the Gracilibacteria, which is represented by one complete, curated genome that completely lacks the glycolysis pathway (Sieber et al., 2019), the absence of these enzymes in a very small number of genomes is likely due to missing genomic information. A second group of enzymes, comprised of fructose bisphosphate aldolase (FBA), enolase (ENO), and phosphoglucose isomerase (PGI), was instead generally more patchily distributed among CPR bacteria and missing in some lineages. Phosphoglycerate mutase (PGM), responsible for converting Glycerate 1,3-P2 to Glycerate 2P in lower glycolysis, fell between the two groups - while present in deeply-rooting clades (thus, a high phylogenetic depth), it is absent in several shallow clades of Parcubacteria, possibly because these forms were too divergent to be recovered with the manual HMM threshold. Finally, several enzymes, including glucokinase/hexokinase, phosphofructokinase (PFK), and pyruvate kinase, exhibited profiles that were highly patchy and lower-depth among CPR lineages. Notably, these enzymes are thought to catalyze irreversible reactions and thus act as important sites of regulation for metabolic flux (Bräsen et al., 2014; Castelle et al., 2018). In particular, glucokinase/hexokinase and PFK were found very infrequently in CPR bacteria, though many have the potential to bypass PFK using a metabolic shunt through the non-oxidative pentose phosphate pathway (Fig 3b) (Kantor et al., 2013). To confirm this result, we also searched genomes for alternative forms of PFK, finding that while some CPR bacteria encode ROK (repressor, open reading frame, kinase) family proteins (TIGR00744), we could not establish close phylogenetic relationships to family members functioning as putative glucokinases. Likewise, we found no evidence for the

alternative versions of ADP-dependent glucokinase/phosphofructokinase employed in the modified glycolytic pathways of some archaea (PF04587) (Tuininga et al., 1999).

To further test the impact of genome incompleteness on the apparent patchiness of glycolytic enzymes across the CPR and investigate whether this pattern is unique, we undertook a comparative analysis of other major bacterial phyla. We reasoned that if high patchiness of glycolysis in CPR bacteria is due primarily to genome incompleteness, enzymes from these organisms should have similar patchiness to their counterparts in genomes from other groups with more typical metabolic platforms. On the contrary, if our initial results are indicative of a true biological signal, we would expect enzymes of CPR bacteria to show consistently higher patchiness than observed across other bacterial phyla. We gathered several thousand genomes from metagenomes that were assembled and binned with similar methods to those used to reconstruct genomes of CPR bacteria, corresponding to large phylogenetic groups - Proteobacteria (n=1090), Firmicutes (n=680), and Bacteroidetes (n=578) (Materials and Methods). To ensure comparability of our results, we used the same methodology for genome completeness assessment, metabolic annotation, and analysis of glycolysis as for the CPR. We show that individual glycolysis enzymes from CPR bacteria generally attain the highest patchiness among the lineages examined, particularly for the enzymes at pathway termini (Additional file 2, Figure S4c). Exceptions include the three glycolysis enzymes that we consider to be a core, essentially universal module across the CPR (TIM, GAPDH, and PGK), and for enolase, where Firmicutes also showed significant patchiness (Additional file 2, Figure S4c). These findings further confirm that the degree of patchiness observed for glycolytic enzymes in CPR bacteria is robust to issues arising from genome incompleteness and is unusual across major bacterial lineages.

To further investigate which specific processes impacted the disparate evolution of glycolytic enzymes in CPR bacteria, we reconstructed single-protein phylogenies and performed gene-species tree reconciliations (Materials and Methods). We reasoned that enzymes whose evolutionary histories were shaped primarily by vertical transfer paired with genomic loss, rather than transfer, would display phylogenetic patterns roughly congruent with our resolved reference species tree, whereas those impacted by horizontal transfer (with either CPR or non-CPR groups) would exhibit incongruent relationships. Gene trees for well-conserved glycolytic capacities like TIM and PGK generally recapitulated phylogenetic groupings at a coarse level (Figure 3c; Additional file 2, Figure S5). However, even for these enzymes, inconsistencies with the species tree were present - for example, some triose phosphate isomerase (TIM) sequences from the Microgenomates, Katanobacteria, and Peregrinibacteria clustered with archaeal reference sequences (Figure 3c). These results were replicated across multiple genomes, and the phylogenetic associations of surrounding ORFs on the same scaffold were verified by BLAST to ensure that the scaffold originated from a CPR organism. Similarly, in the enolase phylogeny, large, monophyletic clusters representing sequences from the Microgenomates and Parcubacteria

1 were resolved; however, other sequences from the Microgenomates and many from Parcubacteria 3 and 4 fell into smaller, fragmented groups that clustered with more distantly related lineages (Figure 3c). Gene trees for other glycolytic enzymes displayed a range of patterns (Additional file 2, Figure S5). On the whole, gene-species tree inconsistencies suggest that lateral gene transfer, either between CPR bacteria and other taxa or among different CPR bacteria, has also impacted the evolution of glycolytic enzymes alongside the gene loss apparent from presence/absence profiles (Figure 3a).

Supporting the possibility of horizontal gene transfer is the observation that multiple distinct enzyme forms underlie the distributions of several glycolytic functions. For example, we recovered unique hits to three individual HMMs representing various versions of PGI - one describing a general, cross-domain version (PF00342), another a bifunctional PGI/phosphomannose isomerase present in some bacteria and archaea (TIGR02128) (Hansen et al., 2004), and, finally, an unrelated cupin-based enzyme originally described from archaea (PF06560) (Hansen et al., 2005; Verhees et al., 2001). Interestingly, all three enzymes were scattered across the broad CPR groups, though very few CPR organisms (~2% of genomes) encoded more than one version. About 15 genomes, mostly belonging to the Nealsonbacteria, encode only the cupin-related version. These sequences form a sibling clade to those from archaeal reference genomes in the corresponding gene tree (Additional file 2, Figure S5). Sequences from CPR organisms with highest similarity to archaeal versions were also recovered for PGM (TIGR00306) and for TIM, although in the latter case sequences did not correspond to a separate HMM (Additional file 2, Figure S5). Similarly, while most CPR bacteria encode a class II FBA enzyme, some, particularly Kaiserbacteria and Woesebacteria, also encode a class I enzyme that functions via a distinct reaction mechanism (Cooper et al., 1996). Finally, in gene tree reconstructions for the class II aldolase, sequences from CPR bacteria do not appear to be monophyletic, with small subgroups dispersed among sequences from other bacteria. Taken together, these results indicate that enzymes of multiple evolutionary origins underlie the distributions of core carbon metabolism, and support the idea that their distributions have been shaped by episodes of lateral gene transfer, potentially from non-CPR bacteria or archaea.

CPR bacteria encode phylogenetically distinct forms of Ni-Fe hydrogenases with variable genomic context

We next investigated the impact of lateral transfer on metabolisms sparsely distributed across the CPR, focusing on the NiFe hydrogenases as a case study because of their possible role in hydrogen economy and/or electron flux (Castelle et al., 2018; Wrighton et al., 2012). Most sequences from CPR bacteria were previously reported to fall within the Group 3b hydrogenases, cytoplasmic enzymes that may catalyze the reversible oxidation of H_2 coupled to regeneration of NADPH or reduction of polysulfide when available (Silva et al., 2000; van Haaster et al., 2008). Here, a revised gene tree that broadly samples the CPR reveals the presence of two subclades,

which we term *hyd1* and *hyd2,* forming a larger clade of Group 3 hydrogenase from CPR organisms (Figure 4a). Both groups are related to, but distinct from, Group 3b versions in other bacteria and archaea, particularly *hyd2*, which is separated from its sibling clades by a relatively long branch (Figure 4a).

Biochemically characterized Group 3b NiFe hydrogenases are known to be tetrameric enzymes (Pedroni et al., 1995). To examine whether subunit associations were consistent across hydrogenase classes, we probed the genomic context of the large subunits from CPR bacteria using a paired HMM-protein clustering approach (Materials and Methods). Intriguingly, while both enzyme types were generally associated with the small subunit hydrogenase (fam019) in addition to the catalytic subunit, only *hyd1* co-located with genes encoding protein families resembling the two other subunits involved in $NAD(P)^+$-binding (gamma, fam034) and electron transfer (beta, fam012) (Figure 4a). HMM searches revealed that these subunits also have homology to anaerobic sulfide reductase A and B, suggesting that the entire complex could be involved in sulfur metabolism through the reduction of reduced sulfur compounds like polysulfide (Ma et al., 1993; Pedroni et al., 1995). However, in some cases, the gamma and beta subunits were not immediately upstream from the gene encoding the small subunit (Figure 4b), and, in others, were not detected at all (Figure 4a). This inconsistency might be due to genome incompleteness or lineage-specific losses within the *hyd1* clade.

Although genomes with *hyd2* also encoded the small subunit protein (fam019), the sequences were consistently truncated (mean 164 amino acids) relative to those associated with *hyd1* and non-CPR bacteria (mean 250 amino acids) (Additional file 2, Figure S6a) (Vignais and Billoud, 2007). Both forms also shared fam002 in their genomic context, some members of which displayed homology to the hydrogenase-associated chaperone *hypC*. Outside these families, immediate genomic context differed for *hyd2*: while an HMM search recovered sequences with the NAD-binding motif (gamma subunit) in the vicinity of some *hyd2*, protein clustering showed that these proteins were neither proximal to, nor on the same strand as the catalytic subunit (Figure 4ab). However, some members of fam013 that were in the genomic context of *hyd2* apparently possessed an NAD(P)-binding domain situated within a larger FAD-binding domain (PF07992). Similarly, while HMM searches did not recover evidence for a putative beta subunit near *hyd2*, we found one protein family (fam390) in proximity to a subset of *hyd2* that contained one of two iron-sulfur-binding domains. These domains were distinct from those associated with the putative beta subunit near *hyd1*. Ultimately, it is unclear whether *hyd2* consistently possesses (or lacks) the gamma and beta subunits and thus its function remains uncertain.

Intriguingly, both *hyd1* and *hyd2* were dispersed across many lineages of the CPR, and some lineages contained both subtypes in closely related but distinct genomes (Figure 4a). For example, genomes from the Roizmanbacteria, which harbored the largest total number of Group 3b-related NiFe hydrogenases (*n=15*), individually contained either *hyd1* or *hyd2* sequences.

Mapping of genome taxonomy onto the 3b-related hydrogenase tree confirmed incongruencies with the CPR species tree (Figure 4a). A similar pattern was observed for sequences from CPR bacteria that fell within a subclade of Group 4 references representing energy-converting hydrogenases-related complexes (Ehr). Notably, the sequences from CPR bacteria were monophyletic and clustered separately from other Ehr proteins, although they also lacked the cysteine residues that bind the metal cofactors in other Group 4 enzymes (Additional file 2, Figure S6b). This observation suggests that Ehr proteins from CPR organisms likely cannot interact with H_2.

1.4 Discussion

Initially described as a radiation of phylum-level clades based on analyses of 16S rRNA divergence (Brown et al., 2015), the CPR was initially suggested to comprise at least 15% of bacterial phylum-level groups (Brown et al., 2015). Subsequent analyses have suggested that its scale potentially matches that of all other bacterial diversity (Hug et al., 2016). Attempts to adjust for lineage specific evolutionary rates have suggested the collapse of the CPR into a single phylum (Parks et al., 2018), but more recent analyses with balanced taxonomic sampling continue to depict it as a large component part of bacterial diversity (Zhu et al. 2019). Here, we combined new and previously reported genomes to construct a robust reference phylogeny for the CPR using two unlinked, concatenated marker sets (Figure 1a, Additional file 2, Figure S2). The reconstructed trees are generally consistent with, and more clearly define, the topology originally described for the CPR (Brown et al., 2015), although definitive resolution of some deep nodes, particularly those connecting divergent groups like the Saccharibacteria, Gracilibacteria, and Absconditabacteria (SR1), remain elusive, possibly due to undersampling of the latter two lineages. Both gene trees support the presence of several monophyletic subgroups within the Parcubacteria, motivating subdivision of this large clade into smaller, taxonomically-relevant units.

Here, we evaluated metabolic platforms across the CPR by mapping genomically-encoded functions onto the reference tree. Analysis of metabolic capacity among CPR organisms presents several challenges, primarily due to the fact that homologs of metabolic genes are often highly divergent compared to known reference sequences. Our custom approach for determining suitable cutoffs for HMMs indicates that manual threshold curation is important when proteins are only distantly related to biochemically characterized versions (Additional file 2, Figures S1 and S3). We found that metabolic platforms for CPR lineages only partially mirror phylogenetic relationships (Figure 1ac), at least for the subset of metabolic traits examined here. In other words, phylogenetically distant lineages often possessed combinations of metabolic capacities that were more similar to each other than to those of more closely related clades (Figure 1c).

Thus, we hypothesize that diverse lineages within the CPR may have converged upon similar metabolic platforms, potentially via combinations of lateral gene transfer and gene loss of genes involved in the same function(s). This finding is intriguing, given that overall protein presence/absence patterns in both CPR and other bacteria generally recapitulate phylogenetic relationships when entire proteomes are considered (Méheust et al., 2019). To account for this difference, we infer that other protein families not included in the current study must show patterns of presence/absence that are generally congruent with the CPR species tree.

Exploration of patterns of gene distribution revealed that patchiness and phylogenetic depth varied for the selected metabolisms and even for enzymes in the same pathway (Figure 2). This finding was validated by additional analyses of how trait patchiness varied with increasing completeness of the underlying genomes, and, for glycolysis in particular, by a comparison to other major bacterial groups. Based on the combination of these analyses, we conclude that incompleteness of genomes from metagenomes used in this study only minimally alters the relative relationships between traits when examining depth and patchiness, and that the unusual patterns observed for CPR organisms are indeed atypical. Similarly, while mis-binning can also complicate any analysis that relies upon metagenome-derived genomes, the similarity of findings for multiple closely-related genomes indicates that it likely does not greatly obscure the major patterns presented here. While the increased availability of complete genomes will best help to further clarify the patterns explored in this study, our general approach to testing the robustness of signal as a function of genome completeness might serve as a valuable way to augment future analyses of gene content in other lineages as well.

We then used gene-species tree reconciliation to validate the prediction that proteins with variable ‘evolutionary profiles’ might have been shaped by different combinations of lateral transfer and vertical inheritance. For a subset of core carbon metabolism, here represented by glycolysis, gene trees were roughly congruent with the reconstructed organismal phylogeny, suggesting that vertical inheritance has primarily shaped distributions of these enzymes (Figure 3). However, the discovery of a divergent subclade of TIM from CPR bacteria that is more closely related to archaeal versions than bacterial ones provides clear evidence of lateral transfer even for the most widely distributed glycolytic enzymes. Interestingly, two enzymes involved in the early steps of the glycolytic pathway (hexokinase/glucokinase and phosphofructokinase) were notably absent in nearly all lineages. Where present, they were likely acquired by lateral gene transfer, potentially following ancestral loss. These sequences separate from those of other bacteria, obscuring the source and suggesting that transfers of phosphofructokinase and hexokinase to CPR were also ancient. In contrast, enolase and pyruvate kinase, the last two steps of the pathway, are only somewhat widespread and show relatively low phylogenetic congruence. This pattern may reflect a mixture of genomic loss in addition to lateral transfer among unrelated CPR organisms.

In archaea, glycolysis is known to be modified in a number of ways, including metabolic shunting (Imanaka et al., 2006) and rewiring of steps through novel enzymes (Siebers and Schönheit, 2005; Verhees et al., 2004). These observations have led to suggestions that evolutionary 'tinkering' has shaped glycolysis at least in some archaeal lineages (Van Der Oost and Siebers, 2007). Paralleling this, we found that several glycolytic steps in CPR bacteria are apparently carried out by different enzyme forms, and, in some cases, by types that are traditionally associated with archaea. This was particularly striking in the case of PGI, which converts Glucose 6-P to Fructose 6-P, where three different enzyme forms accounted for the wide distribution of the function (Figure 3a). Acquisition of variant enzymes may have preceded loss of the ancestral enzyme or occurred afterwards, complementing a loss in function. Overall, our findings suggest that glycolysis among CPR organisms is partly an evolutionary mosaic, as described in at least one eukaryotic organism (the flagellate *Trimastix pyriformis*) (Stechmann et al., 2006), and, further, that gene loss and acquisition may have remodeled their glycolytic pathways over time.

Given the patchy distribution of enzymes involved in upper glycolysis, carbon flux through this portion of the pathway remains unclear. CPR bacteria without glucokinase/hexokinase (hex) or PGI might rely on the uptake of glycolytic intermediates, like fructose 6P or fructose 1,6-P2 from associated cells or released by cell lysis. These compounds could be shunted through the pentose phosphate pathway to bypass the largely absent phosphofructokinase and into the conserved central module of glycolysis (Figure 3a) (Castelle et al., 2018). Alternatively, near universal conservation of TIM and GAPDH across the CPR suggests that either glycerone or G3P could also be important points of input for carbon flow in these organisms. Consistent with this is the fact that CPR organisms encoding Form-III-related RuBisCO are predicted to introduce G3P to central/lower glycolysis as a product of their predicted nucleotide salvage pathway (Sato et al., 2007; Wrighton et al., 2016). The subset of CPR organisms that encode both hexokinase and PGI, on the other hand, could potentially perform a more diverse set of transformations, utilizing glucose precursors taken up from the environment or host. As for lower glycolysis, the observed patchiness in distributions of PGM, enolase, and pyruvate kinase suggests alternative fates for intermediates produced after the step catalyzed by PGK (Figure 3a). In the absence of pyruvate kinase, which was found only in about a third of genomes here, CPR could use phosphoenolpyruvate (PEP) synthetase (PEPS) to instead interconvert PEP and pyruvate or instead generate oxaloacetate (Sauer and Eikmanns, 2005). Of course, with the data presented here, we cannot rule out the possibility that novel, divergent enzymes undetected by our HMM approach functionally substitute for those with patchy or nearly absent distributions among CPR lineages. However, we found no evidence for the presence of archaeal PFK/glucokinase nor strong support for functioning of CPR ROK family proteins as putative glucokinases. Additionally, CPR bacteria are not currently known to employ alternative pathways like the Entner-Doudoroff pathway, as some other bacteria that lack PFK (Conway, 1992). Future work subjecting CPR organisms in culture/co-culture to carbon flux analysis

should help to validate genomic predictions and shed light on the metabolic configurations utilized *in vivo*.

Our second case study investigated the evolutionary history of specialized metabolism in CPR bacteria, focusing on Group 4 and 3b NiFe hydrogenases (Figure 4). These genes, like those putatively involved in nitrite reduction, electron transport, and AMP metabolism (Castelle et al., 2017; Danczak et al., 2017; Jaffe et al., 2019), are sparsely distributed across the CPR and were likely subjected to lateral gene transfer. Notably, we report phylogenetic and genomic evidence for distinct monophyletic clades of Group 3b hydrogenases that are specific to the CPR. This suggests that transfer events were ancient or that these hydrogenase sequences evolved very rapidly. The variable genomic contexts of the 3b-related *hyd1* and *hyd2* suggest at least two evolutionary scenarios: that individual, ancient transfers from non-CPR microorganisms occurred with the associated proteins intact, or that CPR bacteria encoding *hyd2* acquired only the large and small subunit and currently support function with unknown genes. The scattered distribution of both forms, phylogenetically incongruent with the CPR species tree, further suggests that intra-CPR exchange and/or loss also occurred over time. Similarly, we hypothesize that other sparsely distributed protein families among the CPR, like pyruvate:ferredoxin oxidoreductase, cytochrome oxidase, and nirK (nitrite metabolism), may also be the result of lateral transfer followed by further evolution within the CPR. The acquisition of cytochrome oxidase by some Saccharibacteria is presumably an adaptation to aerobic or microaerophilic environments (Castelle et al., 2018; Kantor et al., 2013; Starr et al., 2018).

In contemplating modes of evolution of CPR bacteria, it is important to consider the processes of gene gain and loss in the context of the largely symbiotic lifestyles of these organisms. The dynamic evolution of glycolysis might reflect reduced selection for complete pathways due to metabolic opportunities provided by the host, constraints which probably changed over time. Further, acquisition of new capacities via lateral transfer could have opened new niches, potentially including a change in or adaptation to new hosts in different environments. However, the observation that sequences from CPR bacteria coding for rarer functions are often distinct from those of other bacteria suggests that these transfers probably occurred relatively early in the history of the radiation, or evolved rapidly once acquired. Distantly-related lineages within CPR may have independently undergone loss or gain of the same set of protein families, leading to similarly reduced metabolic platforms over time. These evolutionary constraints may be unique compared to those shaping minimal metabolism in other non-CPR bacterial groups with reduced genomes, like endosymbionts of insects. In contrast to these relatively recently evolved (linked to the appearance of eukaryotic hosts) associations that probably involve irreversible genome reduction trajectories (Moran and Wernegreen, 2000), the potential for CPR organisms to associate with other bacteria raises the possibility of long-established symbioses in which gene sets remain in flux. The resulting pattern of ‘diversity within sparsity’ appears to be characteristic of the CPR.

1.5 Figures

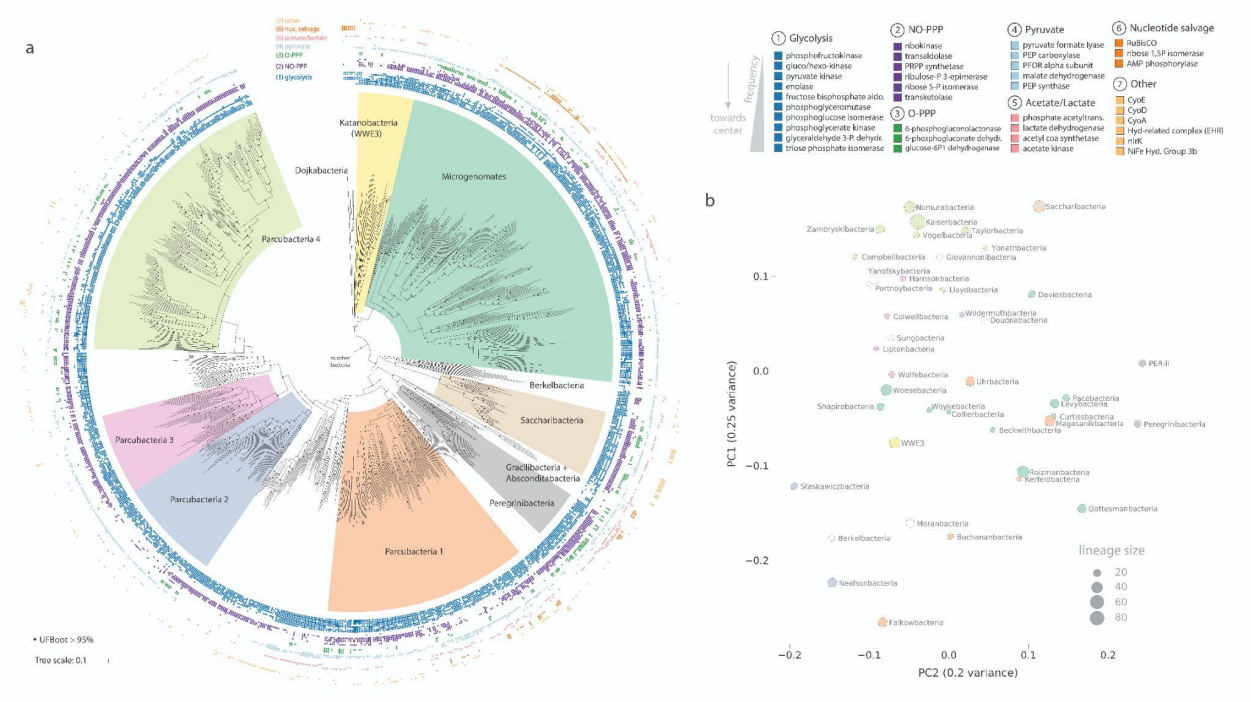

Figure 1. Phylogenetic relationships and metabolic similarity among Candidate Phyla Radiation bacteria. a) Maximum-likelihood tree based on the concatenated set of 16 ribosomal proteins (1427 amino acids, LG+R10 model)**.** Scale bar represents the average number of substitutions per site. Monophyletic subgroups within the Parcubacteria also supported in the concatenated RNA polymerase tree are indicated as Parcubacteria 1-4. Presence/absence of a subset of targeted metabolic traits are indicated as concentric rings. Abbreviations: aldo., aldolase; dehydr., dehydrogenase, PRPP, phosphoribosylpyrophosphate, PEP, phosphoenolpyruvate; PFOR, pyruvate:ferredoxin oxidoreductase; acetyltrans., acetyltransferase; Hyd, hydrogenase. Fully annotated trees with all included lineages are available in Additional file 2, Figure S2. b) Principal coordinates analysis describing similarity between metabolic platforms of CPR lineages with 8 or more representative genomes.

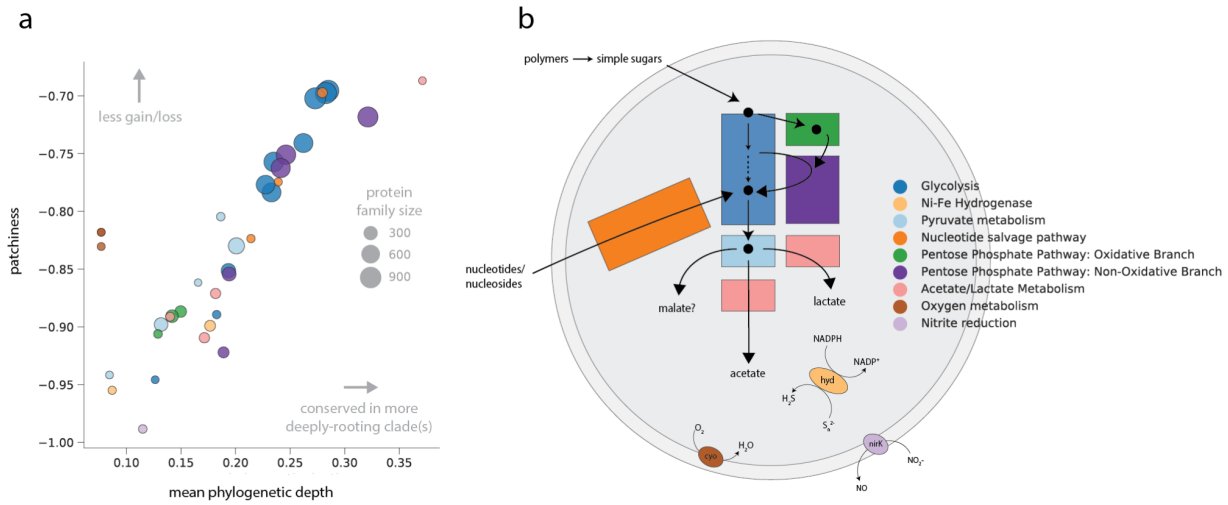

Figure 2. Metabolic traits encoded by CPR bacteria exhibit varying evolutionary profiles, including those in the same pathway (e.g. glycolysis genes). a) Evolutionary profiles generated from phylogenetic depth and patchiness of gene distributions over the rp16 topology. Each point represents a metabolic gene shaded to match the functional category/pathway in b), schematic representing a generalized metabolic platform for CPR bacteria.

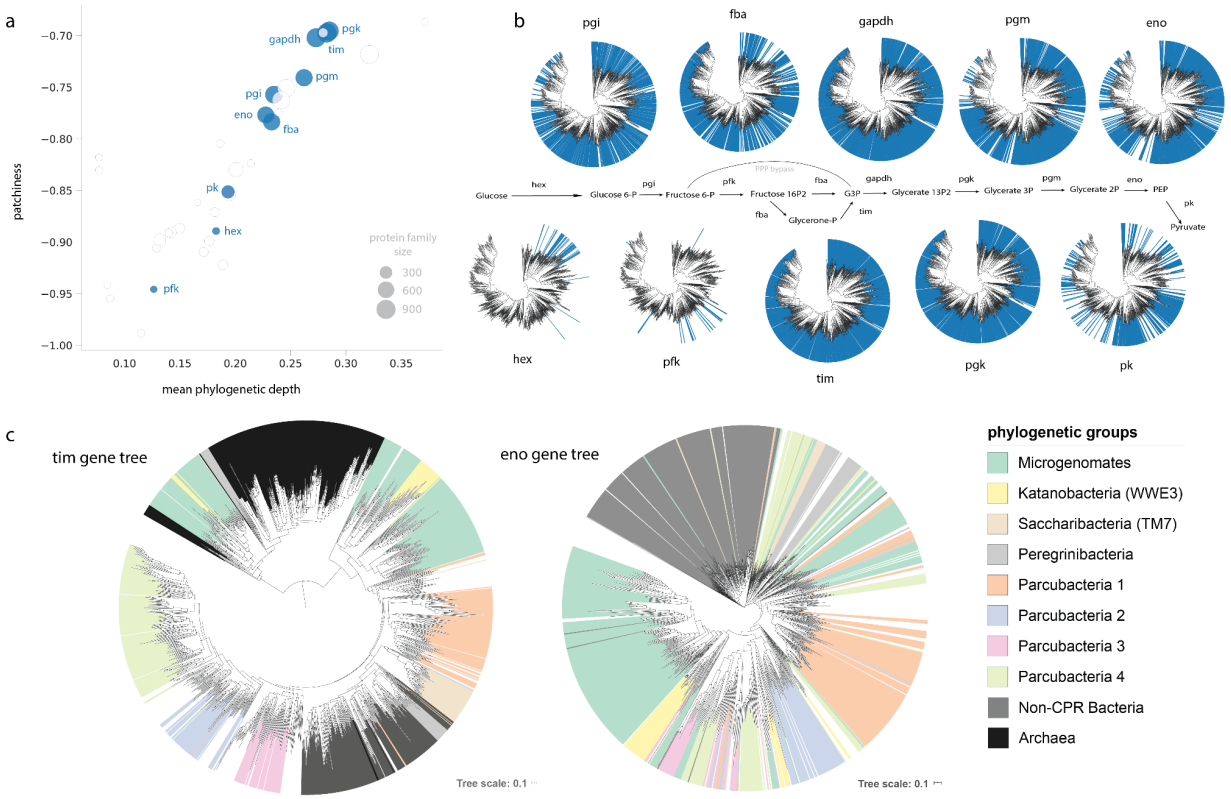

Figure 3. Patterns of distribution and gene trees for glycolytic enzymes across the CPR. a) Evolutionary profiles based on patchiness and phylogenetic depth and b) presence/absence profiles over the rp16 tree. c) Protein-specific molecular phylogenies for triose phosphate isomerase (tim) and enolase (eno). Abbreviations: hex, hexokinase; pfk, phosphofructokinase; pk, pyruvate kinase; fba, fructose bisphosphate aldolase; eno, enolase; pgi, phosphoglucose isomerase; pgm, phosphoglycerate mutase; tim, triose phosphate isomerase; gapdh, glyceraldehyde 3-phosphate dehydrogenase; pgk, phosphoglycerate kinase; G3P, glyceraldehyde 3-phosphate; PEP, phosphoenolpyruvate; PPP, pentose phosphate pathway. Scale bars represent the average number of substitutions per site.

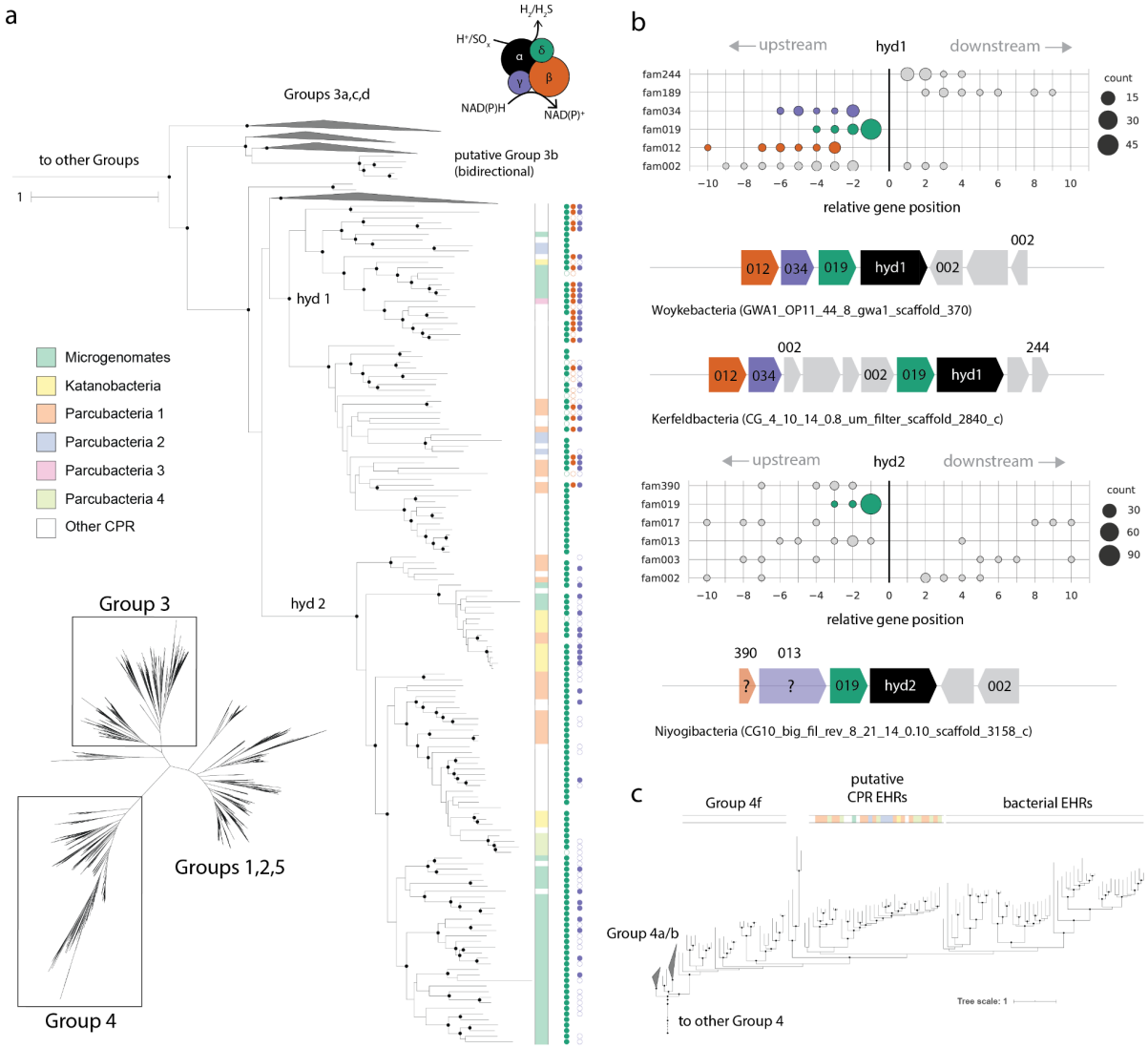

Figure 4: NiFe hydrogenase enzymes encoded by CPR bacteria. a) inset of the unrooted large subunit hydrogenase tree showing putative Group 3b hydrogenases across the CPR, along with presence/absence of HMM hits corresponding to other subunits b) genomic context for hydrogenase gene clusters, where position 0 corresponds to the location of the ORF encoding the large subunit of the NiFe hydrogenase. Only protein families on the same strand as the large subunit are represented in the plots, whereas genome diagrams below the charts include all proximal families regardless of strand orientation. c) Inset of large subunit tree within Group 4 hydrogenases. EHR: Energy-converting hydrogenase related complexes. Scale bars represent the average number of substitutions per site.

2. Lateral gene transfer shapes the distribution of RuBisCO among Candidate Phyla Radiation bacteria and DPANN archaea

Alexander L. Jaffe, Cindy J. Castelle, Christopher L. Dupont, and Jillian F. Banfield

Published in *Molecular Biology and Evolution*, 2019.

Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is considered to be the most abundant enzyme on Earth. Despite this, its full diversity and distribution across the domains of life remain to be determined. Here, we leverage a large set of bacterial, archaeal, and viral genomes recovered from the environment to expand our understanding of existing RuBisCO diversity and the evolutionary processes responsible for its distribution. Specifically, we report a new type of RuBisCO present in Candidate Phyla Radiation (CPR) bacteria that is related to the archaeal Form III enzyme and contains the amino acid residues necessary for carboxylase activity. Genome-level metabolic analyses supported the inference that these RuBisCO function in a CO2-incorporating pathway that consumes nucleotides. Importantly, some Gottesmanbacteria (CPR) also encode a phosphoribulokinase that may augment carbon metabolism through a partial Calvin-Benson-Bassham Cycle. Based on the scattered distribution of RuBisCO and its discordant evolutionary history, we conclude that this enzyme has been extensively laterally transferred across the CPR bacteria and DPANN archaea. We also report RuBisCO-like proteins in phage genomes from diverse environments. These sequences cluster with proteins in the Beckwithbacteria (CPR), implicating phage as a possible mechanism of RuBisCO transfer. Finally, we synthesize our metabolic and evolutionary analyses to suggest that lateral gene transfer of RuBisCO may have facilitated major shifts in carbon metabolism in several important bacterial and archaeal lineages.

***N.B**. All main figures for this manuscript can be found below in section 2.5. All supplementary files (including figures and tables) can be found online with the published manuscript.*

2.1 Introduction

Forms I and II Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) are central to carbon fixation via the Calvin-Benson-Bassham (CBB) Cycle in algae, plants, and some bacteria. Forms III and II/III RuBisCO, discovered in Archaea, are believed to add CO2 to ribulose 1,5-bisphosphate (RuBP) in a two-step reaction from nucleotides like adenosine monophosphate (AMP) (Aono et al., 2015; Sato et al., 2007). These "bona-fide" RuBisCO enzymes were historically considered to be domain specific. In contrast, RuBisCO-like proteins (Form IV) in both Bacteria and Archaea perform functions distinct from carbon fixation and may be involved in methionine salvage, sulfur metabolism, and D-apiose catabolism (Carter et al., 2018; Tabita et al., 2008). While the origin of the RuBisCO superfamily is still unclear (Ashida et al., 2005; Erb and Zarzycki, 2018; Tabita et al., 2007), phylogenetic analysis and enzymatic characterization have suggested that the modern distribution of "bona fide" RuBisCO can be explained by vertical and lateral transfer from an archaeal, Form III ancestor (Tabita et al., 2007).

Recently, metagenomic studies of diverse environments have introduced additional complexity to evolutionary considerations by uncovering new RuBisCO diversity. First, new genomes from Candidate Phyla Radiation (CPR) bacteria and DPANN archaea (a group originally defined based on the lineages Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota, but now including others) contain a hybrid II/III RuBisCO similar to that found in the archaeon *Methanococcoides burtonii* (Castelle et al., 2015; Wrighton et al., 2012). One version of this enzyme was shown to complement photoautotrophic growth in a bacterial RuBisCO deletion strain (Wrighton et al., 2012, 2016). An additional RuBisCO form related to the archaeal Form III, called the "III-like," was reported from genomes of other CPR and some DPANN archaea, further expanding enzyme diversity in these groups (Castelle et al., 2015; Castelle and Banfield, 2018; Wrighton et al., 2016). The III-like enzyme is also predicted to function in the CO2-incorporating AMP pathway (Wrighton et al., 2016).

The presence of RuBisCO in some CPR bacteria and DPANN archaea was interesting because these organisms have small genomes and lack many core metabolic functions. The common absence of pathways for synthesis of nucleotides, amino acids, and lipids has led to speculation that they live symbiotic or syntrophic lifestyles (Brown et al., 2015; Wrighton et al., 2012). Despite this, recent research has shown that some members have a wide range of fermentative capabilities and the potential to play roles in carbon, nitrogen, sulfur, and hydrogen cycling (Danczak et al., 2017; Wrighton et al., 2012). A similar putative ecology has been suggested for the DPANN archaea (Castelle et al., 2015). The recovery of RuBisCO, functioning in an adenosine monophosphate (AMP) pathway, expanded possible metabolic modes for some of these organisms and suggested that CPR and DPANN encoding this enzyme could derive energy

and/or resources from ribose produced by other community members (Castelle and Banfield, 2018; Wrighton et al., 2016).

The discovery of the reductive hexulose pathway (RHP) provided new insight into the variety of ways that RuBisCO can be configured for sugar metabolism. The RHP pathway, differing only in a few steps from the CBB cycle, also employs RuBisCO along with phosphoribulokinase (PRK) to regenerate RuBP and fix carbon dioxide in some methanogenic archaea (Kono et al., 2017). In light of these discoveries, the full diversity, distribution, and possible functionality of RuBisCO in divergent groups like the CPR/DPANN remains an open and important question. In the last few years, additional work has recovered CPR and DPANN genomes from a much wider array of environmental types, including additional groundwater locations, the deep subsurface, hydrocarbon-impacted environments, and the ocean (Hernsdorf et al., 2017; Parks et al., 2017; Tully et al., 2018). Here, we examined over 300 genomes from metagenomes from these environments to further elucidate diversity, potential functions, and the evolutionary history of RuBisCO in members of the bacterial Candidate Phyla Radiation and DPANN archaea. First, we expand the distribution of forms to new phylogenetic groups and propose a new type that, in some cases, might act in concert with phosphoribulokinase to augment carbon metabolism. Additionally, we describe a clade of putative RuBisCO-like proteins encoded by bacteriophage from diverse environments. Drawing on these observations and previous analyses, we suggest that lateral gene transfer may have been largely responsible for the distribution of this enzyme among the CPR/DPANN. These lateral transfers could, in the presence of genes from other pathways, introduce new RuBisCO-based metabolic capacity to genomically reduced lineages.

2.2 Materials and Methods

Genome collection and annotation

We gathered a set of ~4,000 CPR/DPANN genomes from metagenomes from several previous studies of groundwater, soil, ocean, and subsurface environments. Additionally, we binned several new genomes from a sediment from Rifle, CO (Anantharaman et al., 2016) and the water column in the Baltic Sea (Asplund-Samuelsson et al., 2016). Binning methods and taxonomic assignments followed those described in Anantharaman et al. (Anantharaman et al., 2016). Several genome fragments were manually curated, making use of unplaced paired reads to increase their length. This was necessary to test for bacterial affiliation based on comparison of the encoded genes with genes in known CPR genomes.

Proteins were predicted for each genome using Prodigal (“meta” mode) (Hyatt et al., 2010). Preliminary functional predictions were established using a pipeline based on KEGG Orthology

(Kanehisa and Goto, 2000). All vs. all global search of proteins in each KO from the KEGG database was performed using usearch (Edgar, 2010), and protein percent identity was used as input to MCL clustering (Van Dongen, 2008) with inflation parameter of 1.1. For each resulting cluster, the proteins were aligned using MAFFT version 7 (Katoh and Standley, 2013), and Hidden Markov Models (HMMs) were constructed using the HMMER suite (Finn et al., 2011). Predicted proteins from each CPR/DPANN genome in this study were scanned using hmmsearch (Finn et al., 2011), and annotation was assigned according to the best HMM hit, providing it was above a pre-defined KEGG Orthology noise cutoff.

RuBisCO analysis

We extracted above-threshold hits for RuBisCO large chain (K01601), yielding a final set of genomes encoding the enzyme. To analyze the number of non-redundant genomes containing RuBisCO, we repeated the above analysis with a set of ~3000 high quality genomes from various environments. These genomes were de-replicated at 99% secondary ANI using dRep (-comp 20) (Olm et al., 2017) and then analyzed for presence of RuBisCO.

To expand the breadth of our main RuBisCO set, we identified RuBisCO sequences (many of which were unbinned) from sediment and groundwater metagenomes (Anantharaman et al., 2016; Hernsdorf et al., 2017; Probst et al., 2017). We excluded sequences shorter than 200 amino acids in length to remove fragmented proteins. Phylum-level taxonomy for these sequences was assigned based on the closest affiliation of the encoded sequences. These sequences were added to those from genomes and the entire set was de-replicated (USEARCH, -id 0.99 -sort length) (Edgar, 2010). Sourcing for de-replicated sequences can be found in **Table S1**. Next, we combined the full set with reference RuBisCO from NCBI and aligned it using MAFFT (default parameters) (Katoh and Standley, 2013). Alignments were trimmed by removing columns with >95% gaps. The unmasked alignment file of de-replicated RuBisCO sequences with metadata is attached as Supplementary File 1. We next constructed a maximum likelihood tree with RAxML-HPC BlackBox (v. 8.2.10) as implemented on cipres.org (default parameters with rapid bootstrapping) (Stamatakis et al., 2008) and subsequently assigned each RuBisCO sequence to previously identified Forms based on phylogenetic clustering with reference sequences. Binned sequences excluded from the de-replicated set were re-inserted into the tree and classified for downstream analyses. Sequences in ambiguous phylogenetic positions were annotated as "unknown." Custom HMMs were constructed using recovered sequences for each RuBisCO form with the HMMER suite (Finn et al., 2011) and were subsequently self-tested and manually refined to exclude low-scoring sequences.

Collection and analysis of viral sequences

To explore the possibility that phage encode RuBisCO, we generated a database of putative phage genome fragments using sequences from IMG/VR (img.jgi.doe.gov/vr/) and several previous metagenomic studies of groundwater and subsurface environments (Anantharaman et al., 2016; Probst et al., 2017). Contigs from the latter metagenomes were assigned a putative phage origin if the majority of encoded genes had no identifiable sequence similarity to genes in bacterial (or archaea). Only sequences >10 kbp in length were included to improve the confidence of phage assignments. Predicted proteins from putative phage contigs were interrogated using the above RuBisCO HMMs, and those with significant HMM hits at or above a score of 100 and e<<0.05 were retained for further analysis. Genome fragments with RuBisCO-related sequences were further evaluated to confirm the presence of additional (e.g., structural) genes indicative of phage classification. Once manually verified, the putative RuBisCO proteins of phage origin were de-replicated at 99% identity and incorporated into the phylogenetic analysis. Phage terminase proteins were extracted using existing annotations (in the case of IMG/VR fragments) or BLAST-based annotations (in the case of groundwater/subsurface fragments). To establish putative identity, we then aligned recovered phage terminases with reference proteins and created a tree with RaxML. Finally, several additional phage genome fragments encoding RuBisCO-related sequences were manually curated to increase their length.

Residue analysis and protein modeling

To analyze the biochemically-relevant characteristics of the RuBisCO sequences included in the de-replicated set, including those in the putative phage category, we extracted 12 residues known to be important for catalytic activity and 7 important for substrate binding from each sequence (Saito et al., 2009; Tabita et al., 2008). A sequence logo of these 19 sites for each Form was constructed using WebLogo (Crooks et al., 2004). Additionally, we modeled exemplary Form III, Form III-c, Form IV-like, and PRK proteins using the I-TASSER suite (zhanglab.ccmb.med.umich.edu/I-TASSER/) (Yang et al., 2015).

Pathway and contig-level analyses

Finally, we identified two sets of proteins involved in RuBisCO-mediated carbon metabolism (**Table S2**) within the binned genomes using the KEGG annotation results. Of particular interest were AMP phosphorylase (*deoA*) and R15P isomerase (*e2b2*), thought to be involved in AMP metabolism (Wrighton et al., 2016), and phosphoribulokinase (PRK), a marker gene for the CBB pathway. We examined the distribution of these genes, plus others associated with the CBB pathway, across genomes as well as their genomic context. Specifically, for the three genes likely involved in AMP metabolism, the proximity of the genes on the same contig was calculated by taking the sum of the lengths of the intervals between the genes. Gene fusions were identified by

noting abnormally long RuBisCO sequences and examination of the domain structure through NCBI blastp (blast.ncbi.nlm.nih.gov/Blast.cgi). As with RuBisCO, recovered protein sequences for PRK, AMP phosphorylase (*deoA)*, R15P isomerase (*e2b2),* and CBB enzymes were aligned with MAFFT and trimmed as described above. Corresponding trees were then inferred with RAxML-HPC BlackBox (v. 8.2.10) as implemented on cipres.org (default parameters with rapid bootstrapping) (Stamatakis et al., 2008) and visualized using iTOL (Letunic and Bork, 2016). For PRK, NCBI reference sequences, close BLAST hits to identified CPR PRK homologs, and sequences from the close homolog uridine kinase (udk) were gathered and added to the protein set before alignment.

2.3 Results

Metagenomics expands the diversity of RuBisCO forms and reveals putative new enzyme types in CPR and phage

The large majority of CPR and DPANN RuBisCO sequences analyzed fell into clearly defined phylogenetic groups **(Fig. 1a)**, four of which (II/III, III, III-like, IV) correspond to the "Forms" defined in previous literature. A small number of sequences from both the CPR and DPANN (including the generically labelled branches in **Fig. 1a**) resolved in ambiguous phylogenetic positions, largely due to low support for internal nodes. Among the clearly defined groups, we observed the following results:

II/III. Our analysis recovered new sequences of the Form II/III RuBisCO, originally thought to be exclusively found in Archaea (Alonso et al., 2009). Here, we broaden the phylum-level distribution of the group with additional representatives from the DPANN groups Woesearchaeaota and Micrarchaeota **(Fig. 2)**. Form II/III RuBisCO of CPR and DPANN partition into two subgroups based on the presence or absence of a 29 amino acid (or longer) insertion, the biochemical implications of which are currently unknown (Wrighton et al., 2016). All but one of the full-length Micrarchaeaota and Woesearchaeota sequences recovered in this study contained insertions in the expected region.

III. We identified archaeal Form III RuBisCO sequences in a variety of DPANN, and, potentially, CPR genomes. We refer to this as Form III-b to distinguish it from Form III-a, a divergent group described by Kono et al. (Kono et al., 2017), and the Form III-like proteins (Wrighton et al., 2016). Most of the newly reported Form III-b sequences contained the critical substrate binding and catalytic residues for RuBisCO function but also largely shared residue identity with canonical archaeal versions **(Fig. 1c)**. Notably, we found this enzyme form in three DPANN groups - the Diapherotrites, the Micrarchaeota, and the Woesearchaeota - previously not

known to harbor Form III-b RuBisCO, extending the presence of these enzymes beyond Pacearchaeota and Aenigmarchaeota (Castelle et al., 2015), two major groups in the DPANN **(Fig. 2)**. Approximately 70 DPANN sequences fell outside the clade containing characterized Form III-b, but were assigned to this type given low support for separating branches and overall closer relatedness to III-b than III-like sequences (DPANN (Form III-b), **Fig. 1a**).

We identified several previously published sequences assigned to genomes from the Levybacteria and Amesbacteria phyla (CPR superphylum Microgenomates) deeply nested within the archaeal Form III-b sequences. In addition, we recovered a set of unbinned sequences with highest similarity to these same binned Levybacteria and Amesbacteria genomes. The Levybacteria sequences group loosely with sequences from Aenigmarchaeota and a reference sequence from the archaeon *Methanoperedens nitroreducens* (~80% identity). In contrast, the Amesbacteria sequences grouped most closely to those from a clade from Pacearchaeota. To verify the binning of these genome fragments, we examined the top BLAST hits to well annotated genes on each genome fragment bearing a RuBisCO gene. For the fragment putatively assigned to Amesbacteria, BLAST affiliation of non-hypothetical genes was inconclusive - some genes had only low identity with archaeal genes, others had identity to both bacterial and archaeal genes, while one gene encoding a serine acetyltransferase had ~70% bacterial identity. In addition to RuBisCO, the ~34 kbp putative Levybacteria genome fragment encoded a large CRISPR-Cas locus, a novel transposase, and a phage/plasmid primase, all of which may indicate mobile genetic material. BLAST affiliation of well-annotated genes was also inconclusive for this fragment - if not of bacterial origin, the genome fragment could be phage or plasmid. We additionally recovered one small genomic fragment, possibly related to the Moisslbacteria, that also included the III-b enzyme. Ultimately, further research is needed to confirm that Form III-b RuBisCO is encoded in genomes of some CPR bacteria.

III-like. Many RuBisCO sequences fell into a deep-branching, monophyletic clade that is divergent from the archaeal Form III and referred to as "III-like" (Wrighton et al., 2016) **(Fig. 1a)**. Here, we added five Dojkabacteria (WS6, a group within the CPR) sequences from hydrocarbon-impacted environments and nearly 40 DPANN sequences from Woesearchaeota and Pacearchaeota from multiple environments **(Fig. 2)**. As reported previously (Wrighton et al., 2016), the "III-like" sequences recovered in this analysis appear to have insertions mostly 11 amino acids in length.

III-c. We report here an additional deep-branching clade of RuBisCO sequences with overall sequence similarity closest to Form III-b RuBisCO (~50%) (**Fig. 1a).** While support was low for internal branches separating this clade from III-b sequences, best maximum-likelihood trees suggested a distinct phylogenetic status. This group, which we term "Form III-c", contained nearly 50 sequences from the Dojkabacteria and several groups in the Microgenomates and Parcubacteria superphyla. Form III-c was particularly abundant in the Gottesmanbacteria and

Kuenenbacteria, the latter of which appears to harbor this form exclusively **(Fig. 2)**. We identified organisms bearing this enzyme type in almost every environment included in our study, indicating that it may compose an important and previously unrecognized aspect of RuBisCO diversity. To predict the biochemical potential of this type, we examined 12 residues known to be important for carboxylase activity and 7 important for substrate binding from each sequence (Saito et al., 2009; Tabita et al., 2008). This analysis indicated that Form III-c RuBisCO sequences contain critical residues largely identical to those reported for Form III-b enzymes; however, a minority of sequences encoded modifications to the conserved aspartic acid in catalytic site #6 (CA6) that could alter its chemical properties **(Fig. 1c)**. Additionally, modeling through I-TASSER revealed that secondary structure of an III-c sequence was consistent with that of existing Form III-b templates **(Fig. 1b)**.

IV and IV-like. Form IV (RLP) proteins are a clade of highly divergent RuBisCO with low sequence similarity (~30%) to 'bona fide' RuBisCO enzymes and divergent functions (Hanson and Tabita, 2001; Tabita et al., 2008). Our phylogenetic analyses identified 8 Form IV (RLP) RuBisCOs in CPR genomes, all in the genomes of bacteria from the Parcubacteria superphylum **(Fig. 2)**. The Parcubacteria Form IV sequences composed a monophyletic clade nesting within the previously described IV-Photo type RLP. However, the key catalytic/substrate binding residues were only partially conserved **(Fig. 1c)**. We recovered only one RLP from a Micrarchaeota (DPANN) genome **(Fig. 2)**, despite the fact that archaeal sequences appear to be at the root the RuBisCO/RLP superfamily (Tabita et al., 2007).

Using manually curated Hidden Markov Models (HMMs) constructed from recovered CPR and DPANN RuBisCO sequences, we also identified approximately 80 non-redundant sequences related to Form IV RuBisCO on putative phage as well one unusually large, curated phage genome (~200 kbp) **(Fig. S1a)**. The phage sequences appeared highly divergent from other RLPs, but most shared some key residues found in canonical Form IV RuBisCOs **(Fig. 1c)** and scored highly on HMMs constructed from verified Form IV RuBisCO (score > 200, e-val << 0.05). Additionally, modeling of one sequence revealed a secondary structure consistent with large chain templates but missing several conserved residues, as expected for Form IV enzymes **(Fig. 1b)**. Analysis of co-encoded phage terminase proteins indicated that at least some of the viral sequences were from members of the Myoviridae. Two other RLP-related sequences attributed to Beckwithbacteria grouped with those from putative/confirmed phage and together appeared as an outgroup to all other RLPs **(Fig. 1a)**. Analysis of the genes encoded on the manually curated Beckwithbacteria fragment with IV-like RuBisCO supported its bacterial origin **(Fig. S1b)**.

Form III-c RuBisCO is encoded in close proximity to genes of the CO_2-incorporating AMP pathway

To test the hypothesis that the Form III-c RuBisCO participates in the previously described AMP pathway, we narrowed our focus on the binned genomes containing this form of the enzyme. The AMP pathway employs AMP phosphorylase (*deoA*) and ribose-1,5 -bisphosphate (R15P) isomerase (*e2b2*) to provide RuBisCO with RuBP substrate, incorporating CO_2 to produce two 3-Phosphoglyceric acid molecules (3-PGA) which are fed into glycolysis (Aono et al., 2015; Sato et al., 2007). Thirty-nine genomes encoding Form III-c RuBisCO (67%) also contained homologs for *e2b2* and *deoA*. Most Gottesmanbacteria (OP11) contained homologs to *deoA* at a threshold lower than originally used in the KEGG analysis, as these sequences generally had a conserved deletion of ~80 amino acids at the start of the protein compared to other CPR proteins. Despite this, we predict that these proteins are divergent *deoA* homologs based on phylogenetic placement and re-alignment with reference sequences. Among Gottesmanbacteria genomes, the percentage containing all three genes associated with AMP metabolism was higher (~91%). Thus, the new Form III-c RuBisCO, especially among Gottesmanbacteria, is consistently associated with *deoA* and *e2b2* homologs, as reported previously for CPR genomes with other forms of the enzyme (Wrighton et al., 2016) **(Fig. S2a)**.

To analyze possible function of the newly recovered RuBisCOs, we examined the genomic proximity of *deoA* and *e2b2* homologs with RuBisCO in all genomes containing Form III-b, III-c, III-like, and II/III enzymes. Forty-four genomes contained fragments encoding all three genes on the same stretch of assembled DNA **(Fig. S2b)**. Most fragments encoded RuBisCO between the other genes, although a minority of genomes appeared to contain rearrangements **(Fig. S2c)**, as previously reported for a Form II/III-bearing PER-1 genome (Wrighton et al., 2016). Next, for cases where all three genes occurred on the same contig, we defined a metric called "pathway proximity" (sum of the genomic distance between the three genes, see Methods). This analysis suggested that genes involved in AMP metabolism are more frequently and more proximally co-located in the genomes of organisms bearing the Form III-c RuBisCO than in genomes with other forms, even though these RuBisCO were present on assembled fragments of similar length **(Fig. 3a)**. Two outlier Form III-c genomes from a previous study (Parks et al., 2017) encoded extremely distant *deoA* genes on long fragments, possibly the result of genetic rearrangement or errors in genome assembly. Interestingly, some genomes bearing Form III-c RuBisCO encoded the three genes consecutively, where in other cases the RuBisCO gene was fused to the isomerase **(Fig. 3b, Fig. S3)**. The close proximity and fusion support the conclusion that this RuBisCO form participates in the CO_2-incorporating AMP pathway.

RuBisCO phylogeny is incongruent with that of e2b2 *and* deoA

The discovery of Form III-like, and now III-c, RuBisCO in the CPR bacteria suggested that lateral gene transfer may have played a role in shaping the distribution of Form III-related enzymes across the tree of life (Erb and Zarzycki, 2018; Wrighton et al., 2016). To evaluate the extent of lateral gene transfer involving the genes of the AMP pathway, we compared the phylogeny of RuBisCO with that of the other pathway components (Wrighton et al., 2016). With the possible exception of the Peregrinibacteria, we found that the trees for CPR and DPANN AMP phosphorylase (*deoA)* and R15P isomerase (*e2b2)* recapitulate organism phylogeny and so were largely incongruent with that of RuBisCO **(Fig. S4)**. Among the monophyletic Gottesmanbacteria, genomes with the same RuBisCO Form (III-c, III-like) appeared to cluster together, as would be expected if lateral gene transfer of various RuBisCO followed vertical inheritance of *deoA* and *e2b2* by different sub-phylum level lineages. Although still undersampled, there is some indication that RuBisCO-correlated phylogenetic clustering will emerge for Parcubacteria (OD1) and Dojkabacteria (WS6), which at the phylum level contain high RuBisCO diversity.

CPR bacteria bearing Form III-c RuBisCO also encode phosphoribulokinase

We examined whether recovered CPR RuBisCO, including the Form III-c, might participate in a form of the CBB pathway by searching all binned genomes for homologs of phosphoribulokinase (PRK). PRK is a key CBB marker gene that is critical for regenerating RuBP substrate. Our analysis recovered 31 genomes encoding PRK homologs among the Gottesmanbacteria and Peregrinibacteria, the first PRK homologs reported in the CPR. PRK sequences were encoded by Gottesmanbacteria harboring Form III-like and III-c RuBisCO, whereas Peregrinibacteria PRK were associated with the typical Form II/III enzyme. Phylogenetic analysis revealed that Gottesmanbacteria PRK sequences comprised a well-supported monophyletic clade nesting within a group of sequences from recently isolated cyanobacterial genomes and, more broadly, a larger clade of archaeal PRK homologs **(Fig. 4, Fig. S5)**. These putative archaeal and cyanobacterial sequences appeared to be distinct from classical versions and have not yet been assayed for functional activity. Similarly, two recovered Peregrinibacteria PRK formed a monophyletic clade sister to additional divergent cyanobacterial sequences **(Fig. 4, Fig. S5)**. To test whether CPR genomes containing PRK have the full genomic repertoire for the CBB cycle, we searched genomic bins for nine other genes involved in this pathway (**Table S2**). Several Gottesmanbacteria bins contained near-complete CBB pathways, lacking only the gene for sedoheptulose 1,7 bisphosphatase (SBP, 08 in **Fig. 4**). Additionally, instead of separate genes for fructose 1,6-bisphosphate aldolase (FBA) and fructose 1,6-bisphosphatase (FBPase), most Gottesmanbacteria genomes encoded an enzyme most similar to a bifunctional version found in some thermophilic, chemoautotrophic bacteria and archaea (Say and Fuchs, 2010) **(Fig. 4. Fig S7)**. One Peregrinibacteria genome contained all genes

involved in the CBB cycle with the exception of FBPase (06), although additional genome sampling/reconstruction is necessary to definitively designate this enzyme as missing.

2.4 Discussion

New diversity in the RuBisCO superfamily and its implications for phylogenetic distribution and metabolic function among the CPR and DPANN

Through increased metagenomic sampling of diverse environments, we provide new information about the distribution of RuBisCO in major bacterial and archaeal groups and expand RuBisCO superfamily diversity. Our results also allow a quantitative assessment of RuBisCO diversity, revealing that the Pacearchaeota and Dojkabacteria in particular frequently encode various forms of the enzyme across many environmental types **(Fig. 2)**. This suggests that RuBisCO may be an important metabolic enzyme for these groups, which appear to have the most minimal metabolic and biosynthetic capacities among the DPANN and CPR radiations (Castelle and Banfield, 2018).

At present, Form III-c sequences occur only in several CPR lineages, contrasting with other Form III-related enzymes with archaeal representatives. Specifically, Form III-a is only known in methanogens, Form III-b is known to occur in archaea and possibly several CPR lineages, as well as another bacterium (*Ammonifex degensii*) (Berg et al., 2010), and Form III-like enzymes appear to be widely (but sparsely) distributed in both DPANN archaea and CPR bacteria. However, like the other Form III-related enzymes, the association of Form III-c RuBisCO with *e2b2* and *deoA* suggests this enzyme may also function in an AMP metabolism pathway. This pathway, originally described for the Form III-b enzyme in *Thermococcus kodakarensis*, relies on two proteins to provide RuBisCO with its substrate molecule, ribulose-1,5-bisphosphate (RuBP) (Aono et al., 2015; Sato et al., 2007). First, an AMP phosphorylase encoded by the *deoA* gene catalyzes the release of ribose-1,5-bisphosphate (R15P) which is then subsequently converted to RuBP by a R15P isomerase encoded by *e2b2* (Aono et al., 2015; Sato et al., 2007). Next, the RuBisCO incorporates H_2O and CO_2 with this substrate to create two molecules of 3-phosphoglycerate (3-PGA), which in turn can be diverted into central carbon metabolism (Aono et al., 2015; Sato et al., 2007). Among the CPR, this pathway is thought to provide a simple mechanism for ribose salvage that may facilitate the syntrophic ecology of these organisms (Castelle and Banfield, 2018; Wrighton et al., 2016). Contig-level analyses revealed a notable spatial association and occasional fusion of genes involved in AMP metabolism in genomes bearing the Form III-c RuBisCO, supporting the association of the enzyme with this pathway. However, it is critical that future studies characterize the specific biochemistry of this

new form and its possible function in CPR bacteria. While residue analysis of the Form III-c RuBisCO suggests that these enzymes encoded the minimum set of catalytic and substrate binding sites necessary for carboxylase activity, the associated metal cation and the impact of non-active site catalysis remain unknown.

Form I and Form II RuBisCO function in the Calvin-Benson-Bassham cycle, which relies on phosphoribulokinase (PRK) to regenerate RuBP before carbon fixation by RuBisCO **(Fig. 4)**. The presence of PRK in Gottesmanbacteria raises the possibility that a CBB-like pathway may operate in carbon assimilation in these organisms. Lacking from their genomic repertoires, however, is sedoheptulose 1-7 bisphosphatase (SBPase) **(Fig. 4)**. In plants, SBPase catalyzes the dephosphorylation of sedoheptulose 1,7 bisphosphate and is important for regulation of intermediate molecules in the CBB cycle (Harrison et al., 1997). Among Cyanobacteria, it has been shown that a single enzyme often functions as both an SBPase and an FBPase, catalyzing a similar reaction on fructose bisphosphate in the second branch of the cycle **(Fig. 4)** (Feng et al., 2014; Gerbling et al., 1986). Similarly, bifunctional activity has been demonstrated for other bacterial FBPase enzymes in both the CBB (*Ralstonia eutropha*) and a ribulose monophosphate cycle (*Bacillus methanolicus*) (Stolzenberger et al., 2013; Yoo and Bowien, 1995). Complicating the possibility of a bifunctional FBPase/SBPase in the Gottesmanbacteria is the observation that most of these genomes encode a single enzyme most similar to a bifunctional fructose 1,6-bisphosphate aldolase/phosphatase, instead of separate FBPase and FBA. In the archaeal and bacterial lineages in which bifunctional FBA/FBPases have been characterized, these enzymes are thought to play a role in gluconeogenesis instead of the CBB pathway (Say and Fuchs, 2010). Thus, the association of this gene in the classical CBB in Gottesmanbacteria would require tripartite function as a FBPase, FBA, and SBPase. As such, the functioning of a CBB-like pathway in CPR remains uncertain.

An alternative inference is that PRK contributes to carbon metabolism in Gottesmanbacteria by providing additional RuBP as substrate for RuBisCO functioning in the AMP pathway. The same may be true of the two Peregrinibacteria genomes encoding the PRK but missing FBPase **(Fig. 4)**. In this scenario, components of the oxidative pentose phosphate pathway could convert glucose-6P into ribulose-5-P, which could then be converted to RuBP by the PRK **(Fig. S6)**. Going forward, it is critical that the PRK from both lineages, as well as the Gottesmanbacteria FBPase/FBA, be characterized biochemically, especially given that these sequences are divergent from well-studied enzymes **(Fig. S7)**. In any case, the presence of PRK, RuBisCO, and a putative bifunctional FBPase/FBA in CPR genomes suggests that these organisms may have acquired fundamental components of carbon metabolism by lateral transfer. This extends the prior observations of transfer of the bifunctional FBA/FBPase to bacteria (Say and Fuchs, 2010) as well as the general occurrence of transfer among CPR bacteria (Jaffe et al., 2016).

Finally, we report Form IV RuBisCO-like proteins (RLPs) in the genomes of several CPR bacteria and a DPANN archaeon. Previously described RLPs fall into 6 clades and have distinct patterns of active site substitutions that likely affect their functionality (Tabita et al., 2007). Sequences from Parcubacteria and Micrarchaeota were mostly closely related to a known RLP clade called the IV-Photo, which is implicated in sulfur metabolism/stress response in green sulfur bacteria like *C. tepidum* (Hanson and Tabita, 2001). While active site residue divergence complicates the inference of function based on phylogenetic placement, it is possible that Form IV sequences found in CPR bacteria also function in oxidative stress response. Our analysis also revealed the presence of a large, highly divergent clade of RuBisCO sequences related to Form IV/RLP in the genomes of bacteriophage **(Fig. 1a)**, some of which were classified as Myoviridae. That these sequences were recovered from various environments suggests that these phage proteins may be a widespread and to date underappreciated reservoir of diversity in the RuBisCO superfamily. Sequence analysis revealed that these putative RuBisCO-like proteins encoded key residues divergent from known type IVs, leaving possible functionality unclear. However, future work may uncover "bona fide" RuBisCO-like proteins on phage genomes, supporting the inference that these enzymes are widely laterally transferred across lineages. Previous work has shown that marine phage can impact host carbon metabolism through auxilary expression of other photosynthetic genes (Crummett et al., 2016; Thompson et al., 2011). Regardless of type, phage-associated proteins with homology to RuBisCO should be assessed functionally to evaluate whether they have to the potential to augment host metabolism during infection.

Sparse distribution of the RuBisCO superfamily suggests that lateral gene transfer shapes the distribution of multiple forms among CPR, DPANN, and bacteriophage

Regardless of the ancestral function of RuBisCO superfamily (Ashida et al., 2005; Erb and Zarzycki, 2018; Tabita et al., 2007), ancient lateral gene transfer is likely an important process underlying the current distribution of RuBisCO in bacteria and archaea. Transfers probably drove the evolution of Form IV as well as Forms I and II from the ancestral Form III (Tabita et al., 2007, 2008), ultimately resulting in diverse RuBisCO types that now occur in Archaea, Bacteria, and Eukaryotes. Results of the current study support this inference and extend it, suggesting that lateral gene transfer has also played a role in distributing recently recognized RuBisCO forms across Archaea and Bacteria. First, the discovery of Form III-c RuBisCO in CPR bacteria suggests gene transfer between CPR bacteria and archaea, as this new Form is most closely related to the archaeal Form III-b. The recovery of canonical archaeal Form III-b proteins within previously published Amesbacteria and Levybacteria bins (both CPR), if verified, would support this conclusion. Additionally, previous findings of a Dojkabacteria (WS6) genome harboring both Form III-like and Form II/III RuBisCO and a divergent Form III-b RuBisCO enzyme in the Firmicute *Ammonifex degensii* are best explained by gene acquisition via lateral transfer (Berg et al., 2010; Hernsdorf et al., 2017).

Broader evolutionary patterning supports the idea that lateral gene transfer has played an underappreciated role in the shaping the evolution of RuBisCO among CPR and DPANN. Our results reveal a relatively wide but sparse distribution of RuBisCO across CPR/DPANN lineages, with most lineages containing very low frequencies of the enzyme **(Fig. 2)**. The non-congruency of RuBisCO phylogeny with those of *deoA* and *e2b2*, which appear to have been largely vertically transmitted in CPR lineages **(Fig. S4),** suggests divergent evolutionary histories of these functionally related genes. Thus, we conclude that the distribution of RuBisCO diversity in the CPR/DPANN is more likely to be explained by lateral transfer than extensive gene loss.

The results presented in this study give new insights on the evolution of the RuBisCO superfamily as a whole. **Fig. 5** is a schematic diagram that integrates ideas of Tabita and colleagues (Tabita et al., 2007) with inferences arising from our results, detailing one of at least several possible scenarios by which RuBisCO was distributed across the tree of life. Previous work has suggested that an archaeon, possibly an ancestor of the Methanomicrobia, laterally transferred a Form III enzyme to a bacterial ancestor (Step 1 in Fig. 5) where it subsequently evolved to generate both the Form I and Form II enzymes (Step 2) (Schönheit et al., 2016; Tabita et al., 2007). The findings of the current and prior studies (Tabita et al., 2007; Wrighton et al., 2016) indicate that an ancestral Form III sequence may then have diverged from a common ancestor into at least three Form III-related types**,** including the Form III-b (traditional archaeal form). Specifically, an additional transfer of an ancestral Form III or III-b enzyme from Archaea to a CPR bacterium may have led to the evolution of the III-like and newly reported III-c Forms (Steps 3, 4), as previously hypothesized (Erb and Zarzycki, 2018). Subsequent transfers of both forms to the Dojkabacteria (Step 5), Parcubacteria (Step 6), and of Form III-like to the common ancestor of Pacearchaeota and Woesearchaeota (Step 7) would recapitulate the current distribution of of these enzymes across the tree of life. However, with the current evidence, we cannot rule out the possibility that the III-like RuBisCO evolved within DPANN archaea and was transferred in the reverse direction to the CPR. Interestingly, Form III-like enzymes occur only in some CPR lineages and DPANN archaea, without any known representatives outside these radiations. Many DPANN lineages also encode the classical archaeal Form III-b, currently thought to have originated in the Methanomicrobia (Tabita et al., 2007). Recent phylogenomic studies of the Archaea have inferred a root in between DPANN and all other groups (Williams et al., 2017), requiring a transfer of Form III-b to DPANN from another lineage to explain this form's extant distribution (Step 8).

Our results also suggest the importance of lateral transfer processes in shaping the distribution of Form IV RuBisCO. For example, the Parcubacteria and Micrarchaeota Form IV enzymes are similar to those in characterized green sulfur bacteria, and may indicate transfer from this source (Steps 9, 12 in **Fig. 5**). Previous work has suggested that the IV-Photo enzyme is mobile and may have been transferred from Chlorobi to both Gammaproteobacteria and Alphaproteobacteria

(Steps 10, 11) (Tabita et al., 2007). Our results extend the breadth of Form IV-Photo RuBisCO distribution to new branches of the tree of life and add to the possible instances of its gene transfer, including one across domains. However, it is still unclear which lineage of bacteria among those that are known to bear this version of the enzyme is likely to have been the original source for the transfers to the CPR and DPANN. Similarly, we recovered RuBisCO enzymes related to the Form IV encoded in the genome of at least one Beckwithbacteria and many scaffolds of bacteriophage origin, including one unusually large, manually curated phage genome. One possibility is that a phage acquired a copy of the ancestral Form IV enzyme, termed the DeepYkr (Tabita et al., 2007), from a bacterial ancestor (Step 13), ultimately evolving into a divergent clade. One of these divergent sequences may then have been transferred to Beckwithbacteria (Step 14). The reverse scenario is also possible, in which members of the Myoviridae acquired this enzyme from Beckwithbacteria, possibly as prophages. However, to date, we know of no cases of >200 kbp phage genomes integrating into small (generally < 1 Mbp) CPR genomes. The discovery of RuBisCO homologs in phage also provides a possible mechanism for the widespread lateral gene transfer of this enzyme observable across the tree of life (**Fig. 5**) (Canchaya et al., 2003). However, as of yet, phage encoding 'bona fide' RuBisCO forms have not been identified.

The Form II/III enzymes are presently distributed among the CPR, DPANN, and at least one methanogenic archaeal lineage (Alonso et al., 2009; Wrighton et al., 2012). Given that this form is most closely related to Form II (**Fig. 1a**), we speculate here that the Form II/III enzyme evolved in a CPR lineage (possibly the Dojkabacteria) after transfer of a Form II sequence (Step 15). Form II/III could then be transferred to several other CPR lineages and one or more archaeal lineages (Step 16). Notably, no CPR with Form II enzymes have been reported to date.

Finally, there are two possible explanations for the apparent discordance between the phylogenetic pattern showing Forms II and II/III branching together and separate from Form I (**Fig. 1a**) and the pathway association of these Forms (I and II in CBB vs. II/III in the AMP pathway). If Form II/III preserves its ancestral function in the AMP pathway then the CBB pathway function in Forms I and II must have arisen by convergent evolution. Alternatively, the CBB pathway function in Forms I and II shared a common ancestor and Form II/III reverted back to function in the AMP pathway, possibly due to loss of the other CBB pathway enzymes. We suggest that convergent evolution of the more complex CBB pathway (which also requires PRK and transketolase, as well as various glycolysis enzymes) is less likely than reversion due to gene loss, especially given that gene loss is likely to have been widespread in the CPR. DPANN archaea, which generally do not have PRK, and methanogens could then have acquired the Form II/III RuBisCO by lateral transfer (**Fig. 5**).

Conclusion

In conclusion, we show that CPR bacteria, DPANN archaea, and bacteriophage harbor RuBisCO diversity that broadens our understanding of the distribution of this enzyme across the tree of life. The wide but sparse distribution of RuBisCO within the CPR and DPANN may be the consequence of extensive lateral gene transfer as well as gene loss. Further, some transfers may have catalyzed major shifts in carbon metabolism in bacterial lineages with limited metabolic repertoires. Specifically, lateral transfer of RuBisCO could have conferred a "missing puzzle piece" for organisms already bearing AMP phosphorylase and R15P isomerase, completing the genomic repertoire for the CO_2-incorporating AMP metabolism present in extant lineages. Likewise, Gottesmanbacteria may have evolved a partial CBB cycle or augmentation to the AMP pathway by linking laterally acquired RuBisCO and PRK to genes in the oxidative pentose phosphate pathway (**Fig. 4, Fig. S6**). Clearly, metagenomic studies of diverse environments can help to shed light on phylogenetic distribution and also to extend models of evolution for even well-studied enzymes.

2.5 Figures

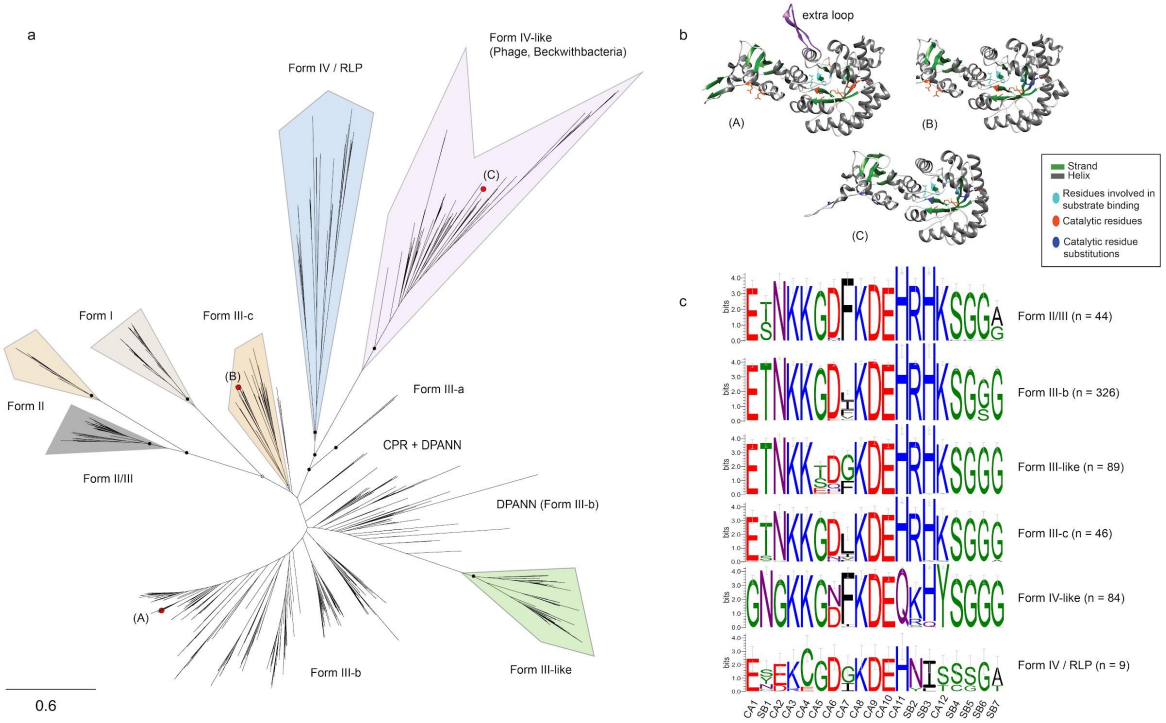

Figure 1. Known RuBisCO diversity is expanded by metagenomics. (a) Maximum-likelihood tree for de-replicated RuBisCO large-chain sequences, delineated by "Form." (A), (B), (C) show the phylogenetic position of protein sequences modeled in (b). Scale bar represents the number of substitutions per site. Closed black circles indicate bootstrap support values >70%, while open circles represent those >50%. For simplicity, bootstrap support below Form level is not shown. (c) Aggregate sequence logos for CPR, DPANN, and phage protein sequences in each phylogenetic "Form," describing key catalytic activity (CA) and substrate binding (SB) residues in RuBisCO sequences. Logo colors represent residues with similar chemical properties.

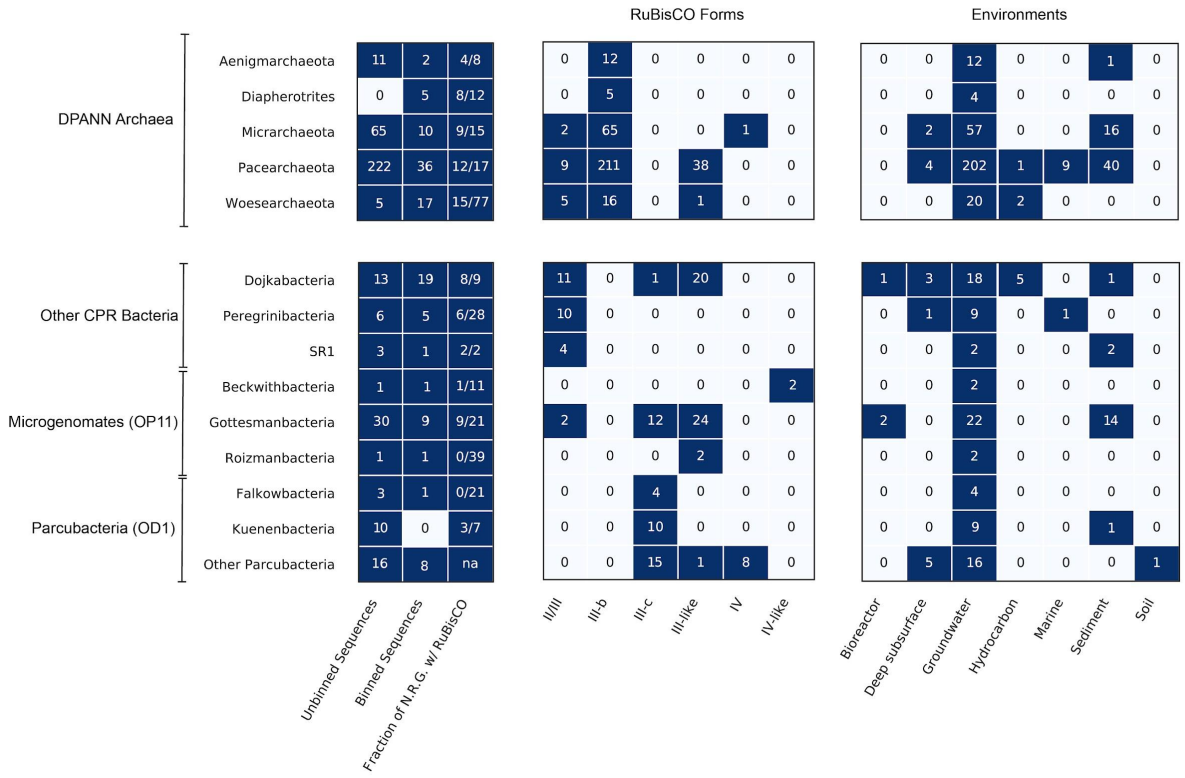

Figure 2. RuBisCO diversity among the CPR bacteria and DPANN archaea. Boxes represent counts of de-replicated protein sequences used in the phylogenetic analysis (Unbinned and Binned Sequences) as well as the number of those sequences that fell into described RuBisCO Forms and environments. Fraction of non-redundant genomes (N.R.G.) with RuBisCO describes the proportion of genomes per phylum encoding RuBisCO after dereplication at 99% ANI (see Methods).

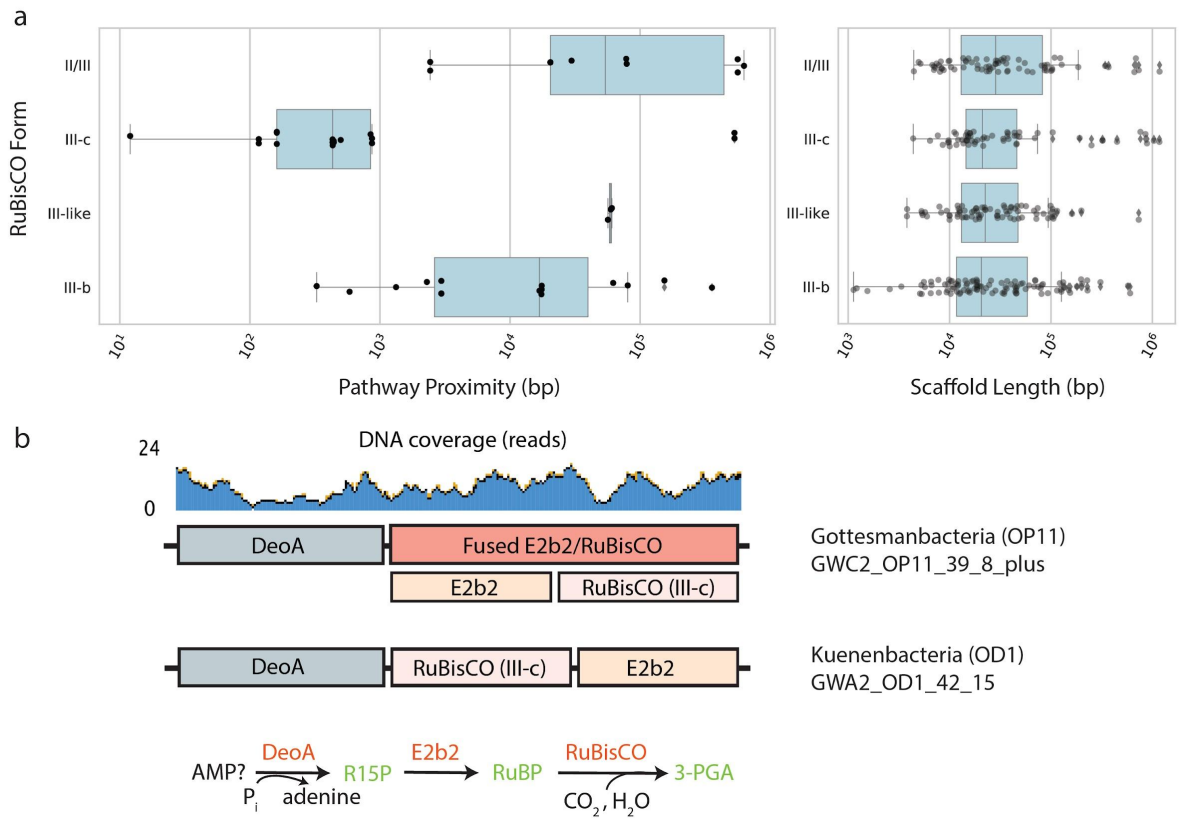

Figure 3. Genomic context of Form III-c RuBisCO in CPR and DPANN. (a) Proximity of genes on genome fragments encoding all three components of the CO_2-incorporating AMP pathway, and the length of all binned fragments containing the specified RuBisCO Form. (b) Genomic diagram of AMP components in CPR genomes with fused RuBisCO and consecutive ordering of genes on the chromosome.

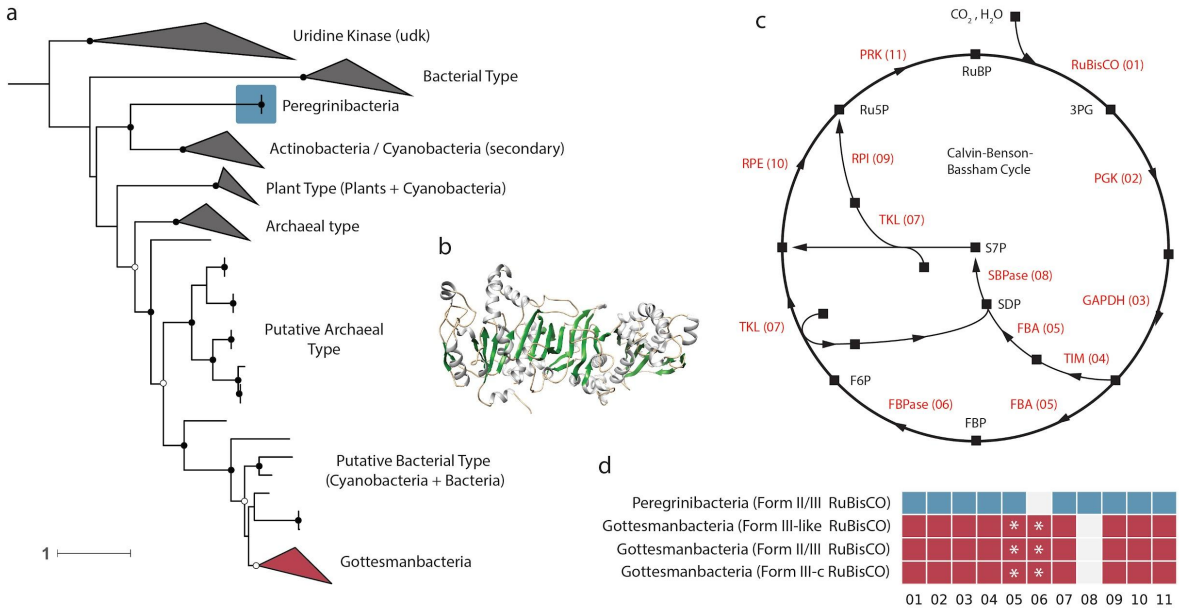

Fig. 4. Some CPR bacteria encode a putative phosphoribulokinase (PRK). (a) Maximum-likelihood tree showing phylogenetic position of putative PRK in CPR phyla. Scale bar represents the number of substitutions per site. Closed black circles indicate bootstrap support values >70%, while open circles represent those >50%. See Fig. S5 for fully labeled tree. (b) Example protein model of putative PRK in the CPR phylum Gottesmanbacteria. (c) Schematic of the Calvin-Benson-Bassham Cycle. Squares represent molecular intermediates, while arrows represent enzymatic steps. Abbreviations: PGK, phosphoglycerate kinase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; TIM, triosephosphate isomerase; FBA, fructose-bisphosphate aldolase; FPBase, fructose 1,6-bisphosphatase; TKL, transketolase; SBPase, sedoheptulose-bisphosphatase; RPI, ribose-5-phosphate isomerase; RPE, ribulose 5-phosphate 3-epimerase. (d) Genomic repertoires of CPR bacteria encoding PRK. Numbers refer to enzymatic steps in (c). Asterisks indicate genomes that harbor an enzyme with highest homology to a bifunctional FBA/FBPase instead of separate FBA and FBPase.

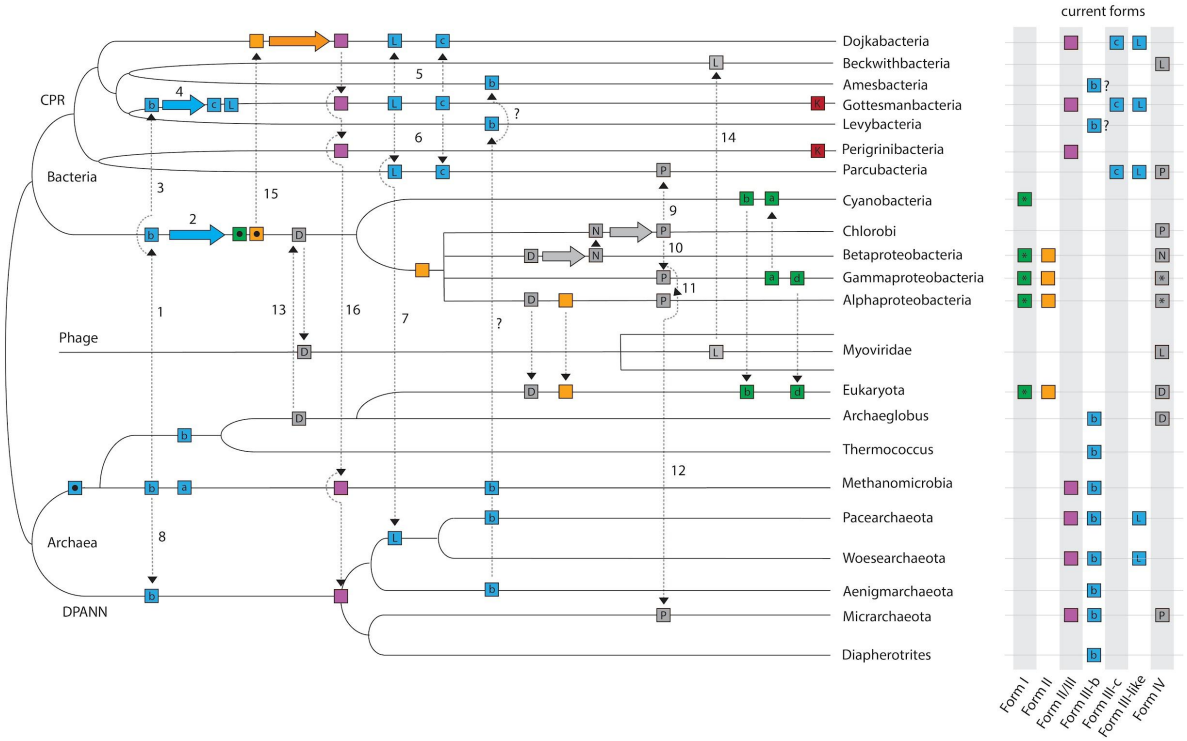

Fig. 5. Conceptual diagram illustrating the role of lateral gene transfer in the evolution of RuBisCO and PRK (K). Distinct RuBisCO Forms are represented by boxes of different colors, and Form subtypes are indicated by a letter within the box if applicable. Asterisks indicate multiple subtypes. Form III enzymes are expanded for additional clarity and are abbreviated as follows: a, Form III-a; b, Form III-b; c, Form III-c, L, Form III-like. Form IV abbreviations are as follows: D, DeepYkr; N, NonPhoto; P, Photo (see Tabita et al., 2007); L, Form IV-like (this study). Ancestral sequences are represented by black dots inside boxes. Dotted arrows represent possible lateral gene transfers, solid colored arrows represent evolution of RuBisCO within a lineage. Step numbers are referenced in the Discussion section. **N.B.** This tree does not convey time-calibrated information and is arranged to optimize conceptual understanding over accurate evolutionary relationships.

3. Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in CPR bacteria

Alexander L. Jaffe, Alex D. Thomas, Christine He, Ray Keren, Luis E. Valentin-Alvarado, Patrick Munk, Keith Bouma-Gregson, Ibrahim F. Farag, Yuki Amano, Rohan Sachdeva, Patrick T. West, and Jillian F. Banfield

Published in *Mbio,* 2021.

Candidate Phyla Radiation (CPR) bacteria are small, likely episymbiotic organisms found across Earth's ecosystems. Despite their prevalence, the distribution of CPR lineages across habitats and the genomic signatures of transitions among these habitats remain unclear. Here, we expand the genome inventory for Absconditabacteria (SR1), Gracilibacteria, and Saccharibacteria (TM7), CPR bacteria known to occur in both animal-associated and environmental microbiomes, and investigate variation in gene content with habitat of origin. By overlaying phylogeny with habitat information, we show that bacteria from these three lineages have undergone multiple transitions from environmental habitats into animal microbiomes. Based on co-occurrence analyses of hundreds of metagenomes, we extend the prior suggestion that certain Saccharibacteria have broad bacterial host ranges and constrain possible host relationships for Absconditabacteria and Gracilibacteria. Full-proteome analyses show that animal-associated Saccharibacteria have smaller gene repertoires than their environmental counterparts and are enriched in numerous protein families, including those likely functioning in amino acid metabolism, phage defense, and detoxification of peroxide. In contrast, some freshwater Saccharibacteria encode a putative rhodopsin. For protein families exhibiting the clearest patterns of differential habitat distribution, we compared protein and species phylogenies to estimate the incidence of lateral gene transfer and genomic loss occurring over the species tree. These analyses suggest that habitat transitions were likely not accompanied by large transfer or loss events, but rather were associated with continuous proteome remodeling. Thus, we speculate that CPR habitat transitions were driven largely by availability of suitable host taxa, and were reinforced by acquisition and loss of some capacities.

***N.B**. All main figures for this manuscript can be found below in section 3.5. All supplementary files (including figures and tables) can be found online with the published manuscript.*

3.1 Introduction

The Candidate Phyla Radiation (CPR) is a phylogenetically diverse clade of bacteria characterized by reduced metabolisms, potentially episymbiotic lifestyles, and ultrasmall cells (He et al., 2015, 2021; Luef et al., 2015). While the first high-quality CPR genomes were primarily from groundwater, sediment, and wastewater (Albertsen et al., 2013; Kantor et al., 2013; Wrighton et al., 2012), subsequently genomes have been recovered from diverse environmental and animal-associated habitats, including humans. Intriguingly, from dozens of major CPR lineages, only three – Candidatus Absconditabacteria (formerly SR1), Gracilibacteria (formerly BD1-5 and GN02), and Saccharibacteria (formerly TM7) – are consistently associated with animal oral cavities and digestive tracts (McLean et al., 2020). The Saccharibacteria are perhaps the most deeply studied of all CPR lineages to date, likely due to their widespread presence in human oral microbiomes and association with disease states such as gingivitis and periodontitis (Abusleme et al., 2013; Bor et al., 2019). On the other hand, Absconditabacteria and Gracilibacteria remain deeply undersampled, potentially due to their rarity in microbial communities or their use of an alternative genetic code that may confound some gene content analyses (Campbell et al., 2013; Hanke et al., 2014; Wrighton et al., 2012).

Absconditabacteria, Gracilibacteria, and Saccharibacteria are predicted to be obligate fermenters, dependent on other microorganisms (hosts) for components such as lipids, nucleic acids, and many amino acids (Kantor et al., 2013; Wrighton et al., 2012). Despite a generally reduced metabolic platform, CPR bacteria display substantial variation in their genetic capacities, even within lineages (Jaffe et al., 2020; Méheust et al., 2019). For example, some Gracilibacteria lack essentially all genes of the glycolysis and pentose-phosphate pathways and the TCA cycle (Sieber et al., 2019). In contrast to many CPR, soil-associated Saccharibacteria encode numerous genes related to oxygen metabolism (Nicolas et al., 2021; Starr et al., 2018). Pangenome analyses have shown genetic evidence for niche partitioning among Saccharibacteria from the same body site (Shaiber et al., 2020). However, the lack of comprehensive genomic sampling of these three CPR lineages across habitats, particularly from environmental biomes, has left unclear the full extent to which CPR gene inventories vary with habitat type, and, relatedly, the extent to which changes in metabolic capacities might have been reshaped during periods of environmental transition. Of particular interest is whether rapid gene acquisitions (e.g., via lateral gene transfer) or losses enabled habitat switches, or if these changes occurred gradually following habitat change.

The availability of suitable hosts may also drive the colonization of new environments by CPR bacteria (Shaiber et al., 2020). While there has been significant progress in characterizing the relationship between Saccharibacteria and Actinobacteria in the oral habitat (He et al., 2015; Murugkar et al., 2020; Utter et al., 2020), other CPR-host relationships remain unclear. Elucidation of environmental transitions among CPR lineages will require both thorough analysis of functional repertoires as well as a more comprehensive understanding of associations

with other microorganisms. Here, we improve existing sampling of CPR genomes and their surrounding communities to examine patterns of distribution, abundance, and gene content in different microbiome types. We also make use of whole-community co-occurrence patterns to shed light on the potential host range of CPR bacteria in their associated ecosystems. In combination, our analyses shed light on habitat shifts in three CPR lineages and the evolutionary processes likely underlying them.

3.2 Materials and Methods

Genome database preparation and curation

To compile an environmentally comprehensive set of genomes from the selected CPR lineages, we first queried four genomic information databases - GTDB (https://gtdb.ecogenomic.org/), NCBI assembly (https://www.ncbi.nlm.nih.gov/assembly), PATRIC (https://www.patricbrc.org/), and IMG (https://img.jgi.doe.gov/) - for records corresponding to the Absconditabacteria, Gracilibacteria, and Saccharibacteria genomes. Genomes gathered from these databases were combined with those drawn from several recent publications as well as genomes newly binned from metagenomic samples of sulfidic springs, an advanced treatment system for potable reuse of wastewater, human saliva, cyanobacterial mats, fecal material from primates, baboons, pigs, goats, cattle, and rhinoceros, several deep subsurface aquifers, dairy-impacted groundwater and associated enrichments, multiple bioreactors, soil, and sediment (Table S1). Assembly, annotation, and binning procedures followed those from Anantharaman et al. (Anantharaman et al., 2016). In some cases, manual binning of the alternatively coded Absconditabacteria was aided by a strategy in which a known Absconditabacteria gene was blasted against predicted metagenome scaffolds to find ‘seed’ scaffolds, whose coverage and GC profile were used to probe remaining scaffolds for those with similar characteristics. For newly binned genomes, genes were predicted for scaffolds > 1 kb using prodigal (“meta” mode) and annotated using USEARCH against the KEGG, UniProt, and UniRef100 databases. Bins were ‘polished’ by removing potentially contaminating scaffolds with phylogenetic profiles that deviated from consensus taxonomy at the phylum level. One genome was further manually curated to remove scaffolding errors identified by read mapping, following the procedures outlined in (Chen et al., 2020).

We removed exact redundancy from the combined genome set by identifying identical genome records and selecting one representative for downstream analyses. We then computed contamination and completeness for the genome set using a set of 43 marker genes sensitive to described lineage-specific losses in the CPR (Anantharaman et al., 2016; Brown et al., 2015) using the custom workflow in CheckM (Parks et al., 2015). Results were used to secondarily

filter the genome set to those with ≥ 70% of the 43 marker genes present and ≤ 10% of marker genes duplicated. The resulting genomes were then de-replicated at 99% ANI using dRep (-sa 0.99 -comp 70 -con 10) (Olm et al., 2017), yielding a set of 389 non-redundant genomes from a starting set of 868 Existing metadata were used to assign both "broad" and "narrow" habitat of origin for each non-redundant genome. The "engineered" habitat category was defined to include human-made or industrial systems like wastewater treatment, bioreactors, and water impacted by farming/mining. Curated metadata, along with accession/source information for each genome in the final set, is available in Table S1. All newly binned genomes are available through Zenodo (Data and Software Availability).

Functional annotation and phylogenomics

We predicted genes for each genome using prodigal ("single" mode), adjusting the translation table (*-g 25*) for Gracilibacteria and Absconditabacteria, which are known to utilize an alternative genetic code (Campbell et al., 2013; Hanke et al., 2014). Predicted proteins were concatenated and functionally annotated using kofamscan (Aramaki et al., 2020). Results with an e-value ≤ 1e-6 were retained and subsequently filtered to yield the highest scoring hit for each individual protein.

To create a species tree for the CPR groups of interest, functional annotations from kofamscan were queried for 16 syntenic ribosomal proteins (rp16). Marker genes were combined with those from a set of representative sequences of major, phylogenetically proximal CPR lineages (Jaffe et al., 2020). Sequences corresponding to each ribosomal protein were separately aligned with MAFFT and subsequently trimmed for phylogenetically informative regions using BMGE (*-m BLOSUM30*) (Criscuolo and Gribaldo, 2010). We then concatenated individual protein alignments, retaining only genomes for which at least 8 of 16 syntenic ribosomal proteins were present. A maximum-likelihood tree was then inferred for the concatenated rp16 (1976 amino acids) set using ultrafast bootstrap and IQTREE's extended Free-Rate model selection (*-m MFP -st AA -bb 1000*) (Nguyen et al., 2015). The maximum likelihood tree is available as File S1. The tree and associated metadata were visualized in iTol (Letunic and Bork, 2016) where well-supported, monophyletic subclades were manually identified within Gracilibacteria and Saccharibacteria for use in downstream analysis.

Abundance analysis

To assess the global abundance of Absconditabacteria, Gracilibacteria, and Saccharibacteria, we manually compiled the original read data associated with each genome in the analysis set, where available. We included only those genomes from short-read, shotgun metagenomics of microbial entire communities (genomes derived from single cell experiments, stable isotope probing experiments, "mini" metagenomes, long-read sequencing experiments, and co-cultures were

excluded). For each sequencing experiment, we downloaded the corresponding raw reads and, where appropriate, filtered out animal-associated reads by mapping to the host genome using bbduk (*qhdist=1)*. Sequencing experiments downloaded from the NCBI SRA database were sub-sampled to the average number of reads across all compiled experiments (~36 million) using seqtk (*sample -s 7*) if the starting read pair count exceeded 100 million. We then removed Illumina adapters and other contaminants from the remaining reads and further quality trimmed them using Sickle. The filtered read set was then mapped against all genomes assembled (or co-assembled) from it using bowtie2 (default parameters). For mappings with a non-zero number of read alignments, abundance of each genome was calculated by counting the number of stringently mapped (≥99% identity) using CoverM (*--min-read-percent-identity 0.99*) (https://github.com/wwood/CoverM) and dividing by the total number of reads in the quality-filtered read set. In most cases where genomes were derived from co-assemblies of multiple sequencing experiments, we computed the abundance for each sample individually and then selected the one with the highest value as a 'representative' sample for downstream analyses. To account for lower sequence representation of co-assembled genomes in individual samples, we considered genomes present if at least 10% of their sequence length was covered by reads.

Co-occurrence analyses

Each representative sample was then probed for co-occurrence patterns of CPR and potential host lineages. To account for across-study differences in binning procedures, quality-filtered read sets were instead re-assembled using MEGAHIT (*--min-contig-len 1000*) and analyzed using GraftM (Boyd et al., 2018) with a ribosomal protein S3 (rpS3) gpackage custom built from GTDB (release 05-RS95) (Crits-Christoph et al., 2021). Recovered rpS3 protein sequences in each sample were clustered to form 'species groups' at 99% identity using USEARCH cluster_fast (*-sort length -id 0.99*). For all samples with >0 marker hits, we then performed three downstream analyses to examine patterns of co-occurrence for various taxa. First, we counted the number of unique species groups in each sample taxonomically annotated as Saccharibacteria (*"c__Saccharimonadia"*) and Actinobacteria ("*p__Actinobacteriota"*), dividing the former by the latter to compute a species 'richness ratio' for each sample (where *p__Actinobacteriota* did not equal 0). Animal-associated samples without detectable rpS3 from Actinobacteria were secondarily profiled for ribosomal protein L6 from Actinobacteria using the same methodology as described above.

Second, to examine the co-occurrence of Saccharibacteria and Actinobacteria within a phylogenetic framework, we inferred maximum likelihood trees for the set of rpS3 marker genes recovered across samples. Species group sequences were clustered *across* samples to further reduce redundancy using USEARCH (as described above) and were combined with rpS3 sequences drawn from a taxonomically balanced set of bacterial reference genomes (Jaffe et al.,

2020) as an outgroup. Saccharibacteria and Actinobacteria sequence sets were then aligned, trimmed, and used to build trees as described above for the 16 ribosomal protein tree, with the exception of using trimal (*-gt 0.1*) (Capella-Gutiérrez et al., 2009) instead of BMGE. Species groups that co-occurred in one or more metagenomic samples were then noted. If a given Saccharibacteria species group exclusively co-occurred with an Actinobacteria species group in *at least* one sample, or Actinobacteria species groups belonging to the same order level in *all* samples, those linkages were labelled. Finally, experimental co-cultures of Saccharibacteria and Actinobacteria from previous studies were mapped onto the trees. To do this, we compiled a list of strain pairs and their corresponding genome assemblies (Table S4) and then used GraftM to extract rpS3 sequences from corresponding genome assemblies downloaded from NCBI. We then matched these rpS3 sequences to their closest previously defined species group using blastp (*-evalue 1e-3 -max_target_seqs 10 -num_threads 16 -sorthits 3 -outfmt 6*), prioritizing hits with the highest bitscore and alignment length. Reference rpS3 sequences with no match at ≥99% identity and ≥95% coverage among the species groups were inserted separately into the tree. We then labelled all experimental pairs of species in the linkage diagram.

Third, we profiled a subset of 43 metagenomes containing Gracilibacteria and Absconditabacteria for overall community composition. For each sample, we extracted all contigs bearing rpS3 and mapped the corresponding quality filtered read set to them using bowtie2. Mean coverage for each contig was then computed using CoverM (*contig --min-read-percent-identity .99*) and a minimum covered fraction of 0.10 was again employed. Relative coverage for each order level lineage (as predicted by GraftM) was computed by summing the mean coverage values for all rpS3-bearing contigs belonging to that lineage. Where species groups did not have order-level taxonomic predictions, the lowest available rank was used. Finally, relative coverage values were scaled by first dividing by the lowest relative coverage observed across samples and then taking the base-10 log. For the re-analysis of 22 cyanobacterial mat metagenomes (Bouma-Gregson et al., 2019), the same approach was taken, and coverage profiles for rpS3-bearing scaffolds were correlated using the pearsonr function in the scipy.stats package.

Proteome size, content, enrichment

We subjected all predicted proteins from the genome set to a two-part, de novo protein clustering pipeline recently applied to CPR genomes, in which proteins are first clustered into "subfamilies" and highly similar/overlapping subfamilies are merged using and HMM-HMM comparison approach (--coverage 0.70) (Méheust et al., 2019). For each protein cluster, we recorded the most common KEGG annotation among its member sequences and the percent of sequences bearing this annotation (e.g. 69% of sequences in fam00095 were matched with K00852).

We then performed three subsequent analyses to describe broad proteome features of included CPR. First, we computed proteome size across habitats, defined as the number of predicted ORFs per genome when considering genomes at increasing thresholds of completeness in single copy gene inventories (75%, 80%, 85%, etc.). Second, we examined similarity between proteomes by generating a presence/absence matrix of protein families with 5 or more member sequences. We then used this matrix to compute distance metrics between each genome based on protein content using the ecopy package in Python (method='jaccard', transform='1') and performing a principal coordinates analysis (PCoA) using the skbio package. The first two axes of variation were retained for visualization alongside environmental and phylogenetic metadata. Finally, we used the clustermap function in seaborn (*metric='jaccard', method='average'*) to hierarchically cluster the protein families based on their distribution patterns and plot these patterns across the genome set. For each protein family, we also computed the proportion of genomes encoding at least one member sequence that belonged to each of the three CPR lineages and each broad environmental category (Fig. 4, bottom panel) (see custom code linked in Data and Software Availability).

We next identified protein families that were differentially distributed among genomes from broad environmental categories. For each protein family, we divided the fraction of genomes from a given habitat ('in-group') encoding the family by the same fraction for genomes from all other habitats ('out-group'). In cases where no 'out-group' genome encoded a member protein, the protein family was simply noted as 'exclusive' to the 'in-group' habitat. In all cases, we calculated the Fisher's exact statistic using the *fisher_exact* function in scipy.stats (*alternative='two-sided').* To account for discrepancies in genome sampling among lineages, we determined ratios and corresponding statistical significance values separately for each lineage. All statistical comparisons for a given lineage were corrected for false discovery rate using the *multipletests* function in statsmodels.stats.multitest (*method="fdr_bh"*). Finally, we selected families that were predicted to be enriched or depleted in particular habitats. We considered enriched families to be those with ratios ≥ 5, and depleted families as those that were encoded in 10% or fewer of genomes from a given habitat but present in 50% or more of genomes outside that environmental category. Retaining only those comparisons with corrected Fisher's statistics at FDR ≤ 0.05 resulted in a set of 926 unique, differentially distributed protein families for downstream analysis.

Analysis of putative rhodopsins

Protein sequences from the CPR (fam11249) were combined with a set of reference protein sequences spanning Type 1 bacterial/archaeal rhodopsin and heliorhodopsin (Pushkarev et al., 2018). Sequences were then aligned using MAFFT (*--auto)* and a tree was inferred using iqtree (*-m TEST -st AA -bb 1000*). Alignment columns with 95% or more gaps were trimmed manually in Geneious for the purposes of visualization. Transmembrane domains were identified by

BLASTp searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and conserved residues were defined by manual comparison with an annotated alignment of previously published reference sequences (Hasegawa et al., 2020).

Processes driving protein family evolution

To examine the evolutionary processes shaping the differentially-distributed protein families, we next subjected each family to an automated gene-species tree reconciliation workflow adapted from (Sheridan et al., 2020). Briefly, for each family, truncated sequences (defined as those with lengths less than 2 standard deviations from the family mean) were removed and the remaining sequences aligned with MAFFT (*--retree 2).* Resulting alignments were then trimmed using trimal (*-gt 0.1)* and used to infer maximum-likelihood phylogenetic trees using IQTree with 1000 ultrafast bootstrap replicates (*-bnni -m TEST -st AA -bb 1000 -nt AUTO).* We removed reference sequences from the inferred species tree and rooted it on the branch separating Saccharibacteria from the monophyletic clade containing Gracilibacteria and Absconditabacteria. A random sample of 100 bootstrap replicates were then used to probabilistically reconcile each protein family with the pruned species tree using the ALE package (*ALE_undated*) (Szöllősi et al., 2013). Estimates of missing gene fraction were derived from the CheckM genome completeness estimates described above. We then calculated the total number of originations (horizontal gene transfer from non-CPR, or *de novo* gene formation), within-CPR horizontal transfers, and losses over each non-terminal branch and mapped branch-wise counts for each event to a species-tree cladogram in iTol (Letunic and Bork, 2016).

3.3 Results

Environmental diversity, phylogenetic relationships, and abundance patterns

We gathered an environmentally comprehensive set of Absconditabacteria, Gracilibacteria, and Saccharibacteria by querying multiple databases for genomes assembled in previous studies and assembling new genomes from several additional metagenomic data sources (Table S1, Materials and Methods). Quality filtration of this curated genome set at ≥ 70% completeness and ≤ 10% contamination, plus subsequent de-replication at 99% average nucleotide identity (ANI), yielded a non-redundant set of 389 genomes for downstream analysis (Table S1). Absconditabacteria and Gracilibacteria were less frequently sampled relative to Saccharibacteria, comprising only ~7.5% and ~10.8% of the total genome set, respectively. All three lineages were distributed across a broad range of microbiomes, encompassing various environmental habitats (freshwater, marine, soil, engineered, plant-associated, hypersaline) as well as multiple animal-associated microbiomes (oral and gut) (Fig. 1). Unlike animal-associated Gracilibacteria and Absconditabacteria genomes, which were recovered primarily from human and animal oral samples, animal-associated Saccharibacteria were found in both oral and gut samples.

We extracted 16 syntenic, phylogenetically informative ribosomal proteins from each genome to construct a CPR species tree and evaluate how habitat of origin maps onto phylogeny. Sequences from related CPR bacteria were used as outgroups for tree construction (Materials and Methods). The resolved topology supports monophyly of all three lineages and a sibling relationship between the two alternatively coded lineages, Absconditabacteria and Gracilibacteria (Figure 1a, File S1), consistent with previous findings (Jaffe et al., 2020). For the Absconditabacteria, a single clade of organisms derived from animal-associated microbiomes was deeply nested within genomes from the environment. On the other hand, Gracilibacteria clearly formed two major lineages (GRA 1-2), each with a small subclade comprised of animal-associated genomes (Table S1). For Saccharibacteria, deeply-rooting lineages were also almost exclusively of environmental origin (soil, water, sediment) and animal-associated genomes were strongly clustered into at least three independent subclades (Table S1, Fig. 1a). Two of these three subclades were exclusively composed of animal-associated sequences whereas one (SAC 5), was a mixture of animal-associated, wastewater (potentially of human origin), and a few aquatic sequences. Intriguingly, for both Saccharibacteria and Gracilibacteria, a subset of organisms from the dolphin mouth (Dudek et al., 2017) did not affiliate with those from terrestrial mammals/humans and instead fell within marine/environmental clades (indicated by asterisks in Fig. 1a). In primarily environmental clades (SAC 1 and 4), genomes from soil, freshwater, engineered, and halophilic environments were phylogenetically interspersed, suggesting comparatively wide global distributions for these lineages. Exceptions to this pattern were two clades representing distinct hypersaline environments – a hypersaline lake and salt crust (Finstad et al., 2017; Vavourakis et al., 2018).

We used read mapping to assess the abundance of Absconditabacteria, Gracilibacteria, and Saccharibacteria genomes in the samples from which they were originally reconstructed, focusing only on those organisms from short-read, whole-community sequencing experiments (Materials and Methods). In total, abundance calculation was possible for 297 of the 389 genomes (~76%). Generally, these lineages of CPR bacteria are not dominant members of microbial communities (<1% of reads). However, they were relatively abundant in some engineered, animal-associated, and freshwater environments (Fig. 1b). In rare cases, CPR taxa comprised >10% of reads (Fig. 1b), and in a bioreactor (engineered) reached a maximum of ~22% of reads. Gracilibacteria and Absconditabacteria attained comparable read recruitment to Saccharibacteria and were particularly abundant in some groundwater, engineered, and animal-associated habitats. In contrast to Saccharibacteria, Gracilibacteria and Absconditabacteria have so far only been minimally detected in soil and plant-associated microbiomes. We also compared abundance patterns across animal body sites. As expected based on extensive prior work (Bor et al., 2019; Cross et al., 2019; He et al., 2015), Saccharibacteria exhibited highest read recruitment in the human oral microbiome. However, these bacteria can also comprise a significant fraction of the sequenced DNA in exceptional gut/oral microbiomes from cows, pigs, and dolphins (Fig. 1c), in one case approaching 5% of reads (Table S2). When

detected, Saccharibacteria in the human gut were relatively rare, comprising a median of ~0.1% of reads across samples.

Patterns of co-occurrence constrain CPR host range across environments

Despite recent progress made in experimentally identifying bacterial host ranges for oral Saccharibacteria, little is known about associations in other habitats. Abundance pattern correlations can be informative regarding associations involving obligate symbionts and their microbial hosts (Huddy et al., 2020; Probst et al., 2018); however, such analyses often rely on highly resolved time-series for statistical confidence. Here, we instead examine patterns of co-occurrence within samples to probe potential relationships between CPR bacteria and their microbial hosts. Given recent experimental evidence demonstrating the association of multiple Saccharibacteria strains with various Actinobacteria in the human oral microbiome (Cross et al., 2019; He et al., 2015; Murugkar et al., 2020; Soro et al., 2014; Utter et al., 2020), we predicted that Actinobacteria may be common hosts of Saccharibacteria in microbiomes other than the mouth and asked to what extent co-occurrence data supported this relationship.

We first identified all ribosomal protein S3 (rpS3) sequences from Actinobacteria and Saccharibacteria in the source metagenomes probed in this study for abundance patterns (Fig. 1bc). RpS3 sequences from all samples were clustered into 'species groups' (Materials and Methods). We observed that species groups from Actinobacteria and Saccharibacteria frequently co-occurred in the soil and plant-associated microbiomes as well as several hypersaline microbiomes (Fig. 2a). On the other hand, co-occurrence of the two lineages was less frequent in engineered and freshwater environments relative to other environments. Surprisingly, only ~78% of animal-associated samples containing Saccharibacteria also contained Actinobacteria at abundances high enough to be detected (Fig. 2a). The absence of Actinobacteria in the remaining animal-associated samples was confirmed with an additional marker gene, ribosomal protein L6 (rpL6) (Materials and Methods). Assemblies with well-sampled Saccharibacteria yet no detectible Actinobacteria could suggest that Saccharibacteria have alternative hosts in these samples or are able to (at least periodically) live independently. Alternatively, the lack of Actinobacteria rpS3/rpL6 in these samples could be the result of poor sequence assembly, e.g. due to population heterogeneity or low coverage.

For samples where both Saccharibacteria and Actinobacteria marker genes were detectable, we computed a 'relative richness' metric describing the ratio of distinct Saccharibacteria species groups to Actinobacteria species groups. In most animal-associated microbiomes, Actinobacteria were more species rich (lower richness ratios), as expected if individual Saccharibacteria can associate with multiple hosts (Fig. 2a). Greater species richness of Actinobacteria compared to Saccharibacteria was also observed for many plant-associated, soil, engineered, and freshwater microbiomes. However, some engineered and freshwater samples had richness ratios equal to (equal richness) or greater than 1 (i.e., Saccharibacteria more species rich) (Fig. 2a). Specifically,

we observed that several metagenomes from engineered and freshwater environments contained anywhere from 1-11 Saccharibacteria species but only one detectable Actinobacteria species (Table S3). Thus, if Actinobacteria serve as hosts for Saccharibacteria in these habitats, there may be both exclusive associations and associations linking multiple Saccharibacteria species with a single Actinobacteria host species.

We next tested for more specific possible associations in the animal microbiome, reasoning that if Actinobacteria are common hosts for Saccharibacteria, then exclusive co-occurrence of a particular Saccharibacteria species with a singular Actinobacteria species within a sample might suggest an interaction *in vivo*. We mapped all pairs of Saccharibacteria and Actinobacteria species that co-occurred within a single sample onto species trees constructed from recovered rpS3 sequences (Fig. 2b), including 22 Saccharibacteria-Actinobacteria pairs reported in previous experimental studies (Table S4). In three cases, we found that individual metagenomic samples contained only one assembled Saccharibacteria species group and one Actinobacteria species group ("exclusive co-occurrence - species group", Fig. 2b; Table S3). Two of these cases involved Actinobacteria from the order Actinomycetales, from which multiple Saccharibacteria hosts have already been identified (Bor et al., 2020). We also noted exclusive species-level co-occurrence of a Saccharibacteria species group from the human gut and an Actinobacteria species group from the order Coriobacterales (Table S3). In an additional seven cases, one Saccharibacteria species group occurred with multiple Actinobacteria species groups of the same order-level classification based on rpS3 gene profiling ("exclusive co-occurrence - order", Fig. 2b; Table S3). Five of the seven instances involved pairs of Saccharibacteria and Coriobacterales from termite and swine gut metagenomes. Thus, unlike in human oral environments, Coriobacterales may serve as hosts for Saccharibacteria in gut environments of multiple animal species. More generally, we also observed that Saccharibacteria from the same phylogenetic clade had predicted relationships to phylogenetically unrelated Actinobacteria (Fig. 2b), consistent with previous experimental observations for individual species (Cross et al., 2019).

Compared to Saccharibacteria, host relationships for Gracilibacteria and Absconditabacteria have received little attention. There are preliminary indications that Absconditabacteria may associate with members of the Fusobacteria or Firmicutes in the oral microbiome (Cross et al., 2019) or the gammaproteobacterium *Halochromatium* in certain salt lakes (Moreira et al., 2020). We thus explored co-occurrence patterns in microbial communities containing Absconditabacteria and Gracilibacteria, attempting to further constrain possible host taxa. In animal and human-associated microbiomes, bacteria from several lineages, including Fusobacteria (Fig. 2c), were relatively abundant in nearly all samples that contained Absconditabacteria. Members of the Chitinophagales, Pseudomonadales, and Acidimicrobiales were detected in high abundance in three wastewater samples from similar treatment plants (Martínez Arbas et al., 2021) and one dairy pond sample containing Absconditabacteria (Fig. 2c, Table S5). No clear patterns of potential host co-occurrence were observed for Gracilibacteria, with the exception of the Proteobacterial order Campylobacterales, which co-occurred in 8 of 10 groundwater samples

where Gracilibacteria were found (Fig. 2c). Across all habitat types, only members of the order Burkholderiales (a large order of Gammaproteobacteria) consistently co-occurred with Gracilibacteria; however, these organisms were also abundant in a number of samples without detectable Gracilibacteria, weakening the potential association.

Among the least complex communities that contained Absconditabacteria were cyanobacterial mats from a California river network, where dominant cyanobacterial taxa accounted for ~60-98% relative abundance (Bouma-Gregson et al., 2019). To complement the above co-occurrence analyses, we re-analyzed 22 published metagenomes representing spatially separated mats and discovered that Absconditabacteria were detectable in 12 of them at varying degrees of coverage (0.12-37x). As noted previously, also present in the mats were members of the phyla Bacteroidetes, Betaproteobacteria, and Verrucomicrobia (Bouma-Gregson et al., 2019). Correlation of read coverage profiles across mats provided moderate support for the association of Absconditabacteria and Bacteroidetes. Specifically, many of the strongest species-level correlations, including five of the top ten, involved Bacteroidetes (Table S6).

Gene content of Absconditabacteria, Gracilibacteria, and Saccharibacteria

We next examined how gene content of these CPR lineages varied across environments. We first compared the predicted proteome size of these bacteria across habitats, taking into consideration differing degrees of genome completeness. This analysis revealed that genomes from soil and the rhizosphere (plant-associated) have on average larger predicted proteomes relative to those from animal-associated environments (Fig. 3a). Saccharibacteria from hypersaline environments appear to have the smallest predicted proteomes, although the limited number of high-quality genomes in this category currently limits a firm conclusion. We observed some evidence for variance in predicted proteome sizes among Absconditabacteria and Gracilibacteria, including potentially smaller predicted proteomes among animal-associated Gracilibacteria (Fig. S1). Additional high quality genomes will be required to confirm this trend.

To examine overall proteome similarity as a function of habitat type, we employed a recently developed protein-clustering approach that is agnostic to functional annotation (Méheust et al., 2019) (Table S7, Materials and Methods). Among Saccharibacteria, principal coordinates analysis (PCoA) of presence/absence profiles for all protein families with 5 or more members yielded a primary axis of variation (~12% variance explained) that distinguished animal-associated Saccharibacteria from environmental or plant-associated ones and a secondary axis (~8% variance explained) that distinguished between phylogenetic clades (SAC1-3 vs 4-5). We did not observe strong clustering of Saccharibacteria by specific environmental biome, consistent with the interspersed nature of their phylogenetic relationships (Fig. 1a, 3b). Notably, several SAC5 genomes from wastewater have protein family contents that are intermediate between those of animal-associated Saccharibacteria and Saccharibacteria from the large

environmental clade (indicated by an asterisk in Fig. 3b). This finding may indicate selection within the engineered environments for variants introduced from human waste. PCoAs of predicted proteome content among Absconditabacteria and Gracilibacteria generally showed that, with the exception of dolphin-derived genomes, animal-associated lineages are also distinct from their relatives from environmental biomes (Fig. S2). Overall, our results indicate that the CPR lineages examined here have predicted proteomes whose content and size vary substantially with their environment. This is particularly evident for animal-associated Saccharibacteria, which are notably dissimilar in their protein family content compared to environmental counterparts.

To further examine the distinctions evident in the PCoA analysis, we arrayed presence/absence information for each protein family and hierarchically clustered them based on their distribution patterns across all three CPR phyla. This strategy allowed us to explore specific protein family distributions and to test for groups of co-occurring protein families (modules) that are common to bacteria from a single lineage or are shared by most bacteria from one or more CPR lineages. We first observed one large module that is generally conserved across all genomes. This module is comprised of families for essential cellular functions such as transcription, translation, cell division, and basic energy generating mechanisms (Fig. 4, “core”).

The protein family analysis also revealed multiple modules specific to Gracilibacteria, Absconditabacteria, and modules shared by both lineages but not present in Saccharibacteria, paralleling their phylogenetic relationships (Fig. 1a, 4). Of the ~70 families shared only by Gracilibacteria and Absconditabacteria (M2, Fig. 4), nearly half had no KEGG annotation at the thresholds employed. One family shared by these phyla but not in Saccharibacteria is the ribosomal protein L9, which supports prior findings on the composition of Saccharibacteria ribosomes (Brown et al., 2015). The remaining families also include two that were fairly confidently annotated as the DNA mismatch repair proteins, MutS and MutL (fam01378 and fam00753), nicking endonucleases involved in correction of errors made during replication (Yamamoto et al., 2011) (Table S7). Despite the generally wide conservation of these proteins among Bacteria, we saw no evidence for the presence of either enzyme in Saccharibacteria, suggesting that aspects of DNA repair may vary in this group relative to other CPR. We recovered a module of approximately 60 proteins highly conserved among the Saccharibacteria and only rarely encoded in the other lineages (M5, Fig. 4). This module contained several protein families confidently annotated as core components of glycolysis and the pentose phosphate pathway, including three enzymes present in almost all CPR (Jaffe et al., 2020): glyceraldehyde 3-phosphate dehydrogenase, (GAPDH) triosephosphate isomerase (TIM), and phosphoglycerate kinase (PGK). These results indicate that Gracilibacteria and Absconditabacteria may have extremely patchy, if not entirely lacking, components of core carbon metabolism, even when a high-quality genome set is considered.

For all three lineages of CPR, we also observed numerous small modules with narrow distributions. To test whether these modules represent functions differentially distributed among organisms from different habitats, we computed ratios describing the incidence of each protein family in genomes from one habitat compared to those from all other habitats (Materials and Methods). Enriched families were defined as those with ratios ≥ 5, whereas depleted families were defined as those that were encoded by <10% of genomes in a given habitat, but ≥ 50% of genomes from other habitats. To account for the fact that small families might appear to be differentially distributed due to chance alone, we also stipulated that comparisons be statistically significant ($p \leq 0.05$, two-sided Fisher's exact test corrected for multiple comparisons).

Using this approach, we identified 926 families that were either enriched (n=872) or depleted (n=54) in genomes from one or more broad habitat groups. We identified 45 families enriched in Absconditabacteria from animal-associated environments relative to those from environmental biomes. The majority of these families were either poorly functionally characterized or entirely without a functional annotation at the thresholds employed. Similarly, families enriched in animal-associated Gracilibacteria relative to environmental counterparts were primarily unannotated; among those families with confident annotations was a family likely encoding a phosphate:Na+ symporter (fam04488) and a putative membrane protein (fam06579). Intriguingly, 6 families were co-enriched in both animal-associated Gracilibacteria and Absconditabacteria, suggesting that these sibling lineages might have acquired or retained a small complement of genes that are important in adaptation to animal habitats or their associated bacteria.

Animal-associated Saccharibacteria, on the other hand, encoded 417 unique families that were exclusive or highly enriched relative to those from other habitats. Enriched families largely fell into three major groups (M1, M3, M6; Fig. 4), and the large majority of them, particularly among modules with narrow, lineage-specific distributions, were without functional annotations. However, our analysis also revealed some protein families with broader distributions across multiple clades of animal-associated Saccharibacteria (Fig. 4). Here, among families with functional annotations, we found several apparently involved in the transport of amino acids and dicarboxylates that were highly enriched (ratios ranging from 10.7 to 112.9) in the majority of animal-associated Saccharibacteria (52-58% of genomes across clades) (Table S8). Two of these families, corresponding to a putative amino acid transport permease and substrate-binding protein (fam00393 and fam11477, respectively) were co-located in some genomes along with a ATP-binding protein (subset of fam00001), suggesting that they may function together to uptake amino acids. We also recovered several other functions that were previously predicted to be enriched based on analysis of a smaller set of animal-associated Saccharibacteria (McLean et al., 2020), including phosphoglycerate mutase, glycogen phosphorylase, and a uracil-DNA glycosylase (ratio 8.3-33.5). Lastly, we found that one family encoding the CRISPR-associated protein csn1/cas9 (fam00646) was also enriched among animal-associated genomes (ratio ~12.4

among 28 genomes), consistent with the suggestion that some Saccharibacteria likely acquired their viral defense systems after colonizing animals (Table S8) (McLean et al., 2020).

We identified multiple families that are either enriched or depleted in animal-associated Saccharibacteria that were functionally related to oxidative stress (Table S8). Among enriched families, one (fam00662) set was mostly annotated with low confidence as rubrerythrin, a family of iron-containing proteins generally involved in detoxification of peroxide (Cardenas et al., 2016). Member sequences of this family were present in over a third of animal-associated Saccharibacteria and were highly enriched relative to environmental genomes (fold-enrichment ratio of 36.2), suggesting that acquisition may have conferred an adaptive benefit in the gut and/or oral cavity. In contrast, we also observed that animal-associated Saccharibacteria were significantly depleted in another family confidently annotated as a Fe-Mn family superoxide dismutase (fam01569) and likely involved in radical detoxification. Animal-associated lineages were also strongly depleted for the genes comprising the cytochrome o ubiquinol oxidase operon (fam00281, fam00112, fam01347, fam00624, and fam10494), with very few, if any, animal-associated genomes and more than 50% of environmental genomes encoding each of the five genes. This operon has been previously suggested to confer an advantage in aerophilic environments like soil through detoxification (Kantor et al., 2013) or use of oxygen (Nicolas et al., 2021; Starr et al., 2018)

Among genomes from environmental biomes, we identified a module of approximately 100 protein families, also primarily without functional annotation, that were associated with a subclade of Saccharibacteria recently reconstructed from metagenomes of freshwater lakes and glacier ice (M4, Fig. 4) (Rissanen et al., 2020; Zeng et al., 2020). Notably, among the most widespread families in this module was one in which sequences were annotated as bacteriorhodopsin with low confidence (fam11249). Further analysis indicated that these sequences fall within the bacterial/archaeal Type 1 rhodopsin clade and contain both the retinal-binding lysine associated with light sensitivity and a DTS motif (Fig. S3), suggesting that they may function as proton pumps (Béjà and Lanyi, 2014; Maliar et al., 2020). Distinct rhodopsin sequences were also recovered in the genomes of environmental Absconditabacteria (NDQ motif) and Gracilibacteria (DTE motif), although they were not statistically enriched (Fig. S3). Genomes of soil-associated Saccharibacteria were enriched for nearly 130 protein families largely without strong functional annotations (Fig. 4, Table S8)). Despite their small proteome sizes, Saccharibacteria from hypersaline environments were only statistically depleted in about 15 families at the thresholds employed here. Sequence files for all protein families are provided in the Supplementary Materials (File S2).

Evolutionary processes shaping proteome evolution

The observation that some differentially distributed traits among CPR were apparently lineage specific, whereas others were more widespread, motivated us to examine the relative contributions of gene transfer and loss to proteome evolution. To do so, we first inferred unrooted, maximum-likelihood phylogenies for the sequences in each protein family that was differentially distributed, then compared these phylogenies to the previously reconstructed species tree (Materials and Methods). For each family, the likelihood of transfer and loss events on each branch of the species tree were then estimated using a probabilistic framework that takes into consideration genome incompleteness, variable rates of transfer and loss, and uncertainty in gene tree reconstruction (Sheridan et al., 2020; Szöllősi et al., 2013). The results of this analysis reveal relatively few instances of originations, defined as lateral transfer from outside the three lineages of CPR or *de novo* evolution ('originations', Fig. 5). In the Absconditabacteria and Gracilibacteria, gene-species tree reconciliation revealed that small modules of families of mostly hypothetical proteins were acquired near the base of animal-associated clades (O1-O2, Fig. 5; Table S1). On the other hand, in Saccharibacteria, originations were primarily associated with shallower subclades of animal-associated (and in one case, freshwater) genomes (O3-O6, Fig. 5; Table S1). These findings generally corresponded with the distribution of small, highly enriched modules of largely hypothetical proteins (Fig. 4) and suggest that the distribution of these modules is best explained by lineage-specific acquisition events of relatively few genes at one time, rather than large acquisition events at deeper nodes. Intriguingly, one subclade of animal-associated Saccharibacteria had the highest incidence of originations of all groups in our analysis (O6, Fig. 5; Table S1), suggesting that these genomes may be phylogenetic 'hotspots' for transfer.

While origination events were relatively infrequent in all three CPR lineages, instances of within-CPR transfer and loss were very frequent and dispersed across most interior branches of the tree (Fig. 5). Notably, we detected sporadic losses across internal branches, which is inconsistent with a major gene loss event at the time of adaptation to animal-associated habitats. Surprisingly, we noticed that genomes of non-animal associated Saccharibacteria, particularly those from the SAC1 clade, displayed substantial patterns of loss despite their relatively large proteome sizes. Thus, losses in these environmental lineages were possibly balanced by lateral transfer events over the course of evolution.

3.4 Discussion

Here, we expand sampling of genomes from the Absconditabacteria, Gracilibacteria, and Saccharibacteria, particularly from environmental biomes. The basal positioning of environmental clades in phylogenetic reconstructions provides strong support for the hypothesis that these lineages originated in the environment (Fig. 1a), and potentially migrated into humans and terrestrial animals via consumption of groundwater (He et al., 2021; McLean et al., 2020). Unlike the Absconditabacteria, which appear to have transitioned only once into animal oral cavities and guts, our phylogenetic evidence suggests that Gracilibacteria may have undergone multiple transitions into the animal microbiome in unique phylogenetic clades. In the Saccharibacteria, phylogenetically interspersed environmental and oral/gut Saccharibacteria could reflect independent migrations into the animal environment, consistent with previous work on smaller genome sets (McLean et al., 2020). Alternatively, this pattern may reflect lineage-specific reversion back to environmental niches in some clades (Fig. 1a). We also show preliminary evidence for small lineages of Gracilibacteria and Saccharibacteria that appear to have colonized the dolphin mouth separately from those that colonized the oral environments of terrestrial animals (Fig. 1a). Previous work showing the clear distinction between marine mammal microbiomes and their surrounding seawater/prey support that these CPR bacteria are likely legitimate members of the dolphin oral microbiome, rather than contamination (Bik et al., 2016).

Currently, the mechanisms that enable environmental transition among CPR are unknown. Several observations, including that CPR host-pairs may be taxonomically distinct between oral and gut habitats, raise the question of whether habitat transitions among CPR involve co-migration with their hosts or the acquisition of new hosts. The finding that single CPR species co-occur with a single Actinobacteria species, or several closely related ones, in multiple animal-associated metagenomes contributes further evidence that these associations may be flexible and phylogenetically diverse rather than highly evolutionarily conserved (Cross et al., 2019). Supporting this, some laboratory strains of oral Saccharibacteria can adapt to new hosts after periods of living independently (Bor et al., 2018). The lack of evidence for lateral gene transfer between experimentally profiled pairs (McLean et al., 2020) also suggests that some CPR-host pairs may have established fairly recently.

Host associations for Absconditabacteria, Gracilibacteria, and environmental Saccharibacteria are largely unknown. However, this is changing quickly - for example, a very recent paper demonstrated the co-culture of Saccharibacteria and two species of *Gordonia* (Actinobacteria) from wastewater foam (Batinovic et al., 2021). Similarly, changes in abundance over a sample series from a bioreactor system treating thiocyanate was recently used to suggest that *Microbacterium ginsengisoli* may serve as a host for a co-occurring Saccharibacteria (Huddy et

al., 2020). One Absconditabacteria lineage (*Vampirococcus*) has been predicted to have a host from the Gammaproteobacteria (Moreira et al., 2020) and one Gracilibacteria was suggested to have a *Colwellia* host based on a shared repeat protein motif (Sieber et al., 2019). Given the scant information about possible hosts for these CPR, especially for Absconditabacteria and Gracilibacteria, the patterns of co-occurrence we report for specific organisms provide starting points for host identification via targeted co-isolation.

To evaluate to what extent changes in gene content are associated with habitat transition, we first established core gene sets. These indicated that overall proteome size and content differed between environmental and animal-associated Saccharibacteria, and to some extent Gracilibacteria. Despite overall smaller proteome size, we identified a large number of protein families that were highly enriched among animal-associated CPR from all three lineages. The most striking capacities involve amino acid transport, oxidative stress tolerance, and viral defense, which may enable use of habitat-specific resources or tolerance of habitat-specific stressors. These findings complement previous evidence that prophages are enriched in animal-associated Saccharibacteria relative to environmental counterparts (Shaiber et al., 2020).

Only three lineages of CPR (of potentially dozens) have been consistently recovered in the animal-associated microbiome. Given the enormous diversity of CPR bacteria in drinking water (He et al., 2021), there has likely been ample opportunity for various taxa to disperse into the mouths of terrestrial animals; however, establishment and persistence of these bacteria may have been limited by the absence of a suitable host in oral and gut environments. Thus, we predict that other CPR bacteria - including those from the large Microgenomates and Parcubacteria lineages - have hosts that are infrequent or transient members of the animal microbiome, or have insufficient ability to ‘adapt’ to new hosts upon contact. For example, formation of new associations may be limited by the specificity of pili involved in host interaction or proteins involved in attachment (He et al., 2021; Luef et al., 2015; Shaiber et al., 2020).

It is also interesting to compare processes of habitat transition in CPR with those proposed for other bacteria and for archaea. Our results suggest that Saccharibacteria (and potentially Gracilibacteria) from the human/animal microbiome have smaller genome sizes than related, deeper-branching lineages of environmental origin. This pattern is also apparent for other, free-living groups adapted to the animal microbiome from the environment, like the Elusimicrobia (Méheust et al., 2020) and intracellular symbionts of insects (McCutcheon and Moran, 2011). However, in contrast to findings for Elusimicrobia, where host-associated lineages have common patterns of loss of metabolic capacities compared to relatives from non-host environments (Méheust et al., 2020), patterns of gene loss in animal-associated CPR appear to be heterogeneous and lineage-specific. One possibility is that gene loss in CPR is primarily modulated by strong dependence on host bacteria, whose capacities may vary substantially,

rather than by adaptation to the relatively stable, nutrient-rich animal habitat that likely shaped evolution of some non-CPR bacteria.

Changes in gene content could enable, facilitate, or follow habitat transitions. Our evolutionary reconstructions revealed that habitat-specific differences in gene content are more likely the product of combinations of intra-CPR transfer and loss rather than major acquisition events at time of lineage divergence. Thus, modules enriched in specific lineages were probably acquired via lateral transfer after habitat transition, suggesting that proteome remodeling has been continuous in CPR over evolutionary time. As such, the processes shaping CPR lineage evolution share both similarities and differences with those predicted for other microbes, including Haloarchaeota (Martijn et al., 2020) and ammonia-oxidizing lineages of Thaumarchaeota (Abby et al., 2020; Sheridan et al., 2020), where both large lateral transfer events and gradual patterns of gene loss, gain, and duplication worked together to shape major habitat transitions.

Conclusion

Overall, our findings highlight factors associated with habitat transitions in three CPR lineages that occur in both human/animal and environmental microbiomes. We expand the evidence for niche-based differences in protein content (McLean et al., 2020; Shaiber et al., 2020) and identify a large set of protein families that could guide future studies of CPR symbiosis. Furthermore, patterns of co-occurrence may inform experiments aiming to co-cultivate CPR and their hosts. Our analyses point to a history of continuous genome remodeling accompanying transition into human/animal habitats, rather than rapid gene gain/loss around the time of habitat switches. Thus, habitat transitions in CPR may have been primarily driven by the availability of suitable hosts and reinforced by acquisition and/or loss of genetic capacities. These processes may be distinct from those shaping transitions in other bacteria and archaea that are not obligate symbionts of other microorganisms.

3.5 Figures

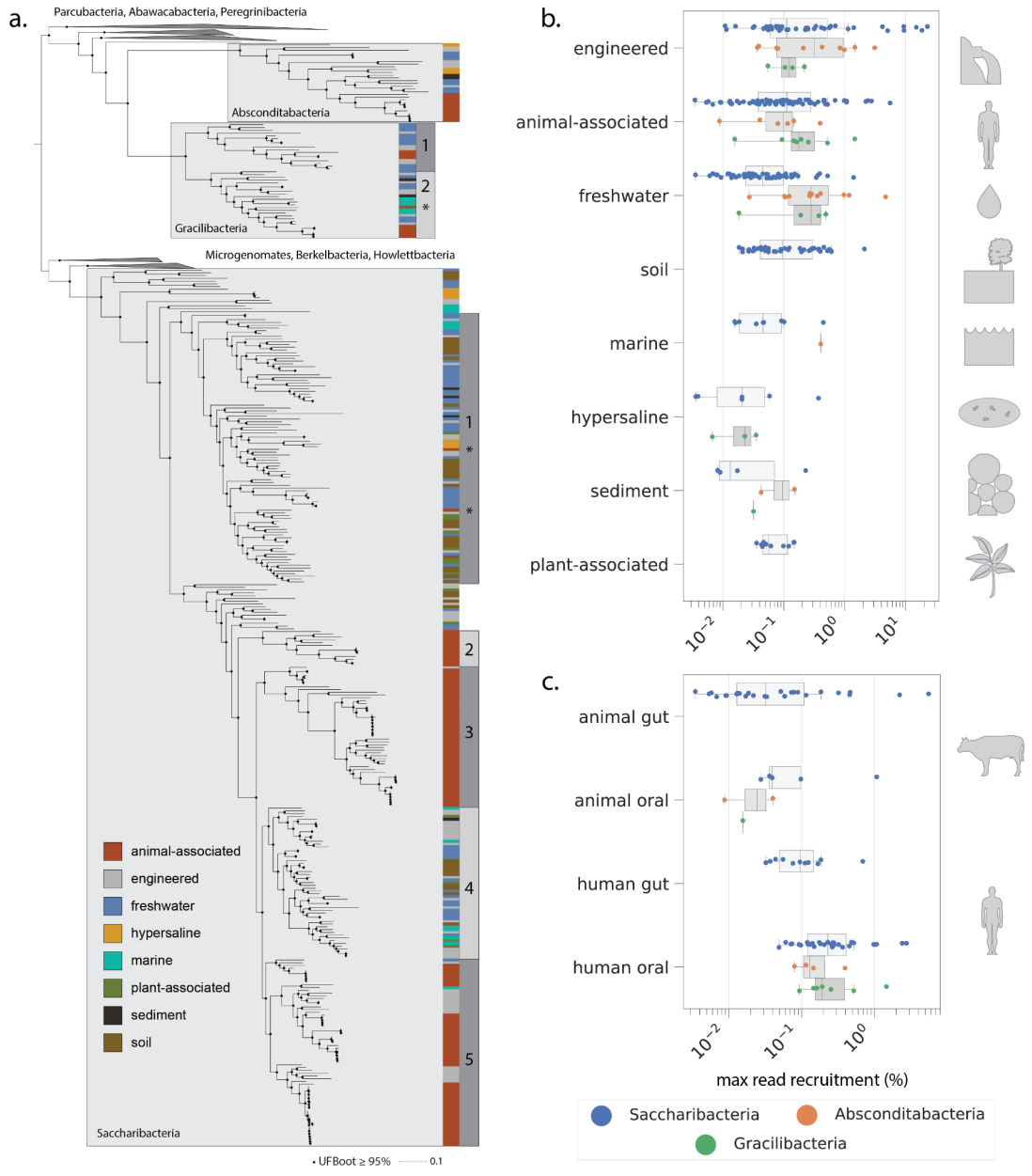

Figure 1. Phylogenetic and environmental patterns for the Absconditabacteria, Gracilibacteria, and Saccharibacteria. a) Maximum-likelihood tree based on 16 concatenated ribosomal proteins (1976 amino acids, LG+R10 model). Scale bar represents the average number of substitutions per site. Habitat of origin and phylogenetic subclade (where applicable) for each genome are indicated to the right of the tree. Asterisks indicate phylogenetic position of a subset of organisms derived from dolphin mouth metagenomes. Percentage of reads per metagenomic sample mapping to individual genomes across b) environments and **c)** body sites of humans and animals.

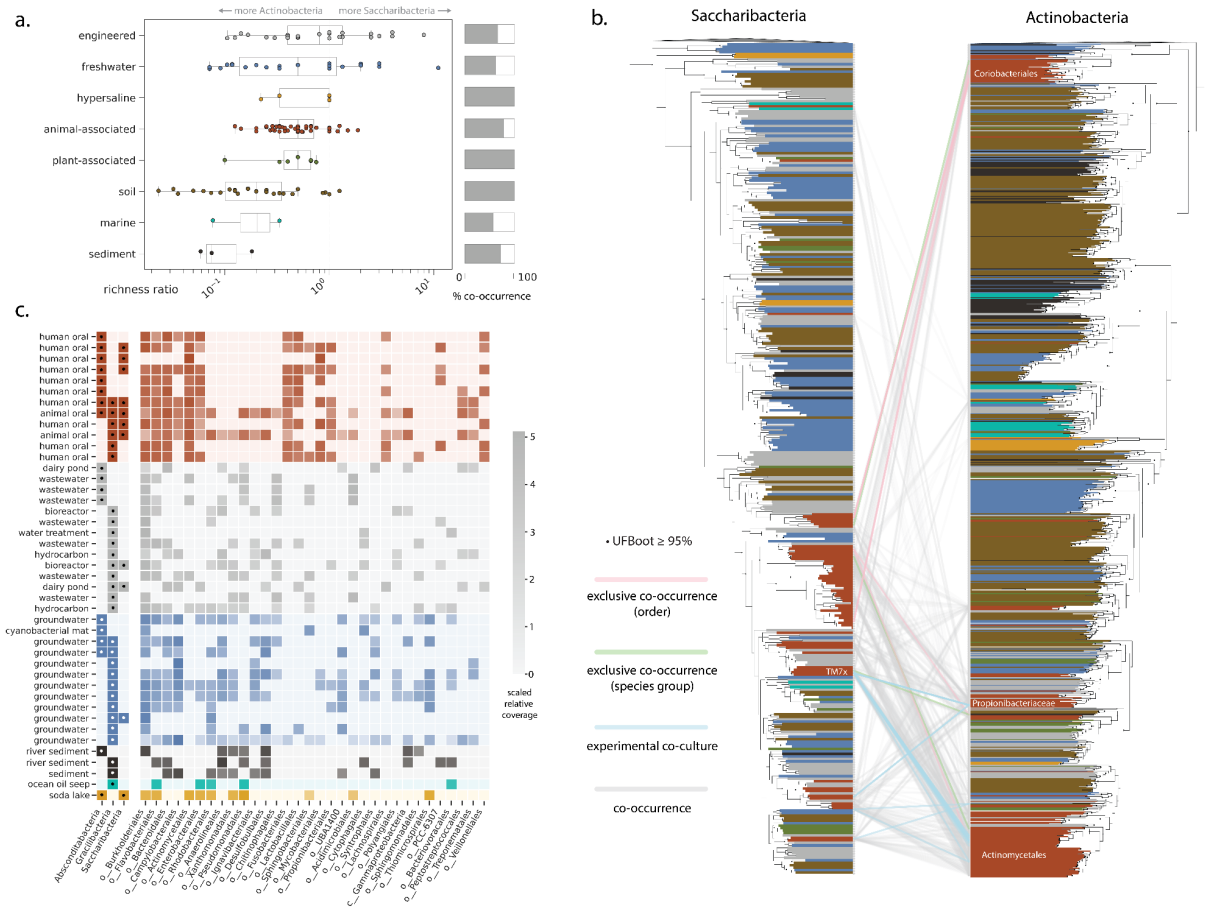

Figure 2. Patterns of co-occurrence between CPR and potential host lineages across environments. a) Relative richness ratio, describing the ratio of distinct Saccharibacteria species groups to Actinobacteria species groups, for each sample and overall co-occurrence percentage across habitat categories. b) Maximum-likelihood trees for Saccharibacteria and Actinobacteria based on ribosomal protein S3 sequences extracted from all source metagenomes. Co-occurrence patterns are shown only for species groups derived from animal-associated metagenomes. c) Community composition (using GTDB taxonomy for non-CPR bacteria) for metagenomic samples containing Absconditabacteria and Gracilibacteria. Cells with dots indicate only presence, whereas those without dots convey log-scaled, normalized relative coverage information. Only potential host lineages present in 8 or more samples are shown.

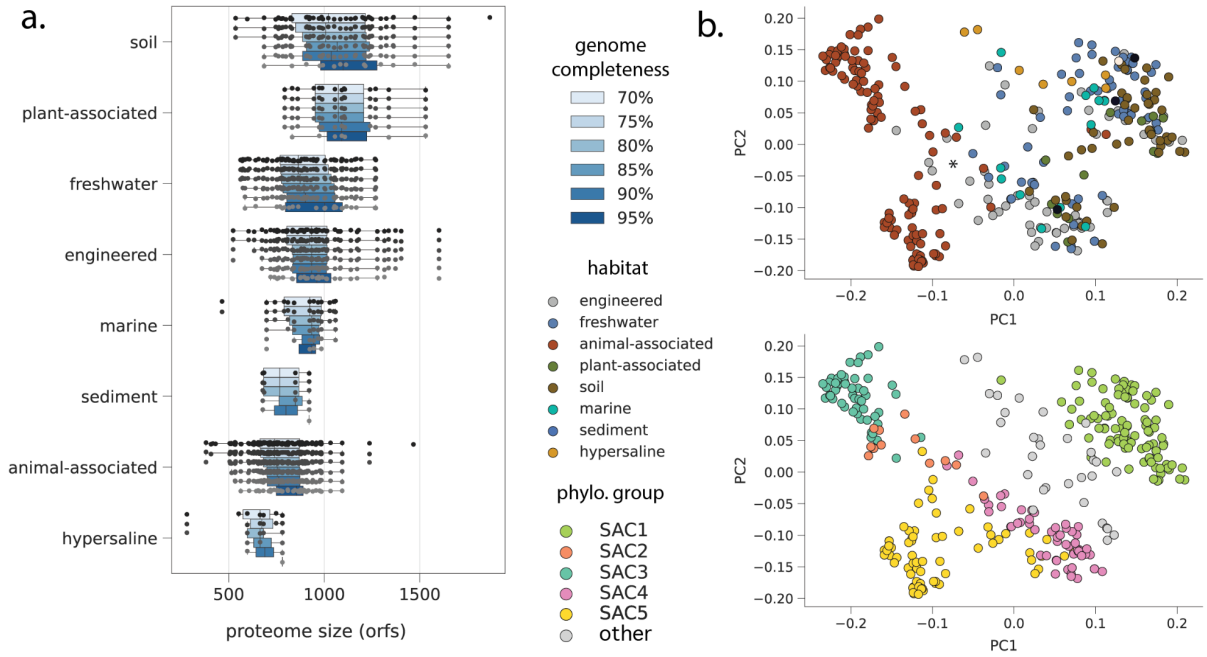

Figure 3. Proteome characteristics for Saccharibacteria. a) Predicted proteome size (open reading frame count) at increasing genome completeness thresholds. b) Overall proteome similarity among Saccharibacteria from different habitat categories (top panel) and phylogenetic clades (bottom panel). PCoAs were computed from presence/absence profiles of all protein clusters with 5 or more member sequences. The primary (PC1) and secondary (PC2) principal coordinates explained 12% and 8% of variance, respectively.

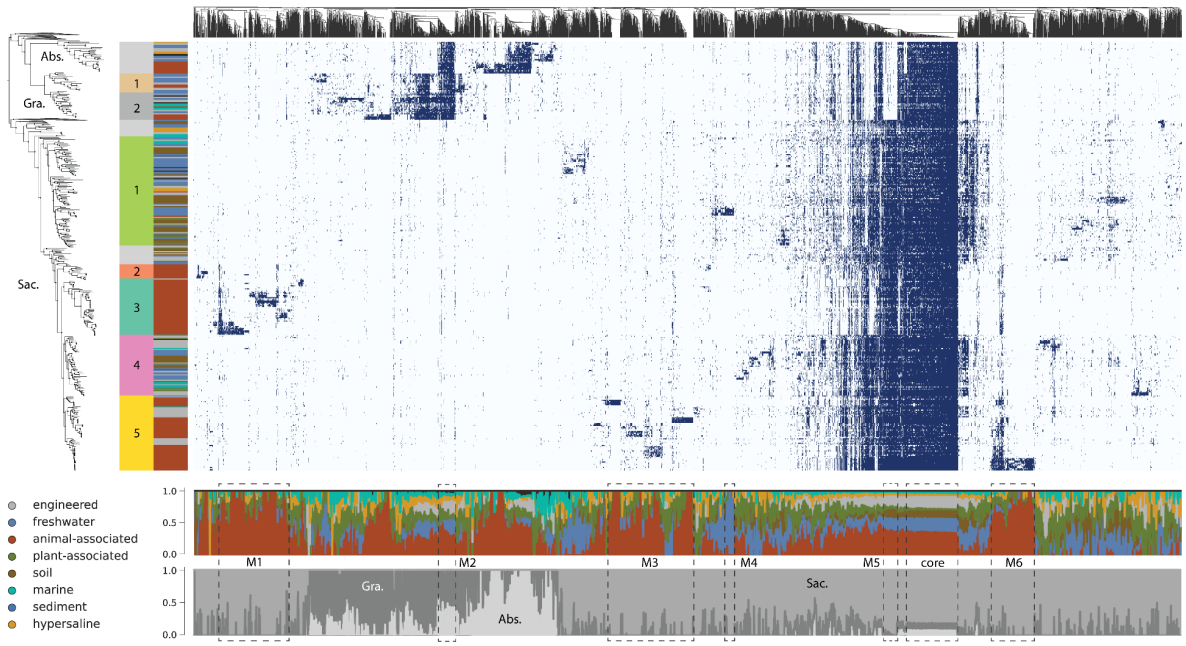

Figure 4. Phylogenetic and environmental distribution of protein families recovered among CPR. Upper panel: Presence/absence profiles for protein families with 5 or more members, with shaded cells indicating presence, and light cells indicating absence. Columns represent protein families, hierarchically clustered by similarity in distribution across the genome set. Rows correspond to genomes, ordered by their phylogenetic position in the species tree (left). Abbreviations: Abs., Absconditabacteria; Gra., Gracilibacteria; Sac., Saccharibacteria. Lower panels: Percentage of genomes encoding individual protein families that belonged to broad habitat groups (top) or taxonomic groups (bottom). Modules of protein families indicated in the text are represented by dotted lines (M1-6 and 'core').

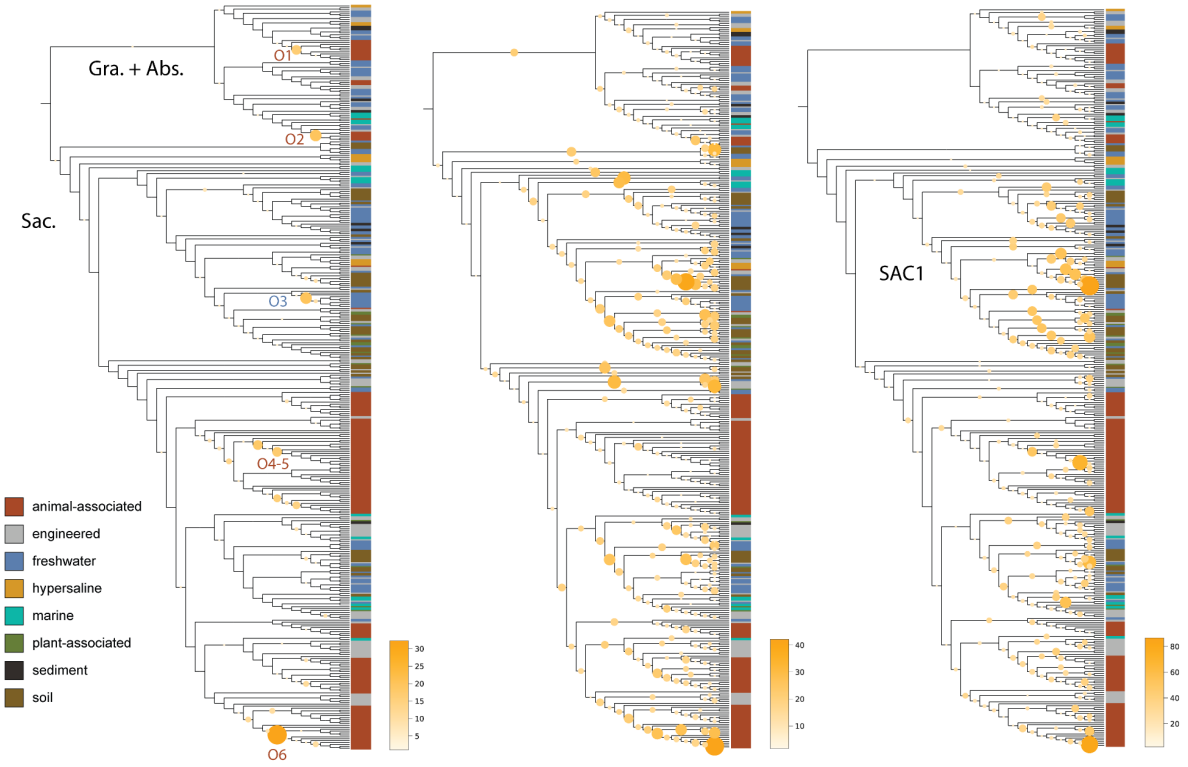

Figure 5. Evolutionary processes shaping proteome evolution in three lineages of CPR bacteria. Each panel displays the species tree from Fig. 1a in cladogram format. The size and color of circles mapped onto interior branches represent the cumulative number of a) originations (defined as either lateral transfer from outside the lineages examined here, or *de novo* evolution) b) transfer among the three CPR lineages included here and c) genomic losses predicted to occur on that branch for all 902 differentially distributed families where gene-species tree reconciliation was possible. Abbreviations: Abs., Absconditabacteria; Gra., Gracilibacteria; Sac., Saccharibacteria. SAC1 indicates a monophyletic clade of Saccharibacteria referenced in the text.

4. Variable impact of geochemical gradients on the functional potential of bacteria and archaea from the permanently stratified Lac Pavin

Alexander L. Jaffe, Cindy J. Castelle, Corinne Bardot, Hermine Billard, Jonathan Colombet, Fanny Perrière, Anne-Hélène Le Jeune, Anne-Catherine Lehours, and Jillian F. Banfield

Unpublished, 2021.

Permanently stratified lakes contain diverse microbial communities that vary with depth, and so can serve as useful models for studying the relationships between microbial community structure and geochemistry. Recent work has shown that stratified lakes can harbor numerous bacteria and archaea from novel lineages, including those from the Candidate Phyla Radiation (CPR). However, the extent to which geochemical stratification differentially impacts metabolic potential in these lineages compared to other organisms has not been defined. Here, we recover nearly 750 draft genomes from microbial communities in Lac Pavin, a deep, stratified lake in central France, and determine the distribution of lineages and functions across the lake's oxygen gradient. Analysis of functional gene content revealed an enrichment of microbial rhodopsins and electron transport chain components in the shallower interface zone, suggesting that light as well as oxygen plays an important role in structuring metabolic capacities in non-CPR organisms. We also find evidence for spatial segregation of some biogeochemical processes, including methane production in the sediments and methane consumption in the water column. In contrast, we detected CPR bacteria throughout the water column that encode form III-related RuBisCO that could contribute to CO_2 fixation. Overall, CPR bacteria do not show strong signals of metabolic differentiation by depth, despite the presence of distinct sets of CPR organisms in different lake compartments. This suggests that environmental gradients in Lac Pavin probably select for the capacities of non-CPR and CPR bacteria to differing extents, possibly because the selection on CPR is indirect and depends primarily on the characteristics of their host cells.

4.1 Introduction

Meromictic lakes are permanently stratified aquatic ecosystems with a global distribution (Hall and Northcote, 2012). Despite variation in size and morphology, these lakes generally have several common physical features, including the presence of two major layers - a shallower, oxygenated layer called the mixolimnion, and a deeper, generally anoxic zone called the monimolimnion (Boehrer et al., 2017). These layers are separated by an oxycline of variable steepness that occurs in the water column, forming an 'interface' zone. A unique feature of meromictic lakes, compared to other similar bodies of water, is that mixing events between the two layers are rare, with seasonal mixing observed primarily within the oxygenated upper layer (Boehrer et al., 2017). Based primarily on marker gene analyses, these compartments are known to host distinct and diverse microbial communities that co-vary with gradients in oxygen, light, and sulfur (Andrei et al., 2015; Lehours et al., 2007), and thus can serve as model systems for probing the relationships between community structure and geochemistry.

While the water columns of meromictic lakes are often dominated by members of well-described bacterial phyla, previous work has shown that these ecosystems can host organisms from other, more divergent lineages, including the Candidate Phyla Radiation (CPR) bacteria and sibling lineages of the Cyanobacteria (Borrel et al., 2010; Peura et al., 2012; Tran et al., 2019), none of which have representatives in pure culture. Candidate Phyla Radiation bacteria are typically ultra-small celled and are predicted to live as obligate episymbionts or parasites of other microorganisms (Brown et al., 2015; Castelle and Banfield, 2018; Moreira et al., 2021). Despite their ubiquity in the microbiomes of diverse aquatic ecosystems, ecological roles of CPR bacteria and the extent to which their gene content is shaped by environmental factors remain poorly understood. These questions could be addressed in part using genome-resolved metagenomics in systems with strong geochemical gradients.

Here, we examine microbial community composition and function in Lac Pavin, a meromictic lake in central France where CPR bacteria are known to attain relatively high diversity and abundance (Borrel et al., 2010). We performed genome-resolved metagenomics to obtain hundreds of draft and high quality genomes for members of diverse lineages of bacteria and archaea, and use these genomes to examine changes in community composition and function across lake depth and oxygen concentration. Together, our analyses shed light on key processes mediated by Lac Pavin's microbial communities and provide insights into how these conditions differentially shape metabolic potential in both CPR bacteria and other lineages.

4.2 Materials and Methods

Sample collection and measurements of oxygen

Prior to sampling, oxygen concentration and saturation were measured through the water column with a SDOT dissolved oxygen data logger (NKE Instrumentation). All water column samples were collected using a 12 liter Niskin bottle. Between 3 and 5 bottles were taken for each water column sample (36-60 liters total) and all water from each depth was pooled and filtered by tangential flow filtration (0.2 micron cartridge) to yield ~1 liter of concentrate per depth. The 12 meter and 57 meter samples from 2019 were pre-filtered before tangential flow filtration with a 20 µm filtration tissue. Filtrate from the 2018 70 meter sample was subsequently re-filtered using a 30 kilodalton cartridge to yield a small size fraction concentrate (sample 70S). Filtrate from each sample was then split into four fractions, each of which was subjected to centrifugation at 18,000 rpm (30 minutes, 4˚C). Supernatant was removed and pellets from each fraction from the same sample were resuspended in TE buffer and pooled. The pooled samples were centrifuged one final time as described above and stored at -20˚C until DNA extraction. In 2019, a sediment core was taken and 5 strata were collected: 3 to 5 cm (SED1), 10 to 13 cm (SED2), 20 to 24 cm (SED3), 32 to 36 cm (SED4), and 54 to 58 cm (SED5). Each stratum was homogenized and stored at 4˚C for DNA extraction.

DNA extraction and sequencing

Water column samples were defrosted and incubated with 25 µl of lysozyme solution (50 mg/ml) for 30 minutes at 37˚C. Next, 30 µl of SDS (10%), 3 µl of proteinase K (20 mg/ml), and 5 µl of RNase A (1 mg/ml) were added and incubated for an additional hour at 37˚C. Finally, 100 µl of NaCL (5 M) and 80 µl of CTAB/NaCL were added and incubated at 65˚C for 10 minutes. To extract genomic DNA, 800 µl of phenol/chloroform/isoamyl alcohol was added to the solution and shaken vigorously by hand before centrifugation at 14,000 rpm for 30 minutes. Isopropanol precipitation was then performed and centrifuged (14,000 rpm, 30 mins) after resting. Finally, extracted DNA was washed with 500 µl of ethanol and centrifuged again before drying and resuspension in TE buffer. DNA quantification was performed using Qubit. Sediment samples were defrosted and centrifuged for 30 seconds at 10,000g to remove the supernatant. For the lysis step, samples were heated to 70˚C and then vortexed for 10 minutes. DNA was then extracted using the Qiagen DNEasy Power soil kit following manufacturer instructions.

Library preparation and metagenomic sequencing were performed at the QB3 (University of California, Berkeley) Functional Genomics Laboratory / Vincent J. Coates Genomics Sequencing Laboratory. Libraries were sequenced with 150 base-pair, paired-end reads on either a Illumina

HiSeq 4000 (2018 and 2019 samples) or a NovaSeq S4 with a target depth of 15-25 gigabasepairs per sample.

Read-based diversity analyses of the microbial community

Phylum-level microbial community composition of each sample was determined from the entire set of metagenomic reads using GraftM (Boyd et al., 2018) and a custom gpackage built from ribosomal protein S3 (rpS3) sequences from GTDB (release 05-RS95) (Crits-Christoph et al., 2021). The numbers of reads mapping to the database of rpS3 sequences were summed by phylum, and then used to determine relative abundance by dividing by the total number of mapped reads across all phyla. To examine overall similarity in community composition across samples, the relative abundance metrics computed above were subsequently used to generate a Bray-Curtis distance matrix (skbio's diversity package in Python) and subjected to a principal coordinates analysis (implemented in skbio's stats module).

Metagenomic assembly, binning, and bin curation

Metagenomic assembly was performed following the procedures of (He et al., 2021). Briefly, sequencing reads were rid of Illumina adapters using BBTools and were trimmed using Sickle (Joshi et al., 2011) (default thresholds). Quality-filtered reads were then assembled using MEGAHIT (v. 1.2.9, default parameters) (Li et al., 2015) and scaffolded using IDBA-UD (Peng et al., 2012). Reads were mapped back to the entire assembly for each sample using bowtie2 (Langmead and Salzberg, 2012) to compute scaffold coverage values. Assembled scaffolds were subsetted to those ≥1000 bp for gene prediction and binning. Genes were predicted using Prodigal (*meta* mode) (Hyatt et al., 2010) and predicted proteins were annotated using USEARCH against the KEGG, UniRef, and UniProt databases.

Genome binning was performed using both a manual and an automated approach. For the manual approach, scaffold information was loaded into ggKbase (ggkbase.berkeley.edu) and scaffolds were binned on the basis of coverage, GC content, taxonomic affiliation, and inventories of bacterial 'single copy' genes. For the automated approach, reads were cross-mapped against every other assembly within each biome type using bowtie2. Sediment and water column samples were binned separately. Coverage tables were generated using the jgi_summarize_bam_contig_depths script (bitbucket.org/berkeleylab/metabat/src/master) and passed to MetaBAT2 (minimum contig size 1500 bp) (Kang et al., n.d.) for automated bin generation. Bins derived from the manual and automated approaches were reconciled with DAS Tool (Sieber et al., 2018) to create the best merged set of bins for downstream analysis.

To prepare for the bin refinement step, preliminary taxonomic classifications were performed using GTDB-Tk (Chaumeil et al., 2019) and classifications refined to align with standard

taxonomies. Bins associated with the Candidate Phyla Radiation bacteria or DPANN archaea were separated and profiled for reduced sets of marker genes sensitive to lineage-specific losses in these groups (Anantharaman et al., 2016). Completeness and redundancy were calculated as the percentage of marker genes present and duplicated, respectively. These quality metrics were combined with those for all other, non-CPR and non-DPANN bins estimated by CheckM (Parks et al., 2015). Bins were then filtered to those with ≥70 completeness and de-replicated at 95% average nucleotide identity (ANI) using dRep to create a secondary set for manual curation. All quality-filtered bins were loaded into Anvi'o (Eren et al., 2015) and visualized individually using the *anvi-refine* command. Bins were refined by removing sets of scaffolds with aberrant coverage profiles across all cross-mapped samples. Completeness and redundancy metrics were also considered. Refined bins were then re-assessed for quality (as above) and additionally profiled with GUNC (Orakov et al., 2021) as an alternative metric of contamination.

Relative abundance calculations for bins

Reads from all samples were re-mapped to the refined bin set using bowtie2. Bin coverage was analyzed using CoverM (https://github.com/wwood/CoverM) with a 95% identity threshold for read mapping and 50% breadth threshold (fraction of genome covered by reads). Relative coverage values for each genome were computed by dividing its mean coverage by the total, summed coverage of all genomes present within a given sample.

KEGG enrichment analysis

Proteins were re-predicted for all refined bins using Prodigal (*single* mode). All proteins that were not from Candidate Phyla Radiation bacteria were subjected to annotation with kofamscan (Aramaki et al., 2020). Protein annotations were filtered first by e-value (≤1e-6) and the highest scoring hit per protein was taken. Presence/absence of each KEGG orthology (KO) was computed per genome and visualized using the clustergram function in seaborn. To determine statistically enriched functions by lake compartment, we drew on a previously developed computational pipeline that tests the frequency of occurrence of a function/protein cluster within genomes from a given lake compartment versus those from outside of that compartment (Jaffe et al., 2021). To determine the degree of enrichment, a ratio describing the frequency of occurrence within genomes from a given compartment relative to the frequency of occurrence in genomes from outside that compartment was computed, and statistical significance assigned using the Fisher's Exact Test (the fisher_exact function in scipy.stats, alternative='two-sided'). In cases where no genomes outside of a given compartment encoded the function of interest, the function was labelled 'exclusive' and no ratio was computed. All comparisons were corrected for false discovery rate using the multipletests function in statsmodels.stats.multitest (method="fdr_bh").

CPR phylogenomics and gene content analyses

Draft genomes from the CPR bacteria were further classified using a previously published phylogenetic framework for this group (Jaffe et al., 2020). Specifically, 16 phylogenetically informative, syntenic ribosomal proteins (Hug et al., 2013) were extracted from each genome (where present) and combined with a comprehensive set of reference sequences from the above study. Each marker protein was aligned individually using MAFFT (Katoh and Standley, 2013) and trimmed using BMGE (*-m BLOSUM30*) (Criscuolo and Gribaldo, 2010). All individual protein alignments were concatenated to generate a supermatrix. Any organisms for which less than 7 of the 16 marker proteins were present were discarded. Finally, a maximum-likelihood phylogenetic tree was inferred using IQTREE (*-m TEST --b 1000*) (Minh et al., 2020). Taxonomic assignments of newly-recovered genomes from Lac Pavin were manually curated using the tree topology and taxonomy of proximal reference sequences.

Gene content of CPR bacteria was analyzed using a two-step *de novo* protein clustering pipeline (Méheust et al., 2019). First, proteins were clustered into 'subfamilies' and overlapping subfamilies were merged (*–coverage 0.75*). For sequences in each protein cluster, the most common KEGG annotation was determined alongside the frequency of this annotation. Clusters were visually arrayed in a heatmap and subjected to enrichment analysis using the statistical pipeline described above for functional enrichment.

Analysis of RuBisCO diversity and abundance

To determine the diversity of RuBisCO genes across the water column and sediments, all metagenomic assemblies were probed with a set of Hidden Markov Models (HMMs) describing the form I, II, II, II/II, and III RuBisCO respectively (Anantharaman et al., 2016). Stringent score cutoffs originally published with the models were applied and the best hit was selected for each individual protein. Scaffolds encoding above-threshold hits were gathered and clustered across samples using dRep at 95% ANI (*cluster -p 20 -pa 0.80 -sa 0.95*) (Olm et al., 2017). A non-redundant, representative set of scaffolds encoding RuBisCO was created by selecting the set of cluster centroids that were binned (i.e., belonged to a draft genome) and/or were the longest. Reads from all samples were re-mapped to this unique set and mean coverage per sample was computed using CoverM, as described above. Only scaffolds with ≥50% breadth were considered 'present' in a given sample and retained for diversity analysis.

Next, putative taxonomy was assigned to all scaffolds encoding RuBisCO using a combination of bin taxonomy (for binned fragments), phylogenetic affiliation of co-encoded proteins, and manual curation. Sequence fragments clearly originating from eukaryotic organisms, such as algae and diatoms, were removed. Classification of newly-recovered RuBisCO sequences was obtained by placing them within a reference RuBisCO large subunit protein phylogeny (Jaffe et

al., 2019). Finally, the relative coverage of each representative scaffold encoding RuBisCO within each sample was computed by dividing mean coverage by the total coverage across all scaffolds. Relative coverage was first summed by phylum-level taxonomy of the encoding scaffold and then averaged across samples within the same lake compartment.

4.3 Results

Diversity of microbial communities across Lac Pavin's water column and sediments

To examine microbial communities in Lac Pavin, we sampled the water column and shallow sediments of Lac Pavin annually from 2017-2019 (Materials and Methods). The first of year sampling focused on the lake's oxycline, which is an 'interface' between the relatively oxygenated mixolimnion (0-40 meters) and the permanently anoxic monimolimnion (~60-90 meters) (Fig. 1b). Subsequent sampling targeted microbial communities in the anoxic zone and the shallow sediments from 3-58 centimeters (Fig. 1bc). We also took one sample from the shallow oxygen maximum (12 meters) as a point of comparison. Water column samples were filtered to the 0.2 - 20 μm size fraction, and whole-community genomic DNA was extracted and sequenced as described in the Materials and Methods. In addition, one post 0.2 μm fraction was used to extend genomic sampling.

First, we surveyed metagenomes for all reads mapping to a ribosomal protein S3 (rps3) database, assigned the reads to phylum-level groups (Materials and Methods), then examined microbial community similarity across sampling depths/years. A principal coordinates analysis (PCoA) suggested that samples from the same lake compartment, e.g. the samples from the anoxic zone, showed highest similarity to each other across years and were distinct from samples from other compartments, consistent with the permanently stratified nature of the lake (Fig. 1c). The community composition of the shallow oxygen maximum clusters with samples from the lake interface (Fig. 1c), despite its relatively high oxygen concentration.

We next investigated the specific phylogenetic groups responsible for driving dissimilarity between microbial communities in the lake. Microbial communities from the water column interface were highly enriched in members of the Actinobacteria relative to other samples, consistent with findings from other lakes with pronounced oxyclines (İnceoğlu et al., 2015; Tran et al., 2019) (Fig. 1d). The shallow regions of the lake also displayed higher abundances of Proteobacteria and Cyanobacteria compared to the anoxic zone of the water column, where members of the Desulfobacterota, Bacteroidetes, and Omnitrophota were particularly abundant (Fig. 1d). Finally, the shallow sediments were enriched for members of the Chloroflexi, Caldisericota, DPANN archaea, and a variety of methanogens (some of which were also found in

the anoxic zone) (Fig. 1d). The shallow sediments also harbored levels of Cyanobacteria comparable to those in the shallower water column, despite the lack of light (Fig. 1d). Notably, members of the Candidate Phyla Radiation were found across the lake's oxygen gradient and shallow sediments, but were barely detected in the region of the oxygen maximum (Fig. 1d).

Recovery of draft quality bacterial and archaeal genomes from Lac Pavin

To more closely examine the phylogenetics and metabolic potential of key community members, including CPR bacteria, we generated metagenomic bins by grouping assembled scaffolds into draft genomes with both manual and automated approaches (Materials and Methods). The resulting bins were de-replicated at the species level (95% ANI) and manually curated using Anvi'o (Materials and Methods). Manual curation improved metrics of redundancy or genome chimerism (as measured by CheckM and GUNC, respectively) in nearly 70% percent of curated bins. Refined bins were then filtered again for quality (≥70% completeness) to yield a final set of 738 non-redundant, draft- or better quality genomes for downstream analysis. Taxonomy assignment via multiple methods indicated that this set included genome information from most major groups identified by the marker gene analysis, including 106 draft genomes from the CPR bacteria, 574 from non-CPR bacteria, and 58 from archaea.

Changes in bacterial and archaeal gene content over an oxygen gradient

We next asked the extent to which changes in community composition across depth in Lac Pavin are accompanied by changes in metabolic potential. We annotated the set of curated draft genomes described above with the full KEGG database and visualized the presence/absence of each genetically-encoded function across the set of genomes detected in each lake compartment and combinations thereof (Figure 2a). While a large group of functions (hereafter, a 'module') were shared across most genomes (region A, Figure 2a), primarily corresponding to essential cellular functions, many modules with narrower distributions were also observable. At least one module appeared to be specific to genomes detected in the lake's sediment (region B, Figure 2a). Several other modules were associated primarily with genomes detected in the interface (multiple modules contained within region C in Figure 2a). We quantified the enrichment of each function within groups of genomes from one ecosystem compartment relative to genomes from other compartments (e.g., the anoxic zone compared to the interface and sediments; Materials and Methods). We identified approximately 1,500 KEGG orthologies that were moderately to highly enriched (≥3 enrichment ratio, p-value <0.05 corrected for multiple comparisons) in genomes from any given lake compartment. Most enriched functions were associated with genomes from the interface between the oxic and anoxic zones (987 enriched families). This finding suggests that, at a high level, metabolic potential changes in tandem with community composition over depth.

We next identified the most statistically enriched genes in any lake compartment. Among these genes were three subunits of cytochrome c oxidase (aerobic metabolism) which were strongly enriched in diverse bacterial genomes from the lake interface (enrichment ratio ~4, fdr < 1e-40), consistent with the oxygenated/microaerophilic conditions of this zone (Fig. 1b). Intriguingly, these genes were not entirely absent from genomes from the anoxic zone and sediment, where oxygen levels are essentially zero (Fig. 2b). Presence of the cytochrome c oxidase in sediment-associated Actinobacteria, Cyanobacteria, and Proteobacteria may represent inactive organisms transported downwards from above or organisms that are facultatively aerobic.

An additional set of highly enriched functions were involved with phototrophy. Specifically, we observed that microbial rhodopsins, a diverse family of light-sensitive ion transporters/sensors (Rozenberg et al., 2021), were highly enriched in various Actinobacteria primarily found in the shallow and interface zones of the lake (Fig. 2c). Rhodopsins were extremely rare in genomes found only in the anoxic zone and sediments, where light intensity is likely very low (Fig. 2c). Similarly, numerous genes related to photosystems I and II were enriched in 17 diverse Cyanobacteria - including members of the Chroococcales, Oscillatoriophycideae, and Synechococcales - almost exclusively found in the shallow and interface zones of the lake (Fig. 2d). Together, these findings point to light as a potential structuring force for not only microbial community composition but also gene content for various bacterial groups in Lac Pavin. Additionally, putative rhodopsins were found in several CPR bacterial genomes; however, they were not statistically enriched in any specific lake compartment.

We also observed that genes associated with methane metabolism (methyl-coenzyme reductase, alpha and beta subunits, *mcr*) were highly enriched in genomes from the lake sediments (Fig. 2e), but also present in a few genomes from the anoxic zone. Specifically, subunits of *mcr* complex were encoded by ten genomes mostly from members of the *Methanosaeta*, *Methanoregula,* and the Methanomicrobiales, implicating them in methane production rather than methane oxidation via reverse methanogenesis. Some of these taxa were previously detected in the lake using amplicon approaches (Borrel et al., 2012; Lehours et al., 2007). We also examined the distribution of genes associated with biological methane oxidation, which is thought to act as the major 'sink' for methane in Lac Pavin (Lopes et al., 2011). These genes were found throughout the water column (Fig. 2f), but primarily in the interface zone encoded by organisms related to the methylotrophic genus *Methylobacter*. The detection of some organisms with the ability to aerobically oxidize methane in the anoxic zone suggests that this process may be occurring under microaerophilic conditions, as has been previously suggested (Biderre-Petit et al., 2011; Lopes et al., 2011).

Diversity and distribution of CPR bacteria in Lac Pavin

Previous work based on amplicon approaches showed that Lac Pavin hosts diverse communities of bacteria from groups now recognized to be part of the CPR, and these communities vary with

depth (Borrel et al., 2010). Here, we draw on assembled genome sequences to examine the phylogenetic diversity of CPR bacteria in Lac Pavin. Our phylogenetic analysis, drawing on a concatenated set of ribosomal proteins (Materials and Methods) revealed that CPR bacteria from Lac Pavin are highly diverse and include members from all sub-radiations recognized within the CPR, including the Absconditabacteria, Saccharibacteria, Peregrinibacteria, Microgenomates and recently delineated clades of the Parcubacteria (Jaffe et al., 2020), each of which each contains numerous sub-lineages (Fig. 3a).

By mapping metagenomic reads back to the draft CPR genomes, we determined the abundance of individual CPR organisms across the lake's oxygen gradient. While CPR bacteria were detectable throughout the water column and shallow sediments, they were most abundant in the anoxic zone. Intriguingly, different species from the same lineage populated each zone (e.g. different species of Berkelbacteria were uniquely associated with the interface, anoxic zone and sediments; Fig. 3b). In contrast, some lineages were only detected in a single compartment - for example, members of the Saccharibacteria were restricted to the lake interface zone (Fig. 3b).

Patterns of gene content in CPR bacteria

We examined how gene content of CPR bacteria varies throughout Lac Pavin's water column. Using a protein clustering approach that is agnostic to gene function (Méheust et al., 2019), we resolved 4405 protein families (≥ 5 proteins in size) among all CPR bacteria and visualized the distribution of each family within the draft genome set. Clustering based on protein family presence/absence revealed a large 'module' (or group of protein families) that were found in nearly all CPR genomes (Fig. 4a). As for non-CPR bacteria, this 'core' module generally contains functions such as transcription, translation, and cell division. Outside of this core module, most other protein families were patchily distributed (Fig. 4a).

While there was some evidence for groups of protein families that are specific to a small number of CPR genomes (narrow, horizontal strips in Fig. 4a), no large modules are associated with organisms from the interface, anoxic zone, sediment or combinations thereof. This suggests that in general, the proteomes of CPR organisms across the water column are not highly differentiated from one another. However, more detailed statistical analysis revealed 35 protein families that are moderately to highly enriched in a specific compartment (enrichment ratio ≥3, fdr <0.05). Most of these families (~75%) were enriched in CPR bacteria from the interface, and included protein families that were involved in the metabolism of malate (K00027), zinc transport (K09815), and metabolism of certain amino acids (K00928). Also notable was a family confidently annotated as an Fe-Mn superoxide dismutase that was moderately enriched (enrichment ratio ~2.7) in CPR bacteria from the interface relative to those from the anoxic zone and sediments. The few families that were statistically enriched in CPR bacteria from the anoxic zone or sediments were mostly un-annotated or poorly annotated. Overall, these results suggest

that while organisms from the interface may have slight differences in their metabolic capacities and stress responses compared to those from the anoxic zone and sediments, gene content in CPR bacteria is largely similar across the water column of Lac Pavin.

CPR bacteria with the metabolic potential to fix carbon via RuBisCO

Given prior research demonstrating the diversity of RuBisCO enzymes among some CPR bacteria (Jaffe et al., 2019; Wrighton et al., 2012, 2016), we hypothesized that some CPR may play a role in carbon fixation in Lac Pavin. HMM searches and subsequent phylogenetic analysis of the RuBisCO large chain sequence revealed two distinct forms of RuBisCO among CPR bacteria in Lac Pavin. Specifically, we identified form II/III RuBisCO in members of the Absconditabacteria and the Komeilibacteria (Parcubacteria radiation) and form III-c in another Komeilibacteria (Fig. 5bcd), a lineage in which the form III-c has not previously been observed. Both forms of RuBisCO likely function by incorporating carbon dioxide with a free nucleotide to form glyceraldehyde-3-phosphate, a glycolytic intermediate (Sato et al., 2007).

We next quantified the relative abundance of CPR RuBisCOs in comparison to other RuBisCO genes encoded by other bacteria and archaea, including those found both in genomic bins as well as in unbinned scaffolds. This analysis revealed that CPR lineages constitute two of five phyla encoding varying forms of RuBisCO in the lake interface and anoxic zone. In these zones, Cyanobacteria, Actinobacteria, and Proteobacteria encode form I or form II genes that likely fix CO_2 via the Calvin Benson Bassham (CBB) cycle (Fig. 5cd). In the interface, Form II/III from the Absconditabacteria composed on average 13.6% (ranging from ~10-17%) of overall coverage of scaffolds with RuBisCO, whereas form II/III from the Komeilibacteria contributed only ~2.2% of overall coverage (Fig. 5c). However, in the anoxic zone, Komeilibacteria accounted for an average of 32.6% of RuBisCO gene coverage, yet this varied dramatically, ranging from 1% to 83% across samples (Fig. 5d). Interestingly, within the anoxic zone just above the sediment-water interface (90 meter sample), RuBisCO from Komeilibacteria was the largest contributor to RuBisCO metabolic potential. In contrast, CPR bacteria with RuBisCO were barely detectable in the lake sediments, where community potential for carbon fixation via RuBisCO was dominated by archaea bearing the Form III-b enzyme (combined 67.9% of overall RuBisCO coverage). There, among the most abundant genes were those encoded by various lineages within the DPANN archaea, including the Pacearchaeota and Diapherotrites (Fig. 5e).

4.4 Discussion

Despite differences in morphology and origin, meromictic lakes generally have similar physical characteristics that shape their microbial communities (Baatar et al., 2016). Our results suggest that like other similar lakes, the availability of oxygen across the water column of Lac Pavin shapes both microbial community composition and their metabolic capacities. Interestingly, we found very little if any evidence for the photosynthetic purple/green sulfur bacteria that frequently occur in the oxycline of meromictic lakes, sometimes at extremely high densities (Hamilton et al., 2014; Meyer et al., 2011; Panwar et al., 2020). While the absence of these bacteria may be due to relatively low light intensities in Lac Pavin's oxycline compared to those in other lakes, there is apparently sufficient light to support some cyanobacteria in this zone (Fig. 1d). Another factor may be the low abundance of sulfide in this zone of the water column. However, our analyses suggest that light impacts metabolic potential of microbial communities in other ways, namely by selecting for bacteria with genes involved in phototrophy (rhodopsin and photosystem proteins) in the shallower interface.

A prior study used amplicon sequencing to document the diversity and distribution of CPR bacteria in Lac Pavin (Borrel et al., 2010). Here, we build upon those results by providing high-quality draft genomes for over 100 CPR bacteria, showing that Lac Pavin harbors a high abundance of these organisms with representatives from virtually all of the major lineages. We also examine the extent to which CPR metabolic potential varies with depth. Broadly speaking, gene content in CPR bacteria, even within specific lineages, can be highly variable (Jaffe et al., 2020; Méheust et al., 2019) raising the question as to what governs gene content in these organisms. One possibility is that ecosystem characteristics, like local geochemistry, drive the distribution of capacities in CPR bacteria. On the other hand, metabolic capacities of the host cell(s) may determine which pathways are present in a given CPR organism. Here, we show that CPR bacteria display relatively little evidence for metabolic differentiation across depth in Lac Pavin, despite very large variations in geochemical conditions. This suggests that within certain habitat types, CPR gene content may be primarily determined by the metabolic capacities of their hosts, rather than geochemistry. In Lac Pavin, this appears to stand in contrast with patterns of gene content observed for other, non-CPR organisms, which are likely shaped more heavily by the availability of oxygen, light, and other key nutrients. Synthesizing these observations, we postulate that geochemistry selects indirectly for CPR bacterial types by selecting for hosts with metabolisms that utilize the spectrum of resources provided across the water column.

One exception to the weak linkage between geochemistry and CPR metabolic traits is the moderate enrichment of superoxide dismutases in CPR bacteria from the partially aerobic interface relative to those from the anoxic water column and sediments. These dismutases may be involved in radical detoxification at the lake interface where oxygen levels are relatively high

compared to deeper compartments (Fig. 1b). Intriguingly, other analyses have also shown that superoxide dismutases are rare in CPR from anoxic human oral microbiomes (Jaffe et al., 2021) yet they do occur in organisms from oxic groundwaters (Chaudhari et al., 2021). Taken together, these findings suggest that superoxide dismutases in CPR are one of the few elements of the CPR pangenome that correlate with oxygen conditions across multiple environment types.

Despite their reduced metabolic platforms, CPR bacteria likely still impact biogeochemistry throughout lakes. Along with genes that shape carbon, sulfur, and nitrogen cycles (Castelle et al., 2018; Danczak et al., 2017; Wrighton et al., 2012), RuBisCO is encoded by some CPR bacteria and could be used to fix carbon dioxide at some rate (Jaffe et al., 2019; Wrighton et al., 2016). Our analyses of RuBisCO in Lac Pavin show that CPR bacteria carry form III-related RuBisCO in the water column, whereas DPANN and other archaea tend to carry them in the sediments. Given the wide environmental prevalence of this RuBisCO form, they are not well characterized biochemically (Flamholz et al., 2019), and thus additional work is required to determine their rates of carbon dioxide incorporation. With this information, it will be possible to better quantify the potential contributions of form III RuBisCO from CPR bacteria to carbon fixation *in situ*.

4.5 Figures

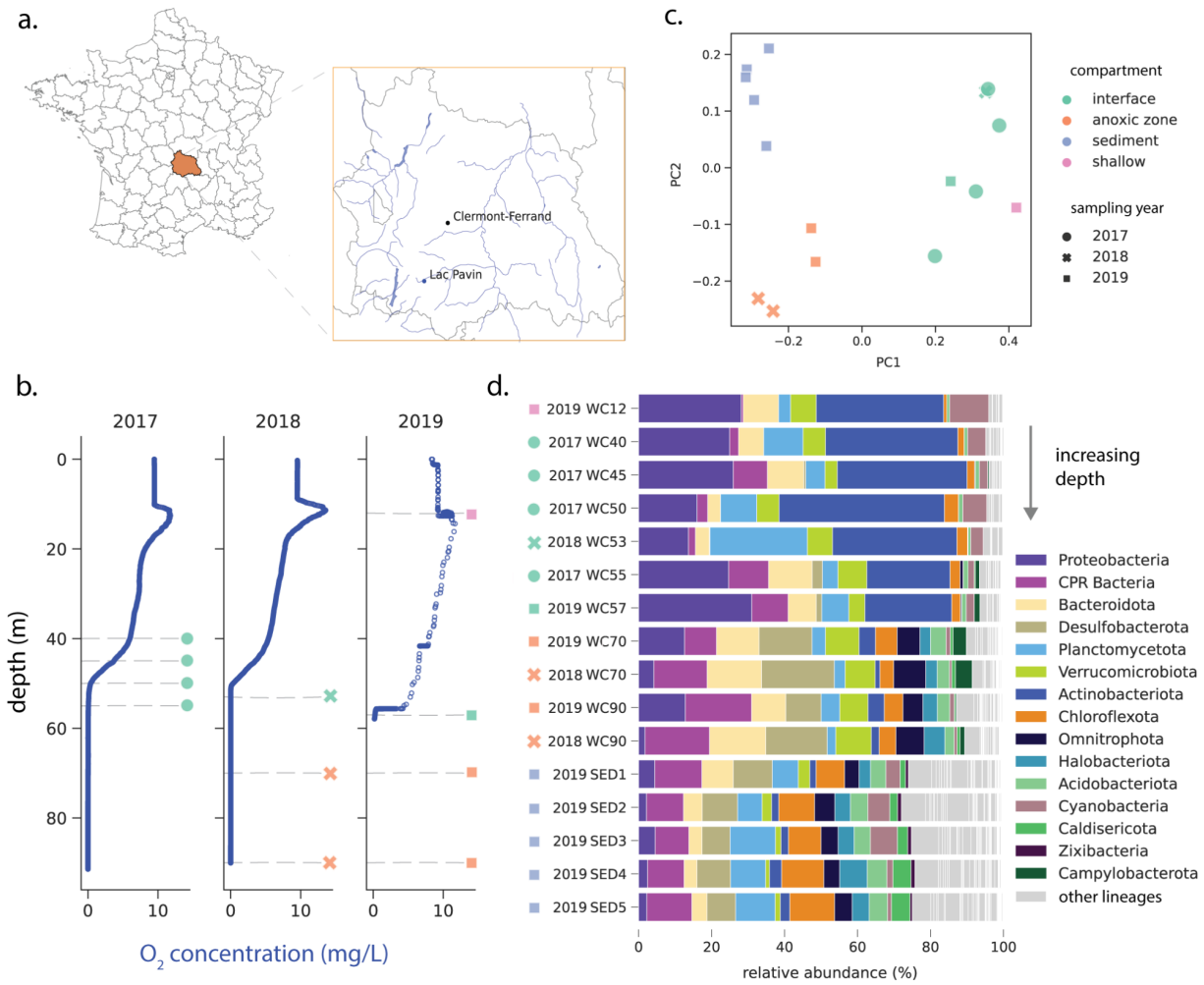

Figure 1. Physical characteristics and microbial communities in Lac Pavin. a) Map showing location of the meromictic Lac Pavin in the Auvergne region of central France. b) Oxygen profiles across the lake water column during each of the three yearly samplings. Depths sampled on a given year are indicated with a dashed grey line, and colored shapes match those in panels cd. c) Principal coordinates analysis of microbial community composition at the phylum level. d) Relative abundance of microbial communities in Lac Pavin at the phylum level, based on read counts of the ribosomal protein S3 gene. Samples are ordered from shallow to deep, including the five sediment samples collected in 2019. **N.B.** For legibility, only the 15 lineages with the highest median abundance across all samples are shown, all others are indicated in grey as "other". Abbreviations: PC, principal coordinate; CPR, Candidate Phyla Radiation, WC, water column; SED, sediment.

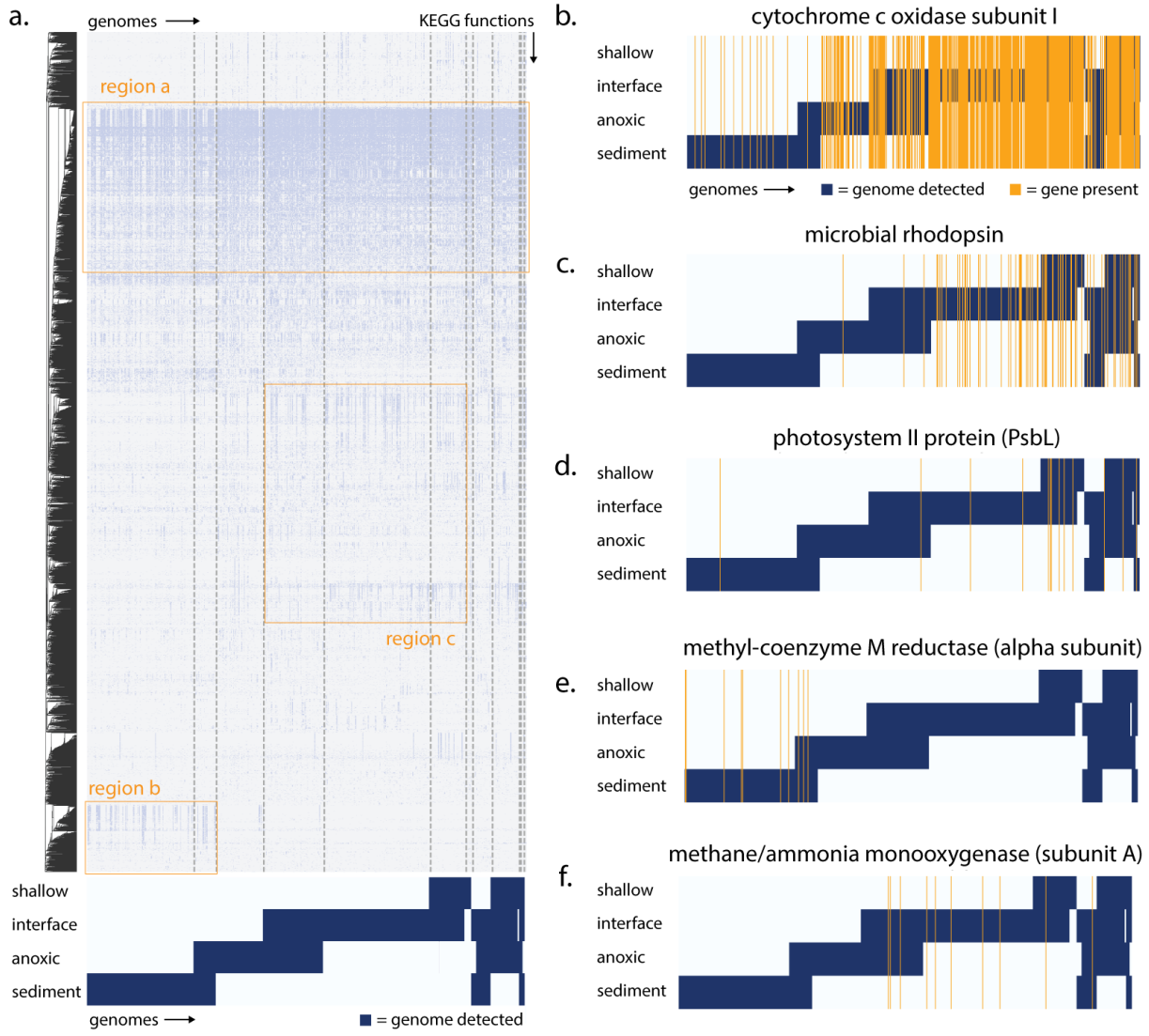

Figure 2. Metabolic potential across the water column and sediments of Lac Pavin. a) Heatmap describing the presence/absence distribution of KEGG functions (rows) across 632 non-CPR bacterial and archaeal genomes (columns). Blue indicates presence, while white indicates absence of a particular KEGG function. The bands below the heatmap indicate the detection of each genome across the water column and sediments by read mapping. b-f) Presence/absence distribution of selected KEGG functions (orange lines) across the genome set. As before, blue bands indicate genome detection across samples.

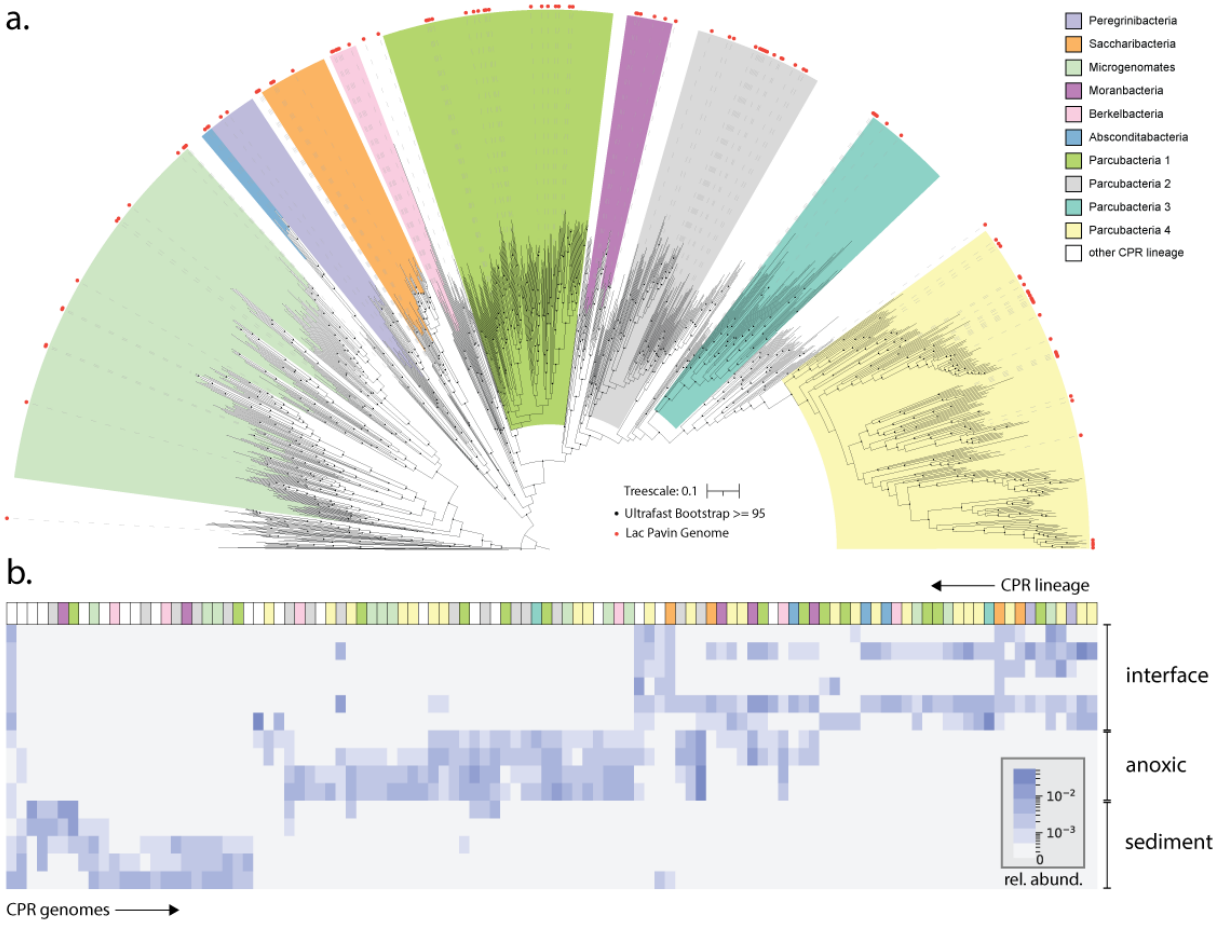

Figure 3. Phylogenomics and abundance distribution of CPR bacteria in Lac Pavin. a) Phylogenetic placement of CPR bacteria from Lac Pavin (red circles) using a set of reference genomes. Major phylogenetic groupings are indicated; however, note that not all are equivalent in terms of rank. Scale bar represents the average number of substitutions per site. b) Relative abundance (coverage) of individual CPR genomes across samples. Lineage assignment (matching the color scheme from panel a) is indicated at the top of panel b.

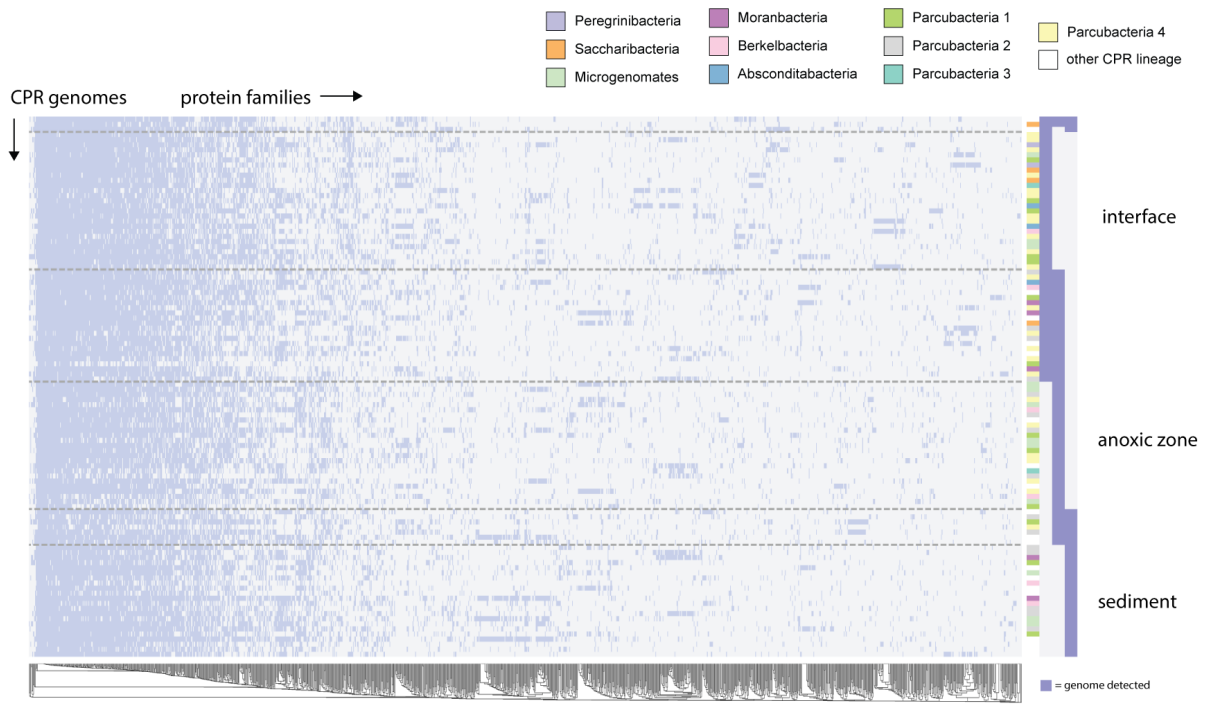

Figure 4. Gene content in the CPR bacteria by depth. a) Heatmap describing the presence/absence of 4,405 protein families across 106 genomes of CPR bacteria. Darker patches indicate presence, while lighter patches indicate absence. Protein families are hierarchically clustered by their distribution pattern across genomes. CPR genomes are organized by their distribution across the lake water column and sediments. In the right panel, darker patches signify that a given genome was detectable in a given compartment, whereas a lighter patch indicates absence/below detectability threshold. Dashed lines distinguish sets of genome with distinct distribution patterns, e.g. the bottom section represents CPR bacteria detected only in the sediments.

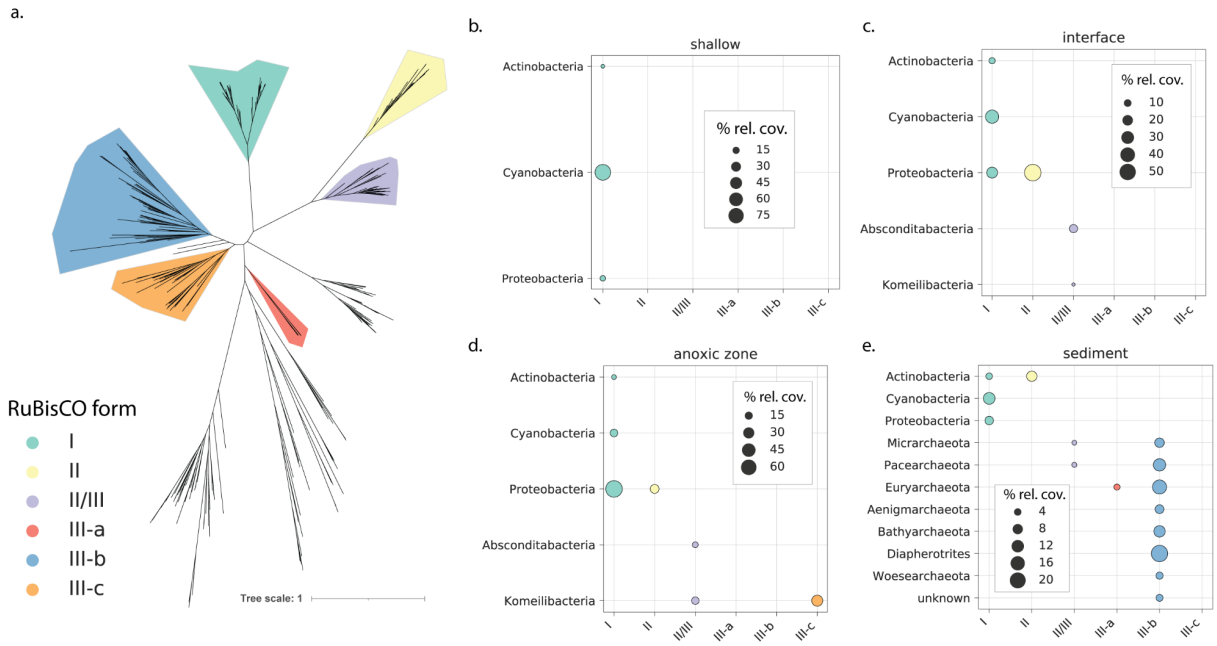

Figure 5. Diversity and distribution of RuBisCO in Lac Pavin. a) Unrooted phylogenetic tree showing relationships among major clades within the RuBisCO superfamily. b-e) Distribution of community metabolic potential for carbon fixation via RuBisCO in each lake compartment. Bubble size represents the percentage of coverage for all scaffolds with RuBisCO attributable to a given lineage/form pair. Percent coverage is averaged across all samples within a lake compartment.

Concluding remarks

Since their discovery via marker gene surveys and subsequent genomic characterization, the largely uncultivated bacterial lineages that would eventually become known as the Candidate Phyla Radiation have posed many intriguing questions for microbial ecologists. For example, what is the full extent of the CPR's phylogenetic and metabolic diversity? What factors have shaped their unique biology, morphology, and gene content? What is the nature and flexibility of the relationships with their microbial hosts? While much remains to be learned about CPR bacteria, the research presented in this dissertation makes some concrete advances towards answering these overarching questions, while also raising some of its own.

A major contribution of this work is clearly mapping out the distribution of core metabolic capacities in CPR bacteria in the framework of a well-resolved internal phylogeny (Chapter 1). Our results suggest that CPR lineages are not highly differentiated at the level of carbon and basic energy metabolism (e.g. glycolysis, the pentose phosphate pathway, acetate/lactate metabolism), and that most probably adopt a fermentative lifestyle, as has been previously suggested (Kantor et al., 2013; Wrighton et al., 2012). Other 'auxiliary' capacities that may function in energy or carbon metabolism, like the RuBisCO or hydrogenases examined here, are scattered across the radiation without strong phylogenetic patterning.

This said, the picture of gene content changes substantially when whole proteomes are considered. At least for the three lineages of CPR bacteria examined in Chapter 3 - the Absconditabacteria, Gracilibacteria, and Saccharibacteria - whole proteome analyses show strong lineage-specific patterning, some of which correlates with the environment of origin. Intriguingly, most genes with narrower distributions across lineages are unable to be confidently annotated, a finding consistent with past work showing that CPR proteomes are highly variable and characterized by tremendous novelty at the protein level (Méheust et al., 2019; Vanni et al., 2021). Future advances in remote homology detection and/or biochemical characterization of hypothetical proteins should help greatly to understand currently cryptic functions that may contribute to CPR metabolism, host relationships, or ecological roles.

The observed pattern of 'diversity within sparsity', referring to the highly variable nature of CPR proteomes despite their overall reduced metabolism, appears to be a unique feature of the CPR, and suggests that the processes shaping their gene content may differ considerably from what has been previously described for other bacteria. The salient biological features of CPR bacteria - including their limited metabolisms and small cell/genome sizes - have fueled speculation that genome reduction has been the dominant process shaping these cells over evolutionary time. A second major contribution of this work is the notion that while genomic loss is certainly evident among CPR, lateral gene transfer has also likely played a major role in shaping gene inventories.

These transfers were probably responsible for distributing both core and auxiliary metabolic functions across the radiation, allowing phylogenetically distant CPR organisms to converge upon similar metabolic platforms. Additionally, some gene acquisitions by CPR probably involved distantly related Bacteria and Archaea, resulting in 'evolutionary mosaicism' of CPR proteomes writ large but also, remarkably, of individual glycolytic enzymes (Chapter 1). Overall, the work presented here points to a complex mix of loss, transfer, and vertical inheritance shaping CPR proteomes over time, suggesting that the evolution of gene content in these organisms cannot solely be characterized by loss/streamlining, which appears to be the dominant process shaping other highly reduced bacterial (endo)symbionts (McCutcheon and Moran, 2011). These findings are particularly interesting to consider in light of newer studies more confidently placing CPR bacteria as a sibling lineage to the Chloroflexi, a free-living lineage of bacteria with replete metabolic capacities (Coleman et al., 2021; Taib et al., 2020).

Despite our growing understanding of *how* gene content has been shaped in CPR bacteria, the underlying drivers of these processes (or, the *why)* are almost entirely unknown. Theoretically, metabolic capacities in CPR could be impacted by two major factors - surrounding environmental conditions or characteristics of their microbial hosts. In the first scenario, ecosystem features or geochemical factors might select for capacities, like the ability to utilize certain sulfur compounds present in the environment. In the second scenario, the capacities of a given CPR might be more tied to the metabolism of the host cell itself and/or the spectrum of resources that are exchanged between them. Currently, these hypotheses are extremely difficult to test directly, given that co-cultures of CPR and their hosts are limited in taxonomic scope and only moderately experimentally tractable. However, signatures of drivers of gene content should also be detectable within genome information, provided that appropriate contrasts are made.

To address this issue, the latter two chapters of this dissertation examined how gene content in CPR varies both across broad habitat types and also within a lake ecosystem with a strong oxygen gradient. In the former study, I find some specific genetic differences between CPR bacteria from 'environmental' habitats compared to those from human/animal microbiomes. However, evolutionary analysis suggested that acquisition of habitat-specific capacities likely followed habitat transitions and thus did not enable them. In the latter study, I show that there is relatively little correlation between CPR metabolic capacity and oxygen concentration in a permanently stratified lake, consistent with similar observations from other ecosystem types (Chaudhari et al., 2021). Together, these findings point to the fact that spatial scale (i.e., across ecosystems vs. within ecosystems) may be relevant for metabolic differentiation in CPR, and that geochemistry is likely a less important factor than host relationships in determining overall gene content. However, I suggest that geochemistry may lead to indirect selection for CPR capacities across environmental gradients by selecting first for host cell types and specific metabolism.

Among the least explored (and thus, most speculative) aspects of CPR evolution is the timing and sequence of major events shaping their diversification. If relationships with microbial hosts have been the primary drivers of gene content in CPR bacteria, did these selection events occur earlier or later in lineage evolution? How flexible have CPR gene sets been over time? Emerging from our work and that of others are several intriguing and potentially contradictory pieces of evidence speaking to these fundamental questions. First is the finding that many CPR proteins, when placed in a phylogenetic context, are highly distinct from reference sequences, typically forming distinct clades (e.g. the NiFe hydrogenases in Chapter 1). This finding suggests that the lateral transfers conferring rarer functions to CPR probably occurred long ago, and is supported by the idea that the CPR lineage itself is likely to be among the earliest evolving groups of bacteria (Coleman et al., 2021). Arising from these combined findings is the intriguing possibility that CPR diversified in tandem with bacterial host lineages, and that aspects of their proteomes were established early in their co-evolution. In this way, gene gain/losses in CPR bacteria may have been somewhat episodic over evolutionary time.

Our analysis in Chapter 3 also suggests that at least some elements of the CPR proteome have been continually remodeled by transfer and loss, likely up until recently. More modern gene gain/loss events may have occurred as new environmental opportunities arose, including the evolution of animals and their microbiomes. However, whether ancient or modern events have imparted a greater effect on extant gene content in CPR bacteria is still an entirely open question.

Overall, the findings of this thesis provide a foundation for future efforts to refine our understanding of the biology of CPR bacteria, and through them, the broader ‘rules of life’ governing ecology and evolution of microbial symbioses on Earth.

References

Abby SS, Kerou M and Schleper C (2020) Ancestral Reconstructions Decipher Major Adaptations of Ammonia-Oxidizing Archaea upon Radiation into Moderate Terrestrial and Marine Environments. *mBio* 11(5). DOI: 10.1128/mBio.02371-20.

Abusleme L, Dupuy AK, Dutzan N, et al. (2013) The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *The ISME journal* 7(5): 1016–1025.

Adam PS, Borrel G, Brochier-Armanet C, et al. (2017) The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *The ISME journal* 11(11): 2407–2425.

Albertsen M, Hugenholtz P, Skarshewski A, et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* 31(6): 533–538.

Alonso H, Blayney MJ, Beck JL, et al. (2009) Substrate-induced assembly of Methanococcoides burtonii D-ribulose-1,5-bisphosphate carboxylase/oxygenase dimers into decamers. *The Journal of biological chemistry* 284(49): 33876–33882.

Anantharaman K, Brown CT, Hug LA, et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature communications* 7: 13219.

Andrei A-Ş, Robeson MS 2nd, Baricz A, et al. (2015) Contrasting taxonomic stratification of microbial communities in two hypersaline meromictic lakes. *The ISME journal* 9(12): 2642–2656.

Aono R, Sato T, Imanaka T, et al. (2015) A pentose bisphosphate pathway for nucleoside degradation in Archaea. *Nature chemical biology* 11(5): 355–360.

Aramaki T, Blanc-Mathieu R, Endo H, et al. (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36(7): 2251–2252.

Ashida H, Danchin A and Yokota A (2005) Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? *Research in microbiology* 156(5-6): 611–618.

Asplund-Samuelsson J, Sundh J, Dupont CL, et al. (2016) Diversity and Expression of Bacterial Metacaspases in an Aquatic Ecosystem. *Frontiers in microbiology* 7: 1043.

Baatar B, Chiang P-W, Rogozin DY, et al. (2016) Bacterial Communities of Three Saline Meromictic Lakes in Central Asia. *PloS one* 11(3): e0150847.

Batinovic S, Rose JJA, Ratcliffe J, et al. (2021) Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams. *Nature microbiology*. DOI: 10.1038/s41564-021-00892-1.

Béjà O and Lanyi JK (2014) Nature's toolkit for microbial rhodopsin ion pumps. *Proceedings of the National Academy of Sciences of the United States of America*.

Berg IA, Kockelkorn D, Ramos-Vera WH, et al. (2010) Autotrophic carbon fixation in archaea. *Nature reviews. Microbiology* 8(6): 447–460.

Biderre-Petit C, Jézéquel D, Dugat-Bony E, et al. (2011) Identification of microbial communities involved in the methane cycle of a freshwater meromictic lake. *FEMS microbiology ecology* 77(3): 533–545.

Bik EM, Costello EK, Switzer AD, et al. (2016) Marine mammals harbor unique microbiotas shaped by and yet distinct from the sea. *Nature communications* 7: 10516.

Boehrer B, von Rohden C and Schultze M (2017) Physical Features of Meromictic Lakes: Stratification and Circulation. In: Gulati RD, Zadereev ES, and Degermendzhi AG (eds) *Ecology of Meromictic Lakes*. Cham: Springer International Publishing, pp. 15–34.

Bor B, McLean JS, Foster KR, et al. (2018) Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. *Proceedings of the National Academy of Sciences of the United States of America* 115(48): 12277–12282.

Bor B, Bedree JK, Shi W, et al. (2019) Saccharibacteria (TM7) in the Human Oral Microbiome. *Journal of dental research* 98(5): 500–509.

Bor B, Collins AJ, Murugkar PP, et al. (2020) Insights Obtained by Culturing Saccharibacteria With Their Bacterial Hosts. *Journal of dental research* 99(6): 685–694.

Borrel G, Lehours A-C, Bardot C, et al. (2010) Members of candidate divisions OP11, OD1 and SR1 are widespread along the water column of the meromictic Lake Pavin (France). *Archives of microbiology* 192(7): 559–567.

Borrel G, Lehours A-C, Crouzet O, et al. (2012) Stratification of Archaea in the deep sediments of a freshwater meromictic lake: vertical shift from methanogenic to uncultured archaeal lineages. *PloS one* 7(8): e43346.

Bouma-Gregson K, Olm MR, Probst AJ, et al. (2019) Impacts of microbial assemblage and environmental conditions on the distribution of anatoxin-a producing cyanobacteria within a river network. *The ISME journal* 13(6): 1618–1634.

Boyd JA, Woodcroft BJ and Tyson GW (2018) GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic acids research* 46(10): e59.

Bräsen C, Esser D, Rauch B, et al. (2014) Carbohydrate metabolism in Archaea: current insights

into unusual enzymes and pathways and their regulation. *Microbiology and molecular biology reviews: MMBR* 78(1): 89–175.

Brown CT, Hug LA, Thomas BC, et al. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523(7559): 208–211.

Campbell JH, O'Donoghue P, Campbell AG, et al. (2013) UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences of the United States of America* 110(14): 5540–5545.

Canchaya C, Fournous G, Chibani-Chennoufi S, et al. (2003) Phage as agents of lateral gene transfer. *Current opinion in microbiology* 6(4): 417–424.

Capella-Gutiérrez S, Silla-Martínez JM and Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15): 1972–1973.

Cardenas JP, Quatrini R and Holmes DS (2016) Aerobic Lineage of the Oxidative Stress Response Protein Rubrerythrin Emerged in an Ancient Microaerobic, (Hyper)Thermophilic Environment. *Frontiers in Microbiology*. DOI: 10.3389/fmicb.2016.01822.

Carter MS, Zhang X, Huang H, et al. (2018) Functional assignment of multiple catabolic pathways for D-apiose. *Nature chemical biology* 14(7): 696–705.

Castelle CJ and Banfield JF (2018) Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* 172(6): 1181–1197.

Castelle CJ, Wrighton KC, Thomas BC, et al. (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current biology: CB* 25(6): 690–701.

Castelle CJ, Brown CT, Thomas BC, et al. (2017) Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the Candidate Phyla Radiation. *Scientific reports* 7: 40101.

Castelle CJ, Brown CT, Anantharaman K, et al. (2018) Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature reviews. Microbiology* 16(10): 629–645.

Chaudhari NM, Overholt WA, Figueroa-Gonzalez PA, et al. (2021) The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. *bioRxiv*. DOI: 10.1101/2021.07.28.454184.

Chaumeil P-A, Mussig AJ, Hugenholtz P, et al. (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* . DOI: 10.1093/bioinformatics/btz848.

Chen L-X, Anantharaman K, Shaiber A, et al. (2020) Accurate and complete genomes from

metagenomes. *Genome research* 30(3): 315–333.

Coleman GA, Davín AA, Mahendrarajah TA, et al. (2021) A rooted phylogeny resolves early bacterial evolution. *Science* 372(6542). DOI: 10.1126/science.abe0511.

Constant P, Chowdhury SP, Hesse L, et al. (2011) Genome data mining and soil survey for the novel group 5 [NiFe]-hydrogenase to explore the diversity and ecological importance of presumptive high-affinity H(2)-oxidizing bacteria. *Applied and environmental microbiology* 77(17): 6027–6035.

Conway T (1992) The Entner-Doudoroff pathway: history, physiology and molecular biology. *FEMS microbiology reviews* 9(1): 1–27.

Cooper SJ, Leonard GA, McSweeney SM, et al. (1996) The crystal structure of a class II fructose-1,6-bisphosphate aldolase shows a novel binuclear metal-binding active site embedded in a familiar fold. *Structure* 4(11): 1303–1315.

Criscuolo A and Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology* 10: 210.

Crits-Christoph A, Diamond S, Al-Shayeb B, et al. (2021) A widely distributed genus of soil Acidobacteria genomically enriched in biosynthetic gene clusters. *bioRxiv*. DOI: 10.1101/2021.05.10.443473.

Crooks GE, Hon G, Chandonia J-M, et al. (2004) WebLogo: a sequence logo generator. *Genome research* 14(6). genome.cshlp.org: 1188–1190.

Cross KL, Campbell JH, Balachandran M, et al. (2019) Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nature biotechnology* 37(11): 1314–1321.

Crummett LT, Puxty RJ, Weihe C, et al. (2016) The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology* 499: 219–229.

Danczak RE, Johnston MD, Kenah C, et al. (2017) Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome* 5(1): 112.

Dudek NK, Sun CL, Burstein D, et al. (2017) Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome. *Current biology: CB* 27(24): 3752–3762.e6.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19). academic.oup.com: 2460–2461.

Erb TJ and Zarzycki J (2018) A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. *Current opinion in biotechnology* 49: 100–107.

Eren AM, Esen ÖC, Quince C, et al. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3: e1319.

Feng L, Sun Y, Deng H, et al. (2014) Structural and biochemical characterization of fructose-1,6/sedoheptulose-1,7-bisphosphatase from the cyanobacterium Synechocystis strain 6803. *The FEBS journal* 281(3): 916–926.

Finn RD, Clements J and Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 39(Web Server issue). academic.oup.com: W29–37.

Finstad KM, Probst AJ, Thomas BC, et al. (2017) Microbial Community Structure and the Persistence of Cyanobacterial Populations in Salt Crusts of the Hyperarid Atacama Desert from Genome-Resolved Metagenomics. *Frontiers in Microbiology*. DOI: 10.3389/fmicb.2017.01435.

Flamholz AI, Prywes N, Moran U, et al. (2019) Revisiting Trade-offs between Rubisco Kinetic Parameters. *Biochemistry* 58(31): 3365–3376.

Gerbling KP, Steup M and Latzko E (1986) Fructose 1,6-Bisphosphatase Form B from Synechococcus leopoliensis Hydrolyzes both Fructose and Sedoheptulose Bisphosphate. *Plant physiology* 80(3): 716–720.

Greening C, Biswas A, Carere CR, et al. (2016) Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *The ISME journal* 10(3): 761–777.

Hall KJ and Northcote TG (2012) Meromictic lakes. *Encyclopedia of lakes and reservoirs*. Springer Netherlands Dordrecht: 519–524.

Hamilton TL, Bovee RJ, Thiel V, et al. (2014) Coupled reductive and oxidative sulfur cycling in the phototrophic plate of a meromictic lake. *Geobiology* 12(5): 451–468.

Hanke A, Hamann E, Sharma R, et al. (2014) 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. *Frontiers in microbiology* 5. Frontiers: 231.

Hansen T, Wendorff D and Schönheit P (2004) Bifunctional phosphoglucose/phosphomannose isomerases from the Archaea Aeropyrum pernix and Thermoplasma acidophilum constitute a novel enzyme family within the phosphoglucose isomerase superfamily. *The Journal of biological chemistry* 279(3): 2262–2272.

Hansen T, Schlichting B, Felgendreher M, et al. (2005) Cupin-type phosphoglucose isomerases (Cupin-PGIs) constitute a novel metal-dependent PGI family representing a convergent line of PGI evolution. *Journal of bacteriology* 187(5): 1621–1631.

Hanson TE and Tabita FR (2001) A ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO)-like protein from Chlorobium tepidum that is involved with sulfur metabolism

and the response to oxidative stress. *Proceedings of the National Academy of Sciences of the United States of America* 98(8): 4397–4402.

Harrison EP, Willingham NM, Lloyd JC, et al. (1997) Reduced sedoheptulose-1,7-bisphosphatase levels in transgenic tobacco lead to decreased photosynthetic capacity and altered carbohydrate accumulation. *Planta* 204(1). Springer-Verlag: 27–36.

Hasegawa M, Hosaka T, Kojima K, et al. (2020) A unique clade of light-driven proton-pumping rhodopsins evolved in the cyanobacterial lineage. *Scientific reports* 10(1): 16752.

He C, Keren R, Whittaker ML, et al. (2021) Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nature microbiology*. DOI: 10.1038/s41564-020-00840-5.

Hernsdorf AW, Amano Y, Miyakawa K, et al. (2017) Potential for microbial H2and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. *The ISME journal* 11(8): 1915–1929.

He X, McLean JS, Edlund A, et al. (2015) Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America* 112(1): 244–249.

Hoang DT, Chernomor O, von Haeseler A, et al. (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular biology and evolution* 35(2): 518–522.

Huddy RJ, Sachdeva R, Kadzinga F, et al. (2020) Thiocyanate and organic carbon inputs drive convergent selection for specific autotrophic Afipia and Thiobacillus strains within complex microbiomes. *Cold Spring Harbor Laboratory*. DOI: 10.1101/2020.04.29.067207.

Hug LA, Castelle CJ, Wrighton KC, et al. (2013) Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* 1(1): 22.

Hug LA, Baker BJ, Anantharaman K, et al. (2016) A new view of the tree of life. *Nature microbiology* 1: 16048.

Hyatt D, Chen G-L, Locascio PF, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11: 119.

Imanaka H, Yamatsu A, Fukui T, et al. (2006) Phosphoenolpyruvate synthase plays an essential role for glycolysis in the modified Embden-Meyerhof pathway in Thermococcus kodakarensis. *Molecular microbiology* 61(4): 898–909.

İnceoğlu Ö, Llirós M, Crowe SA, et al. (2015) Vertical Distribution of Functional Potential and Active Microbial Communities in Meromictic Lake Kivu. *Microbial ecology* 70(3): 596–611.

Jaffe AL, Corel E, Pathmanathan JS, et al. (2016) Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins. *Environmental microbiology* 18(12): 5072–5081.

Jaffe AL, Castelle CJ, Dupont CL, et al. (2019) Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea. *Molecular biology and evolution* 36(3): 435–446.

Jaffe AL, Castelle CJ, Matheus Carnevali PB, et al. (2020) The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biology*. DOI: 10.1186/s12915-020-00804-5.

Jaffe AL, Thomas AD, He C, et al. (2021) Patterns of Gene Content and Co-occurrence Constrain the Evolutionary Path toward Animal Association in Candidate Phyla Radiation Bacteria. *mBio*: e0052121.

Joshi NA, Fass J and Others (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software].

Kalyaanamoorthy S, Minh BQ, Wong TKF, et al. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* 14(6): 587–589.

Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28(1). academic.oup.com: 27–30.

Kang D, Li F, Kirton ES, et al. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. DOI: 10.7287/peerj.preprints.27522v1.

Kantor RS, Wrighton KC, Handley KM, et al. (2013) Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* 4(5): e00708–13.

Kantor RS, van Zyl AW, van Hille RP, et al. (2015) Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. *Environmental microbiology* 17(12): 4929–4941.

Katoh K and Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30(4). academic.oup.com: 772–780.

Kono T, Mehrotra S, Endo C, et al. (2017) A RuBisCO-mediated carbon metabolic pathway in methanogenic archaea. *Nature communications* 8: 14007.

Langmead B and Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9(4): 357–359.

Lehours A-C, Evans P, Bardot C, et al. (2007) Phylogenetic diversity of archaea and bacteria in the anoxic zone of a meromictic lake (Lake Pavin, France). *Applied and environmental*

microbiology 73(6): 2016–2019.

Letunic I and Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* 44(W1): W242–5.

Li D, Liu C-M, Luo R, et al. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10): 1674–1676.

Lopes F, Viollier E, Thiam A, et al. (2011) Biogeochemical modelling of anaerobic vs. aerobic methane oxidation in a meromictic crater lake (Lake Pavin, France). *Applied geochemistry: journal of the International Association of Geochemistry and Cosmochemistry* 26(12): 1919–1932.

Luef B, Frischkorn KR, Wrighton KC, et al. (2015) Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature communications* 6: 6372.

Ma K, Schicho RN, Kelly RM, et al. (1993) Hydrogenase of the hyperthermophile Pyrococcus furiosus is an elemental sulfur reductase or sulfhydrogenase: evidence for a sulfur-reducing hydrogenase ancestor. *Proceedings of the National Academy of Sciences of the United States of America* 90(11): 5341–5344.

Maliar N, Okhrimenko IS, Petrovskaya LE, et al. (2020) Novel pH-Sensitive Microbial Rhodopsin from Sphingomonas paucimobilis. *Doklady. Biochemistry and biophysics* 495(1): 342–346.

Martijn J, Schön ME, Lind AE, et al. (2020) Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nature communications* 11(1): 5490.

Martínez Arbas S, Narayanasamy S, Herold M, et al. (2021) Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nature Microbiology* 6(1): 123–135.

Martiny AC, Treseder K and Pusch G (2013) Phylogenetic conservatism of functional traits in microorganisms. *The ISME journal* 7(4): 830–838.

Matheus Carnevali PB, Schulz F, Castelle CJ, et al. (2019) Hydrogen-based metabolism as an ancestral trait in lineages sibling to the Cyanobacteria. *Nature communications* 10(1): 463.

McCutcheon JP and Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nature reviews. Microbiology* 10(1): 13–26.

McLean JS, Bor B, Kerns KA, et al. (2020) Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts. *Cell reports* 32(3): 107939.

Méheust R, Burstein D, Castelle CJ, et al. (2019) The distinction of CPR bacteria from other bacteria based on protein family content. *Nature communications* 10(1): 4173.

Méheust R, Castelle CJ, Matheus Carnevali PB, et al. (2020) Groundwater Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have a novel nitrogenase paralog. *The ISME journal* 14(12): 2907–2922.

Mendler K, Chen H, Parks DH, et al. (2019) AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic acids research* 47(9): 4442–4448.

Meyer KM, Macalady JL, Fulton JM, et al. (2011) Carotenoid biomarkers as an imperfect reflection of the anoxygenic phototrophic community in meromictic Fayetteville Green Lake. *Geobiology* 9(4): 321–329.

Minh BQ, Schmidt HA, Chernomor O, et al. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular biology and evolution* 37(5): 1530–1534.

Moran NA and Wernegreen JJ (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends in ecology & evolution* 15(8): 321–326.

Moreira D, Zivanovic Y, López-Archilla AI, et al. (2020) Reductive evolution and unique infection and feeding mode in the CPR predatory bacterium Vampirococcus lugosii. *Cold Spring Harbor Laboratory*. DOI: 10.1101/2020.11.10.374967.

Moreira D, Zivanovic Y, López-Archilla AI, et al. (2021) Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. *Nature communications* 12(1): 2454.

Murugkar PP, Collins AJ, Chen T, et al. (2020) Isolation and cultivation of candidate phyla radiation Saccharibacteria (TM7) bacteria in coculture with bacterial hosts. *Journal of oral microbiology* 12(1). Taylor & Francis: 1814666.

Nguyen L-T, Schmidt HA, von Haeseler A, et al. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* 32(1): 268–274.

Nicolas AM, Jaffe AL, Nuccio EE, et al. (2021) Soil Candidate Phyla Radiation Bacteria Encode Components of Aerobic Metabolism and Co-occur with Nanoarchaea in the Rare Biosphere of Rhizosphere Grassland Communities. *mSystems* 6(4): e0120520.

Olm MR, Brown CT, Brooks B, et al. (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal* 11(12). nature.com: 2864–2868.

Orakov A, Fullam A, Coelho LP, et al. (2021) GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome biology* 22(1): 178.

Panwar P, Allen MA, Williams TJ, et al. (2020) Influence of the polar light cycle on seasonal dynamics of an Antarctic lake microbial community. *Microbiome* 8(1): 116.

Parks DH, Imelfort M, Skennerton CT, et al. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* 25(7): 1043–1055.

Parks DH, Rinke C, Chuvochina M, et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology* 2(11): 1533–1542.

Parks DH, Chuvochina M, Waite DW, et al. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology* 36(10): 996–1004.

Pedroni P, Della Volpe A, Galli G, et al. (1995) Characterization of the locus encoding the [Ni-Fe] sulfhydrogenase from the archaeon Pyrococcus furiosus: evidence for a relationship to bacterial sulfite reductases. *Microbiology* 141 (Pt 2): 449–458.

Peng Y, Leung HCM, Yiu SM, et al. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28(11): 1420–1428.

Peura S, Eiler A, Bertilsson S, et al. (2012) Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1. *The ISME journal* 6(9): 1640–1652.

Price MN, Dehal PS and Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* 5(3): e9490.

Probst AJ, Castelle CJ, Singh A, et al. (2017) Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. *Environmental microbiology* 19(2): 459–474.

Probst AJ, Ladd B, Jarett JK, et al. (2018) Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nature microbiology* 3(3): 328–336.

Pushkarev A, Inoue K, Larom S, et al. (2018) A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. *Nature* 558(7711): 595–599.

Rinke C, Schwientek P, Sczyrba A, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459): 431–437.

Rissanen AJ, Saarela T, Jäntti H, et al. (2020) Vertical stratification patterns of methanotrophs and their genetic controllers in water columns of oxygen-stratified boreal lakes. *FEMS microbiology ecology*. DOI: 10.1093/femsec/fiaa252.

Rozenberg A, Inoue K, Kandori H, et al. (2021) Microbial Rhodopsins: The Last Two Decades. *Annual review of microbiology* 75: 427–447.

Saito Y, Ashida H, Sakiyama T, et al. (2009) Structural and functional similarities between a

ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO)-like protein from Bacillus subtilis and photosynthetic RuBisCO. *The Journal of biological chemistry* 284(19): 13256–13264.

Sato T, Atomi H and Imanaka T (2007) Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science* 315(5814): 1003–1006.

Sauer U and Eikmanns BJ (2005) The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. *FEMS microbiology reviews* 29(4): 765–794.

Say RF and Fuchs G (2010) Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* 464(7291): 1077–1081.

Schönheit P, Buckel W and Martin WF (2016) On the Origin of Heterotrophy. *Trends in microbiology* 24(1): 12–25.

Shaiber A, Willis AD, Delmont TO, et al. (2020) Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome biology* 21(1): 292.

Sheridan PO, Raguideau S, Quince C, et al. (2020) Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. *Nature communications* 11(1): 5494.

Sieber CMK, Probst AJ, Sharrar A, et al. (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature microbiology* 3(7): 836–843.

Sieber CMK, Paul BG, Castelle CJ, et al. (2019) Unusual metabolism and hypervariation in the genome of a Gracilibacteria (BD1-5) from an oil degrading community. *bioRxiv*. DOI: 10.1101/595074.

Siebers B and Schönheit P (2005) Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Current opinion in microbiology* 8(6): 695–705.

Silva PJ, van den Ban EC, Wassink H, et al. (2000) Enzymes of hydrogen metabolism in Pyrococcus furiosus. *European journal of biochemistry / FEBS* 267(22): 6541–6551.

Soro V, Dutton LC, Sprague SV, et al. (2014) Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria. *Applied and environmental microbiology* 80(20): 6480–6489.

Spang A, Saw JH, Jørgensen SL, et al. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521(7551): 173–179.

Stamatakis A, Hoover P and Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. *Systematic biology* 57(5): 758–771.

Starr EP, Shi S, Blazewicz SJ, et al. (2018) Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived

carbon. *Microbiome* 6(1): 122.

Stechmann A, Baumgartner M, Silberman JD, et al. (2006) The glycolytic pathway of Trimastix pyriformis is an evolutionary mosaic. *BMC evolutionary biology* 6: 101.

Stolzenberger J, Lindner SN, Persicke M, et al. (2013) Characterization of fructose 1,6-bisphosphatase and sedoheptulose 1,7-bisphosphatase from the facultative ribulose monophosphate cycle methylotroph Bacillus methanolicus. *Journal of bacteriology* 195(22): 5112–5122.

Szöllősi GJ, Rosikiewicz W, Boussau B, et al. (2013) Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic biology* 62(6). Oxford Academic: 901–912.

Tabita FR, Hanson TE, Li H, et al. (2007) Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *M ICROBIOLOGY AND M OLECULAR B IOLOGY R EVIEWS* 71(4): 576–599.

Tabita FR, Satagopan S, Hanson TE, et al. (2008) Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *Journal of experimental botany* 59(7): 1515–1524.

Taib N, Megrian D, Witwinowski J, et al. (2020) Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nature ecology & evolution* 4(12): 1661–1672.

Thompson LR, Zeng Q, Kelly L, et al. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America* 108(39): E757–64.

Tran PQ, McIntyre PB, Kraemer BM, et al. (2019) Depth-discrete eco-genomics of Lake Tanganyika reveals roles of diverse microbes, including candidate phyla, in tropical freshwater nutrient cycling. *bioRxiv*. DOI: 10.1101/834861.

Tuininga JE, Verhees CH, van der Oost J, et al. (1999) Molecular and biochemical characterization of the ADP-dependent phosphofructokinase from the hyperthermophilic archaeon Pyrococcus furiosus. *The Journal of biological chemistry* 274(30): 21023–21028.

Tully BJ, Graham ED and Heidelberg JF (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific data* 5: 170203.

Tyson GW, Chapman J, Hugenholtz P, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978): 37–43.

Utter DR, He X, Cavanaugh CM, et al. (2020) The saccharibacterium TM7x elicits differential responses across its host range. *The ISME journal*. DOI: 10.1038/s41396-020-00736-6.

Van Der Oost J and Siebers B (2007) The glycolytic pathways of Archaea: evolution by tinkering. *Archaea: evolution, physiology and molecular biology* 22. Wiley Online Library:

247–260.

Van Dongen S (2008) Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications* 30(1). Society for Industrial and Applied Mathematics: 121–141.

van Haaster DJ, Silva PJ, Hagedoorn P-L, et al. (2008) Reinvestigation of the steady-state kinetics and physiological function of the soluble NiFe-hydrogenase I of Pyrococcus furiosus. *Journal of bacteriology* 190(5): 1584–1587.

Vanni C, Schechter MS, Acinas SG, et al. (2021) Unifying the known and unknown microbial coding sequence space. *bioRxiv*. DOI: 10.1101/2020.06.30.180448.

Vavourakis CD, Andrei A-S, Mehrshad M, et al. (2018) A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. *Microbiome* 6(1): 168.

Verhees CH, Huynen MA, Ward DE, et al. (2001) The Phosphoglucose Isomerase from the Hyperthermophilic ArchaeonPyrococcus furiosus Is a Unique Glycolytic Enzyme That Belongs to the Cupin Superfamily. *The Journal of biological chemistry* 276(44): 40926–40932.

Verhees CH, Kengen SWM, Tuininga JE, et al. (2004) The unique features of glycolytic pathways in Archaea. *Biochemical Journal*. DOI: 10.1042/bj3770819.

Vignais PM and Billoud B (2007) Occurrence, classification, and biological function of hydrogenases: an overview. *Chemical reviews* 107(10): 4206–4272.

Wang H-C, Susko E and Roger AJ (2019) The Relative Importance of Modeling Site Pattern Heterogeneity versus Partition-wise Heterotachy in Phylogenomic Inference. *Systematic biology*. DOI: 10.1093/sysbio/syz021.

Williams TA, Szöllősi GJ, Spang A, et al. (2017) Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America* 114(23): E4602–E4611.

Wrighton KC, Thomas BC, Sharon I, et al. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337(6102): 1661–1665.

Wrighton KC, Castelle CJ, Wilkins MJ, et al. (2014) Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *The ISME journal* 8(7): 1452–1463.

Wrighton KC, Castelle CJ, Varaljay VA, et al. (2016) RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. *The ISME journal* 10(11): 2702–2714.

Yamamoto T, Iino H, Kim K, et al. (2011) Evidence for ATP-dependent structural rearrangement of nuclease catalytic site in DNA mismatch repair endonuclease MutL. *The Journal of*

biological chemistry 286(49): 42337–42348.

Yang J, Yan R, Roy A, et al. (2015) The I-TASSER Suite: protein structure and function prediction. *Nature methods* 12(1). nature.com: 7–8.

Yoo JG and Bowien B (1995) Analysis of the cbbF genes from Alcaligenes eutrophus that encode fructose-1,6-/sedoheptulose-1,7-bisphosphatase. *Current microbiology* 31(1): 55–61.

Zeng Y, Chen X, Madsen AM, et al. (2020) Potential Rhodopsin- and Bacteriochlorophyll-Based Dual Phototrophy in a High Arctic Glacier. *mBio* 11(6). DOI: 10.1128/mBio.02641-20.

Zhu Q, Mai U, Pfeiffer W, et al. (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature communications* 10(1): 5477.

Appendix / Glossary of Terms

CPR - Candidate Phyla Radiation

DPANN archaea - a group of archaea originally defined based on the lineages Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota, but now including others.

HMM - Hidden Markov Model