AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

# Patterns of Gene Content and Co-occurrence Constrain the Evolutionary Path toward Animal Association in Candidate Phyla Radiation Bacteria

Alexander L. Jaffe,[a] Alex D. Thomas,[b,c] Christine He,[d] Ray Keren,[e] Luis E. Valentin-Alvarado,[a,d] Patrick Munk,[f] Keith Bouma-Gregson,[g,h] Ibrahim F. Farag,[i] Yuki Amano,[j,k] Rohan Sachdeva,[d,g] Patrick T. West,[l] Jillian F. Banfield[b,d,g,m]

[a]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA
[b]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, California, USA
[c]Rocky Mountain Biological Laboratory, Crested Butte, Colorado, USA
[d]Innovative Genomics Institute, University of California, Berkeley, Berkeley, California, USA
[e]Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, California, USA
[f]National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark
[g]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California, USA
[h]Department of Integrative Biology, University of California, Berkeley, Berkeley, California, USA
[i]School of Marine Science and Policy, University of Delaware, Lewes, Delaware, USA
[j]Nuclear Fuel Cycle Engineering Laboratories, Japan Atomic Energy Agency, Ibaraki, Japan
[k]Horonobe Underground Research Center, Japan Atomic Energy Agency, Hokkaido, Japan
[l]Department of Medicine (Hematology & Blood and Marrow Transplantation), Stanford University, Stanford, California, USA
[m]Chan Zuckerberg Biohub, San Francisco, California, USA

**ABSTRACT** 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.

**IMPORTANCE** Studying the genetic differences between related microorganisms from different environment types can indicate factors associated with their movement among habitats. This is particularly interesting for bacteria from the Candidate Phyla Radiation because their minimal metabolic capabilities require associations with

**Citation** Jaffe AL, Thomas AD, He C, Keren R, Valentin-Alvarado LE, Munk P, Bouma-Gregson K, Farag IF, Amano Y, Sachdeva R, West PT, Banfield JF. 2021. Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in Candidate Phyla Radiation bacteria. mBio 12:e00521-21. https://doi.org/10.1128/mBio.00521-21.

**Editor** Stephen J. Giovannoni, Oregon State University



Address correspondence to Jillian F. Banfield, jbanfield@berkeley.edu.

**Received** 9 March 2021
**Accepted** 14 June 2021
**Published** 13 July 2021

microbial hosts. We found that shifts of Absconditabacteria, Gracilibacteria, and Saccharibacteria between environmental ecosystems and mammalian mouths/guts probably did not involve major episodes of gene gain and loss; rather, gradual genomic change likely followed habitat migration. The results inform our understanding of how little-known microorganisms establish in the human microbiota where they may ultimately impact health.

**KEYWORDS** CPR bacteria, animal microbiome, bacterial evolution, comparative genomics, habitat transition

The Candidate Phyla Radiation (CPR) is a phylogenetically diverse clade of bacteria characterized by reduced metabolisms, potentially episymbiotic lifestyles, and ultrasmall cells (1–3). While the first high-quality CPR genomes were primarily from groundwater, sediment, and wastewater (4–6), 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 (7). 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 (8, 9). 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 (4, 10, 11).

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 (4, 6). Despite a generally reduced metabolic platform, CPR bacteria display substantial variation in their genetic capacities, even within lineages (12, 13). For example, some Gracilibacteria lack essentially all genes of the glycolysis and pentose-phosphate pathways and the tricarboxylic acid (TCA) cycle (14). In contrast to many CPR bacteria, soil-associated Saccharibacteria harbor numerous genes related to oxygen metabolism (15, 16). Pangenome analyses have shown genetic evidence for niche partitioning among Saccharibacteria from the same body site (17). 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 (17). While there has been significant progress in characterizing the relationship between Saccharibacteria and Actinobacteria in the oral habitat (3, 18, 19), other CPR-host relationships remain unclear. Elucidation of environmental transitions among CPR lineages will require both thorough analysis of functional repertoires and a more comprehensive understanding of associations with other microorganisms. Here, we expand 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 investigate 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.

## 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 (Materials and Methods and see Table S1 at https://doi.org/10.5281/zenodo.4560554) (2, 4–7, 10, 15–18, 20–83). Quality filtration of this curated genome set at ≥70% completeness and ≤10% contamination, plus subsequent dereplication at 99% average nucleotide identity (ANI), yielded a nonredundant set of 389 genomes for downstream analysis (see Table S1 at https://doi.org/10.5281/zenodo.4560554). 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 (Fig. 1a; see also File S1 at https://doi.org/10.5281/zenodo.4560554), consistent with previous findings (13). 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 (GRA1 and -2), each with a small subclade comprised of animal-associated genomes (see Table S1 at https://doi.org/10.5281/zenodo.4560554). 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 (Fig. 1a; see also Table S1 at https://doi.org/10.5281/zenodo.4560554). Two of these three subclades were exclusively composed of animal-associated sequences whereas one (SAC5) 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 (24) 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 (SAC1 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 (66, 71).

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 read recruitment comparable to Saccharibacteria and were particularly abundant in some groundwater, engineered, and animal-associated habitats. In contrast to Saccharibacteria, Gracilibacteria and Absconditabacteria have so far been only minimally detected in soil and plant-associated microbiomes. We also compared abundance patterns across animal body sites. As expected based on extensive prior work (3, 8, 45), 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

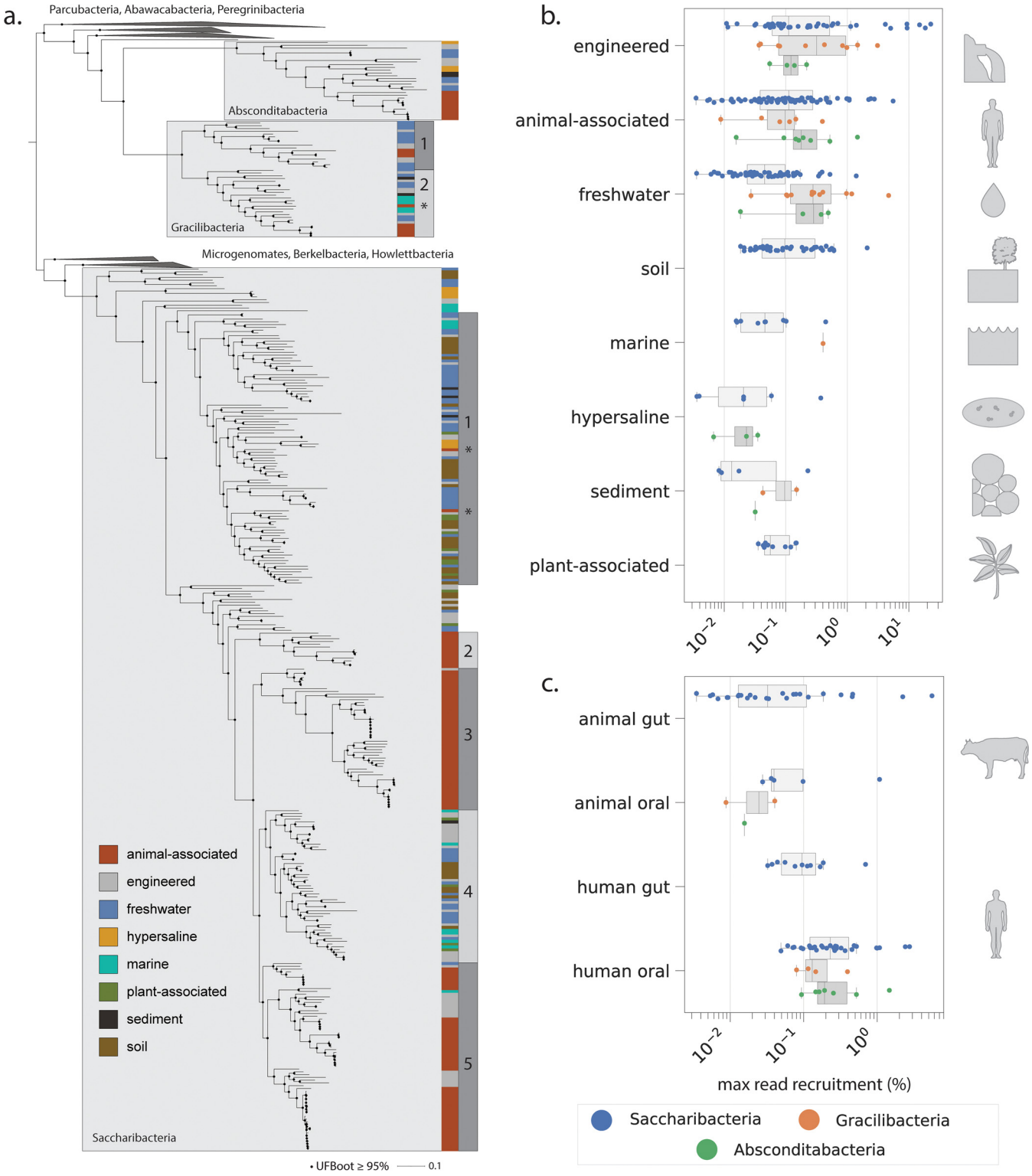


**FIG 1** Phylogenetic and environmental patterns for the Absconditabacteria, Gracilibacteria, and Saccharibacteria. (a) Maximum-likelihood tree based on 16 concatenated ribosomal proteins (1,976 amino acids, LG+R10 model). Scale bar represents the average number of substitutions per site. Habitat of 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. (b and c) Percentage of reads per metagenomic sample mapping to individual genomes across environments (b) and body sites of humans and animals (c).

from cows, pigs, and dolphins (Fig. 1c), in one case approaching 5% of reads (see Table S2 at https://doi.org/10.5281/zenodo.4560554). 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 (39, 56); 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 (3, 18, 19, 45, 84), 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. 1b and c). rpS3 sequences from all samples were clustered into "species groups" (Materials and Methods). We observed that species groups from Actinobacteria and Saccharibacteria frequently cooccurred 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 detectable 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 than of 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 to 11 Saccharibacteria species but only one detectable Actinobacteria species (see Table S3 at https://doi.org/10.5281/zenodo.4560554). 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 cooccurred within a single sample onto species trees constructed from recovered rpS3 sequences (Fig. 2b), including 22 Saccharibacteria-Actinobacteria pairs reported in previous experimental studies (see Table S4 at https://doi.org/10.5281/zenodo.4560554). In three cases, we found that individual metagenomic samples contained only one assembled Saccharibacteria

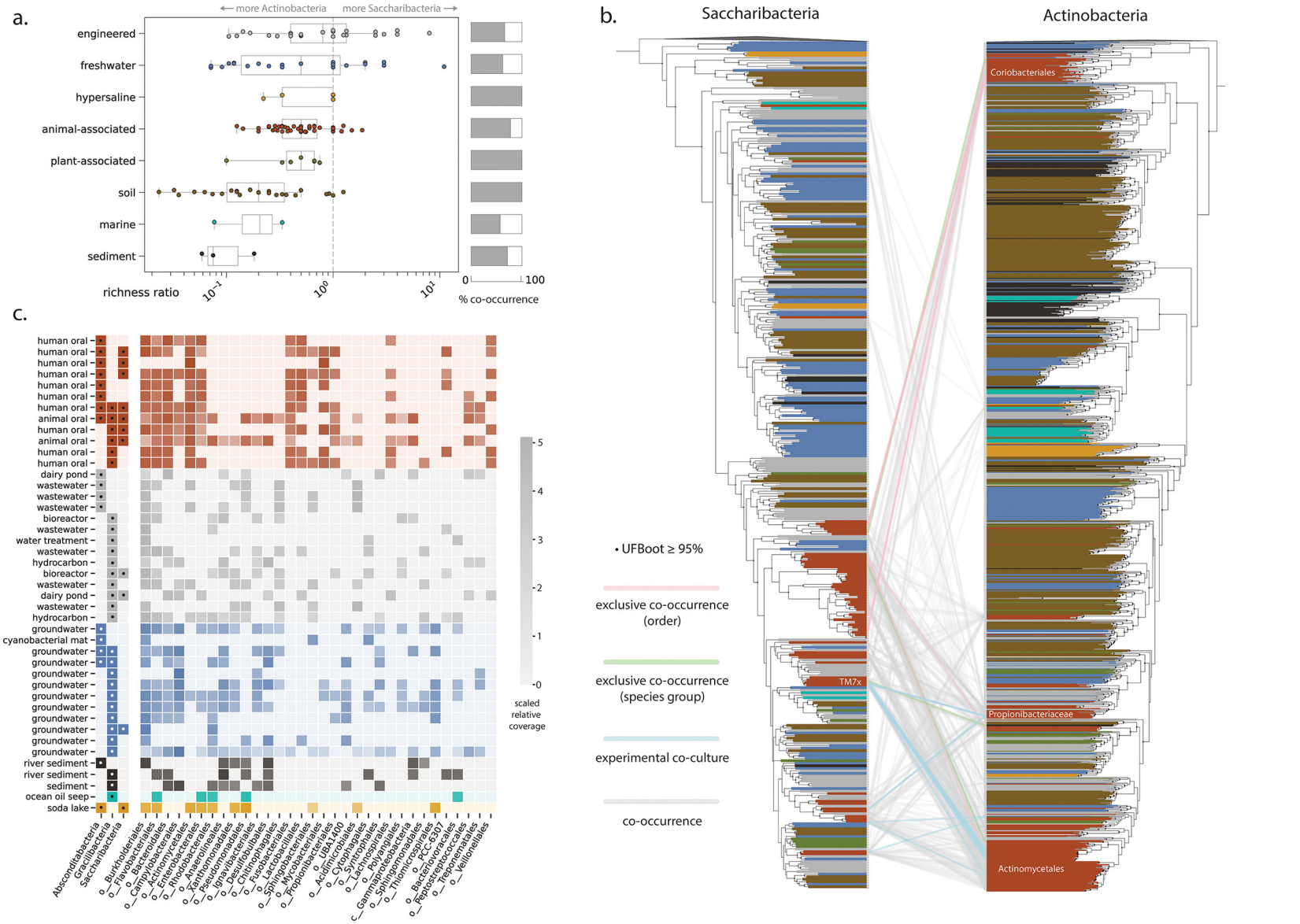


**FIG 2** Patterns of co-occurrence between CPR and potential host lineages across environments. (a) Relative richness ratio, describing the ratio of distinct Saccharibacteria species groups to Actinobacteria species groups, for each sample and overall co-occurrence percentage across habitat categories. (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.

species group and one Actinobacteria species group ("exclusive co-occurrence - species group," Fig. 2b; see also Table S3 at https://doi.org/10.5281/zenodo.4560554). Two of these cases involved Actinobacteria from the order Actinomycetales, from which multiple Saccharibacteria hosts have already been identified (85). 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 Coriobacteriales (see Table S3 at https://doi.org/10.5281/zenodo.4560554). 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; see also Table S3 at https://doi.org/10.5281/zenodo.4560554). Five of the seven instances involved pairs of Saccharibacteria and Coriobacteriales from termite and swine gut metagenomes. Thus, unlike in human oral environments, Coriobacteriales 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 (45).

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 (45) or the gammaproteobacterium *Halochromatium* in certain salt lakes (86). 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 (42) and one dairy pond sample containing Absconditabacteria (Fig. 2c; see also Table S5 at https://doi.org/10.5281/zenodo.4560554). No clear patterns of potential host co-occurrence were observed for Gracilibacteria, with the exception of the proteobacterial order Campylobacterales, which cooccurred 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 cooccurred 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 to 98% of relative abundance (34). To complement the above co-occurrence analyses, we reanalyzed 22 published metagenomes representing spatially separated mats and discovered that Absconditabacteria were detectable in 12 of them at various degrees of coverage (0.12× to 37×). As noted previously, also present in the mats were members of the phyla Bacteroidetes, Betaproteobacteria, and Verrucomicrobia (34). 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 10, involved Bacteroidetes (see Table S6 at https://doi.org/10.5281/zenodo.4560554).

**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 sizes 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 than 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

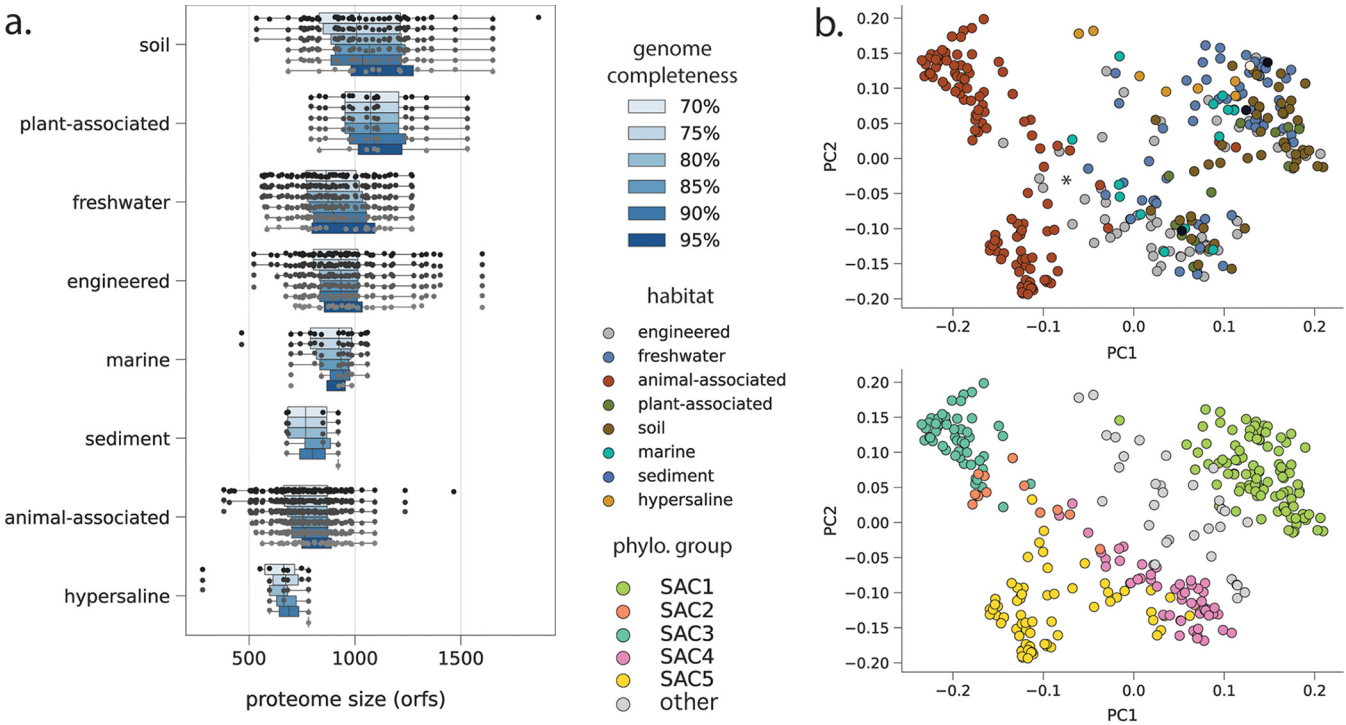


**FIG 3** Proteome characteristics for Saccharibacteria. (a) Predicted proteome size (open reading frame count) at increasing genome completeness thresholds. (b) Overall proteome similarity among Saccharibacteria from different habitat categories (top panel) and phylogenetic clades (bottom panel). PCoAs 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.

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 (see Fig. S1 at https://doi.org/10.5281/zenodo.4560554). 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 (12) (Materials and Methods; see also Table S7 at https://doi.org/10.5281/zenodo.4560554). Among Saccharibacteria, principal-coordinate 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 to -3 versus SAC4 and SAC5). We did not observe strong clustering of Saccharibacteria by specific environmental biome, consistent with the interspersed nature of their phylogenetic relationships (Fig. 1a and Fig. 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 (see Fig. S2 at https://doi.org/10.5281/zenodo.4560554). 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 from environmental counterparts.

To further examine the distinctions evident in the PCoA, 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 cooccurring 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 and Absconditabacteria and modules shared by both lineages but not present in Saccharibacteria, paralleling their phylogenetic relationships (Fig. 1a and Fig. 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 (31). 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 (87) (see Table S7 at https://doi.org/10.5281/zenodo.4560554). 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 bacteria. 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 bacteria (13): 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 of $\geq$5, whereas depleted families were defined as those that were encoded by $<$10% of genomes in a given habitat but $\geq$50% of genomes from other habitats. To account for the fact that small families might appear to be differentially distributed due to chance alone, we also stipulated that comparisons be statistically significant ($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 coenriched 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.

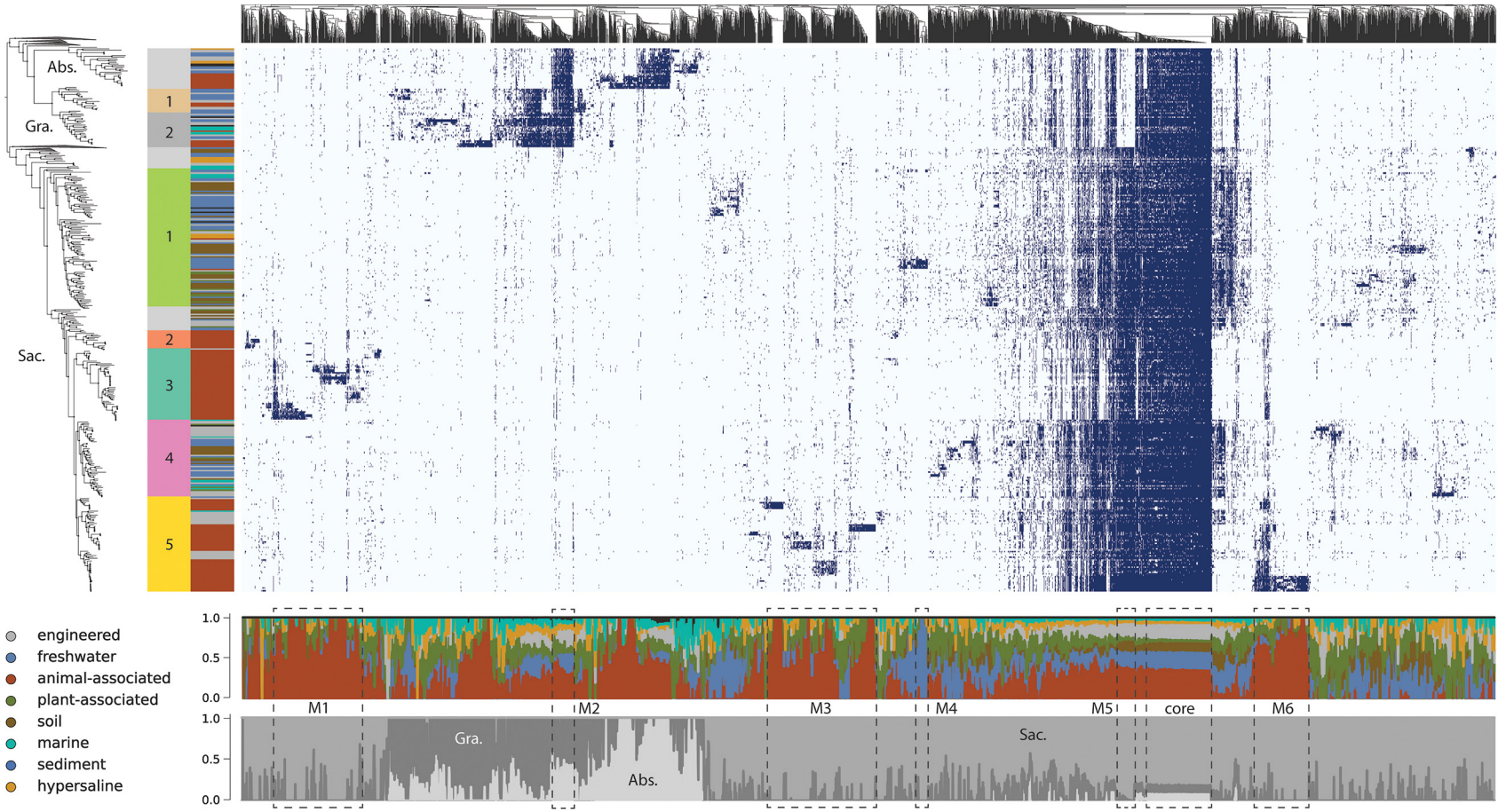


**FIG 4** Phylogenetic and environmental distribution of protein families recovered among CPR bacteria. (Upper panel) Presence/absence profiles for protein families with 5 or more members, with shaded cells indicating presence and light cells indicating absence. Columns represent protein families, hierarchically 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 dashed lines (M1 to -6 and "core").

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 to 58% of genomes across clades) (see Table S8 at https://doi.org/10.5281/zenodo.4560554). Two of these families, corresponding to a putative amino acid transport permease and substrate-binding protein (fam00393 and fam11477, respectively), were colocated in some genomes along with an ATP-binding protein (subset of fam00001), suggesting that they may function together to take up 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 (7), including phosphoglycerate mutase, glycogen phosphorylase, and a uracil-DNA glycosylase (ratio 8.3 to 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 (see Table S8 at https://doi.org/10.5281/zenodo.4560554) (7).

We identified multiple families that are either enriched or depleted in animal-associated Saccharibacteria that were functionally related to oxidative stress (see Table S8 at https://doi.org/10.5281/zenodo.4560554). 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 (88). 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 harboring each of the five genes. This operon has been previously suggested to confer an advantage in aerophilic environments like soil through detoxification (6) or use of oxygen (15, 16).

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) (62, 89). 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 (see Fig. S3 at https://doi.org/10.5281/zenodo.4560554), suggesting that they may function as proton pumps (90, 91). Distinct rhodopsin sequences were also recovered in the genomes of environmental Absconditabacteria (NDQ motif) and Gracilibacteria (DTE motif), although they were not statistically enriched (see Fig. S3 at https://doi.org/10.5281/zenodo.4560554). Genomes of soil-associated Saccharibacteria were enriched for nearly 130 protein families largely without strong functional annotations (Fig. 4; see also Table S8 at https://doi.org/10.5281/zenodo.4560554). 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 File S2 in the supplemental material at https://doi.org/10.5281/zenodo.4560554.

**Evolutionary processes shaping proteome evolution.** The observation that some differentially distributed traits among CPR bacteria 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 and 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 was then estimated using a probabilistic framework that takes into consideration genome incompleteness, variable rates of transfer and loss, and uncertainty in gene tree reconstruction (92, 93). The results of this analysis reveal relatively few instances of originations, defined as lateral transfer from outside the three lineages of CPR bacteria 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 and O2, Fig. 5; see also Table S1 at https://doi.org/10.5281/zenodo.4560554). On the other hand, in Saccharibacteria, originations were primarily associated with shallower subclades of animal-associated (and, in one case, freshwater) genomes (O3 to O6, Fig. 5; see also Table S1 at https://doi.org/10.5281/zenodo.4560554). 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; see also Table S1 at https://doi.org/10.5281/zenodo.4560554), suggesting that these genomes may be phylogenetic "hot spots" 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.

## 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 (2, 7). 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 (7). Alternatively, this pattern may reflect lineage-specific reversion 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 supports the idea that these CPR bacteria are likely legitimate members of the dolphin oral microbiome, rather than contamination (94).

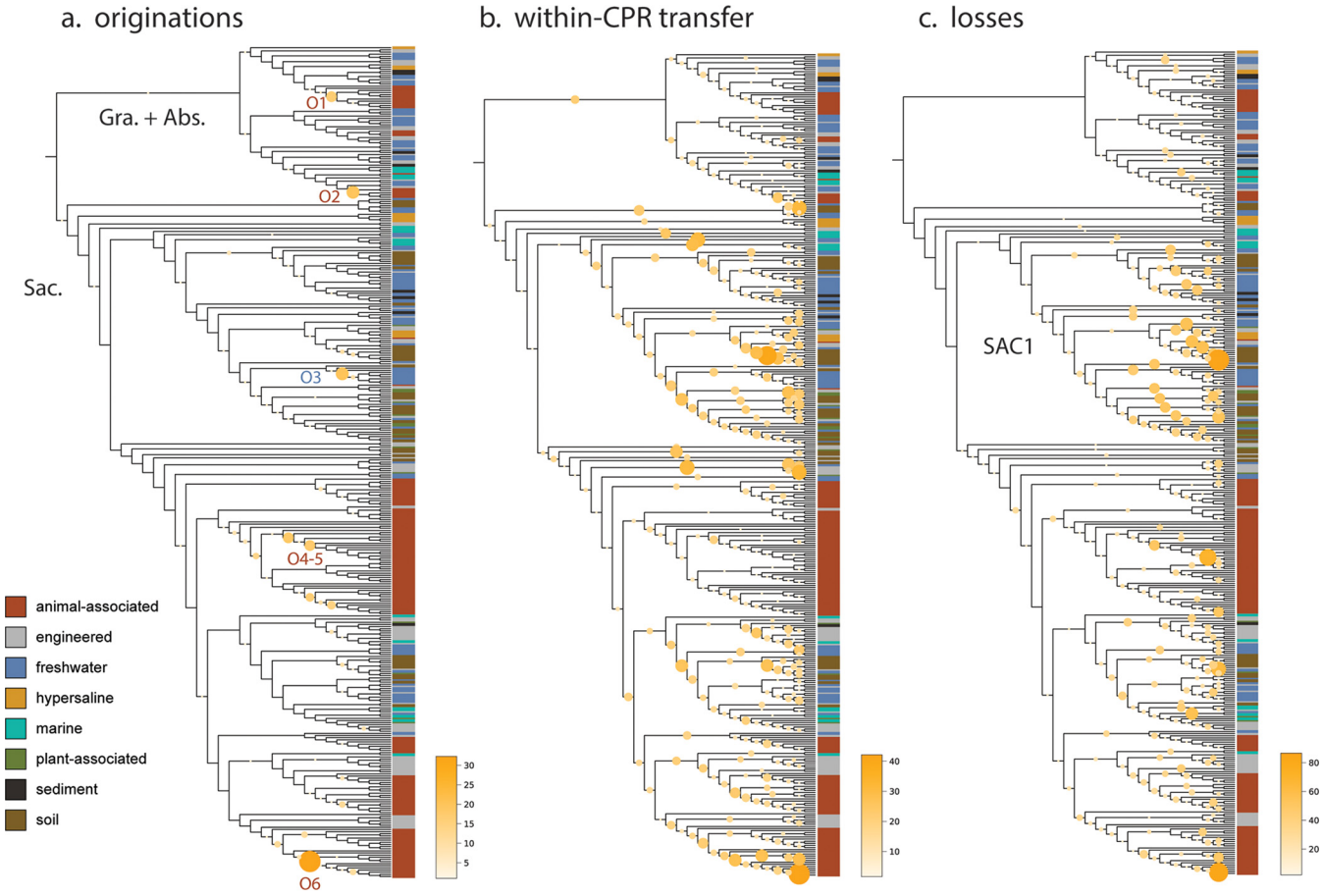


**FIG 5** Evolutionary processes shaping proteome evolution in three lineages of CPR bacteria. Each panel displays the species tree from Fig. 1a in cladogram format. The size and color of circles mapped onto interior branches represent the cumulative number of originations (defined as either lateral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

Currently, the mechanisms that enable environmental transition among CPR bacteria 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 bacteria involve comigration with their hosts or the acquisition of new hosts. The finding that single CPR species cooccur 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 (45). Supporting this, some laboratory strains of oral Saccharibacteria can adapt to new hosts after periods of living independently (95). The lack of evidence for lateral gene transfer between experimentally profiled pairs (7) 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 coculture of Saccharibacteria and two species of Gordonia (Actinobacteria) from wastewater foam (96). Similarly, changes in abundance over a sample series from a bioreactor system treating thiocyanate were recently used to suggest that *Microbacterium ginsengisoli* may serve as a host for a cooccurring Saccharibacteria bacterium (56). One Absconditabacteria lineage (*Vampirococcus*) has been predicted to have a host from the Gammaproteobacteria (86), and one Gracilibacteria was suggested to have a *Colwellia* host based on a shared repeat protein motif (14). Given the scant information about possible hosts for these CPR

bacteria, especially for Absconditabacteria and Gracilibacteria, the patterns of co-occurrence we report for specific organisms provide starting points for host identification via targeted coisolation.

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 bacteria 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 (17).

Only three lineages of CPR bacteria (of potentially dozens) have been consistently recovered in the animal-associated microbiome. Given the enormous diversity of CPR bacteria in drinking water (2), 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 (1, 2, 17).

It is also interesting to compare processes of habitat transition in CPR bacteria 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 (97) and intracellular symbionts of insects (98). However, in contrast to findings for Elusimicrobia, where host-associated lineages have common patterns of loss of metabolic capacities compared to relatives from nonhost environments (97), patterns of gene loss in animal-associated CPR bacteria appear to be heterogeneous and lineage specific. One possibility is that gene loss in CPR bacteria 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 bacteria over evolutionary time. As such, the processes shaping CPR lineage evolution share both similarities with and differences from those predicted for other microbes, including Haloarchaeota (99) and ammonia-oxidizing lineages of Thaumarchaeota (92, 100), 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 (7, 17) 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 cocultivate CPR bacteria 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.

## 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 (see Table S1 at https://doi.org/10.5281/zenodo.4560554). Assembly, annotation, and binning procedures followed those from Anantharaman et al. (33). 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 reference 101.

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 bacteria (31, 33) using the custom workflow in CheckM (102). 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 dereplicated at 99% ANI using dRep (-sa 0.99 -comp 70 -con 10) (103), yielding a set of 389 nonredundant genomes from a starting set of 868. Existing metadata were used to assign both "broad" and "narrow" habitat of origin for each nonredundant 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, are available in Table S1 at https://doi.org/10.5281/zenodo.4560554. All newly binned genomes are available through Zenodo (see "Data and software availability" below).

**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 (10, 11). Predicted proteins were concatenated and functionally annotated using kofamscan (104). Results with an E value of ≤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 (13). Sequences corresponding to each ribosomal protein were separately aligned with MAFFT and subsequently trimmed for phylogenetically informative regions using BMGE (*-m BLOSUM30*) (105). 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 (1,976 amino acids) set using ultrafast bootstrap and IQ-TREE's extended Free-Rate model selection (*-m MFP -st AA -bb 1000*) (106). The maximum-likelihood tree is available as File S1 at https://doi.org/10.5281/zenodo.4560554. The tree and associated metadata were visualized in iTOL (107) 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 cocultures 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 subsampled 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 coassembled) from it using bowtie2 (default parameters). For mappings with a nonzero number of read alignments, abundance of each genome was calculated by counting the number of stringently mapped reads (≥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 coassemblies 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 coassembled 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 reassembled using MEGAHIT (*–min-contig-len 1000*) and analyzed using GraftM (108) with a ribosomal protein S3 (rpS3) gpackage custom built from GTDB (release 05-RS95) (109). 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 (13) 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*) (110) instead of BMGE. Species groups that cooccurred in one or more metagenomic samples were then noted. If a given Saccharibacteria species group exclusively cooccurred 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 labeled. Finally, experimental cocultures 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 (see Table S4 at https://doi.org/10.5281/zenodo.4560554) 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 labeled 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 0.99*), 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 reanalysis of 22 cyanobacterial mat metagenomes (34), 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, and 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 an HMM-HMM comparison approach (–coverage 0.70) (12) (https://github.com/raphael-upmc/proteinClusteringPipeline). For each protein cluster, we recorded the most common KEGG annotation among its member sequences and the percentage 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 bacteria. First, we computed proteome size across habitats, defined as the number of predicted open reading frames (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-coordinate 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 under "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 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 (FDR) 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 $\geq$5 and depleted families to be 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 of $\leq$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 bacteria (fam11249) were combined with a set of reference protein sequences spanning type 1 bacterial/archaeal rhodopsin and heliorhodopsin (111). Sequences were then aligned using MAFFT (*–auto*), and a tree was inferred using IQ-TREE (-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 (112).

**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 reference 92. 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 were aligned with MAFFT (*–retree 2*). Resulting alignments were then trimmed using trimal (*-gt 0.1*) and used to infer maximum-likelihood phylogenetic trees using IQ-TREE with 1,000 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 wasthen used to probabilistically reconcile each protein family with the pruned species tree using the ALE package (*ALE_undated*) (93). 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 bacteria, or *de novo* gene formation), within-CPR horizontal transfers, and losses over each nonterminal branch and mapped branch-wise counts for each event to a species-tree cladogram in iTOL (107).

**Data and software availability.** All accession information for the genomes analyzed in this study is listed in Table S1 at https://doi.org/10.5281/zenodo.4560554. Genomes as well as custom code for the described analyses are also available on GitHub, https://github.com/alexanderjaffe/cpr-crossenv. All supplemental figures, tables, and files are available through Zenodo (https://doi.org/10.5281/zenodo.4560554).

## ACKNOWLEDGMENTS

We thank Shufei Lei, Lily Law, Alex Crits-Christoph, Tom Williams, Oded Béjà, Adair Borges, Raphaël Méheust, Alison Sharrar, Alexa Nicolas, Jett Liu, Basem Al-Shayeb, and Simonetta Gribaldo for informatics support, helpful discussions, and comments on the manuscript. We thank Ariel Amadio, Mircea Podar, Ramunas Stepanauskas, Connor Skennerton, Stefano Campanaro, Cédric Laczny, Paul Wilmes, Clara Chan, Scott E. Miller, Lauren C. Kennedy, Rose S. Kantor, Kara L. Nelson, Lauren Lui, Maliheh Mehrshad, Chris Greening, Mads Albertsen, and Sari Peura for permission to use genomic data that were unpublished at the time of writing. We also thank the Innovative Genomics Institute at UC Berkeley.

Christine He was funded by a Camille & Henry Dreyfus Environmental Chemistry Postdoctoral Fellowship. Patrick Munk was supported by the Danish Veterinary and Food Administration and The Novo Nordisk Foundation (NNF16OC0021856). The Japan Atomic Energy Agency (JAEA) was funded by the Ministry of Economy, Trade and Industry of Japan, as "The Project for Validating Near-field System Assessment Methodology in Geological Disposal System." Keith Bouma-Gregson and cyanobacterial mat sample collection were supported by the National Science Foundation's Eel River Critical Zone Observatory [EAR-1331940], Department of Energy grant [DOE-SC10010566], NSF Division of Environmental Biology [1656009], and US EPA STAR Fellowship [91767101-0]. Rohan Sachdeva and thiocyanate reactor genome construction were supported by a grant from the National Science Foundation (USA) to J.F.B. (EAR-1349278). Alex D. Thomas was

supported by the National Science Foundation Graduate Research Fellowship Program, and soil sample collection was funded by the Watershed Function Scientific Focus Area funded by the US Department of Energy, Office of Science, Office of Biological and Environmental Research [DE-AC02-05CH11231] with facility support from RMBL equipment grant Understanding Genetic Mechanisms [NSF DBI-1315705].

A.L.J. and J.F.B. compiled the data set, performed genome curation and analysis, developed the project, and wrote the manuscript. A.D.T., C.H., R.K., L.E.V.-A., P.M., K.B.-G., I.F.F., Y.A., R.S., and P.T.W. generated data for the study and provided comments on the manuscript.

## REFERENCES

1. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun 6:6372. https://doi.org/10.1038/ncomms7372.
2. He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, Banfield JF. 2021. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol 6:354–365. https://doi.org/10.1038/s41564-020-00840-5.
3. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, Dorrestein PC, Esquenazi E, Hunter RC, Cheng G, Nelson KE, Lux R, Shi W. 2015. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci U S A 112:244–249. https://doi.org/10.1073/pnas.1419038112.
4. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337:1661–1665. https://doi.org/10.1126/science.1224041.
5. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533–538. https://doi.org/10.1038/nbt.2579.
6. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF. 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio 4:e00708-13. https://doi.org/10.1128/mBio.00708-13.
7. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W, He X. 2020. Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to mammalian hosts. Cell Rep 32:107939. https://doi.org/10.1016/j.celrep.2020.107939.
8. Bor B, Bedree JK, Shi W, McLean JS, He X. 2019. Saccharibacteria (TM7) in the human oral microbiome. J Dent Res 98:500–509. https://doi.org/10.1177/0022034519831671.
9. Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson JA, Strausbaugh LD, Gamonal J, Diaz PI. 2013. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. ISME J 7:1016–1025. https://doi.org/10.1038/ismej.2012.174.
10. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Söll D, Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc Natl Acad Sci U S A 110:5540–5545. https://doi.org/10.1073/pnas.1303090110.
11. Hanke A, Hamann E, Sharma R, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R, Makinwa K, Hettich RL, Banfield JF, Tegetmeyer HE, Strous M. 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. Front Microbiol 5:231. https://doi.org/10.3389/fmicb.2014.00231.
12. Méheust R, Burstein D, Castelle CJ, Banfield JF. 2019. The distinction of CPR bacteria from other bacteria based on protein family content. Nat Commun 10:4173. https://doi.org/10.1038/s41467-019-12171-z.
13. Jaffe AL, Castelle CJ, Matheus Carnevali PB, Gribaldo S, Banfield JF. 2020. The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. BMC Biol 18:69. https://doi.org/10.1186/s12915-020-00804-5.
14. Sieber CMK, Paul BG, Castelle CJ, Hu P, Tringe SG, Valentine DL, Andersen GL, Banfield JF. 2019. Unusual metabolism and hypervariation in the genome of a gracilibacterium (Bd1-5) from an oil-degrading community. mBio 10:e02128-19. https://doi.org/10.1128/mBio.02128-19.
15. Starr EP, Shi S, Blazewicz SJ, Probst AJ, Herman DJ, Firestone MK, Banfield JF. 2018. Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived carbon. Microbiome 6:122. https://doi.org/10.1186/s40168-018-0499-z.
16. Nicolas AM, Jaffe AL, Nuccio EE, Taga ME, Firestone MK, Banfield JF. 2020. Unexpected diversity of CPR bacteria and nanoarchaea in the rare biosphere of rhizosphere-associated grassland soil. bioRxiv https://www.biorxiv.org/content/10.1101/2020.07.13.194282v1.abstract.
17. Shaiber A, Willis AD, Delmont TO, Roux S, Chen L-X, Schmid AC, Yousef M, Watson AR, Lolans K, Esen ÖC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren AM. 2020. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol 21:292. https://doi.org/10.1186/s13059-020-02195-w.
18. Murugkar PP, Collins AJ, Chen T, Dewhirst FE. 2020. Isolation and cultivation of candidate phyla radiation Saccharibacteria (TM7) bacteria in co-culture with bacterial hosts. J Oral Microbiol 12:1814666. https://doi.org/10.1080/20002297.2020.1814666.
19. Utter DR, He X, Cavanaugh CM, McLean JS, Bor B. 2020. The saccharibacterium TM7x elicits differential responses across its host range. ISME J 14:3054–3067. https://doi.org/10.1038/s41396-020-00736-6.
20. Lui LM, Nielsen TN, Arkin AP. 2021. A method for achieving complete microbial genomes and improving bins from metagenomics data. PLoS Comput Biol 17:e1008972. https://doi.org/10.1371/journal.pcbi.1008972.
21. Dueholm MS, Albertsen M, Stokholm-Bjerregaard M, McIlroy SJ, Karst SM, Nielsen PH. 2015. Complete genome sequence of the bacterium Aalborg_AAW-1, representing a novel family within the candidate phylum SR1. Genome Announc 3:e00624-15. https://doi.org/10.1128/genomeA.00624-15.
22. Ornaghi M, Prado RM, Ramos TR, Catalano FR, Mottin C, Creevey CJ, Huws SA, Prado IN. 2020. Natural plant-based additives can improve ruminant performance by influencing the rumen microbiome. Res Square https://assets.researchsquare.com/files/rs-29748/v1/83e12471-abdd-43c4-b07e-80e1449f0864.pdf.
23. Cabello-Yeves PJ, Zemskaya TI, Zakharenko AS, Sakirko MV, Ivanov VG, Ghai R, Rodriguez-Valera F. 2020. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. Limnol Oceanogr 65:1471–1488. https://doi.org/10.1002/lno.11401.
24. Dudek NK, Sun CL, Burstein D, Kantor RS, Aliaga Goltsman DS, Bik EM, Thomas BC, Banfield JF, Relman DA. 2017. Novel microbial diversity and functional potential in the marine mammal oral microbiome. Curr Biol 27:3752–3762.e6. https://doi.org/10.1016/j.cub.2017.10.040.
25. Andersen VD, Aarestrup FM, Munk P, Jensen MS, de Knegt LV, Bortolaia V, Knudsen BE, Lukjancenko O, Birkegård AC, Vigre H. 2020. Predicting effects of changed antimicrobial usage on the abundance of antimicrobial resistance genes in finisher' gut microbiomes. Prev Vet Med 174:104853. https://doi.org/10.1016/j.prevetmed.2019.104853.
26. Rehman ZU, Fortunato L, Cheng T, Leiknes T. 2020. Metagenomic analysis of sludge and early-stage biofilm communities of a submerged membrane bioreactor. Sci Total Environ 701:134682. https://doi.org/10.1016/j.scitotenv.2019.134682.
27. Youssef NH, Farag IF, Hahn CR, Premathilake H, Fry E, Hart M, Huffaker K, Bird E, Hambright J, Hoff WD, Elshahed MS. 2019. Candidatus Krumholzibacterium zodletonense gen. nov., sp nov, the first representative of the candidate phylum Krumholzibacteriota phyl. nov. recovered from an

anoxic sulfidic spring using genome resolved metagenomics. Syst Appl Microbiol 42:85–93. https://doi.org/10.1016/j.syapm.2018.11.002.
28. Poghosyan L, Koch H, Frank J, van Kessel MAHJ, Cremers G, van Alen T, Jetten MSM, Op den Camp HJM, Lücker S. 2020. Metagenomic profiling of ammonia- and methane-oxidizing microorganisms in two sequential rapid sand filters. Water Res 185:116288. https://doi.org/10.1016/j.watres.2020.116288.
29. Hernsdorf AW, Amano Y, Miyakawa K, Ise K, Suzuki Y, Anantharaman K, Probst A, Burstein D, Thomas BC, Banfield JF. 2017. Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. ISME J 11:1915–1929. https://doi.org/10.1038/ismej.2017.39.
30. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. https://doi.org/10.1038/nature12352.
31. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211. https://doi.org/10.1038/nature14486.
32. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229.
33. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219. https://doi.org/10.1038/ncomms13219.
34. Bouma-Gregson K, Olm MR, Probst AJ, Anantharaman K, Power ME, Banfield JF. 2019. Impacts of microbial assemblage and environmental conditions on the distribution of anatoxin-a producing cyanobacteria within a river network. ISME J 13:1618–1634. https://doi.org/10.1038/s41396-019-0374-3.
35. Engelberts JP, Robbins SJ, de Goeij JM, Aranda M, Bell SC, Webster NS. 2020. Characterization of a sponge microbiome using an integrative genome-centric approach. ISME J 14:1100–1110. https://doi.org/10.1038/s41396-020-0591-9.
36. Zhou Z, Tran PQ, Kieft K, Anantharaman K. 2020. Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation. ISME J 14:2060–2077. https://doi.org/10.1038/s41396-020-0669-4.
37. Pereira FC, Wasmund K, Cobankovic I, Jehmlich N, Herbold CW, Lee KS, Sziranyi B, Vesely C, Decker T, Stocker R, Warth B, von Bergen M, Wagner M, Berry D. 2020. Rational design of a microbial consortium of mucosal sugar utilizers reduces Clostridiodes difficile colonization. Nat Commun 11:5104. https://doi.org/10.1038/s41467-020-18928-1.
38. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
39. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CMK, Emerson JB, Anantharaman K, Thomas BC, Malmstrom RR, Stieglmeier M, Klingl A, Woyke T, Ryan MC, Banfield JF. 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nat Microbiol 3:328–336. https://doi.org/10.1038/s41564-017-0098-y.
40. Solden LM, Naas AE, Roux S, Daly RA, Collins WB, Nicora CD, Purvine SO, Hoyt DW, Schückel J, Jørgensen B, Willats W, Spalinger DE, Firkins JL, Lipton MS, Sullivan MB, Pope PB, Wrighton KC. 2018. Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. Nat Microbiol 3:1274–1284. https://doi.org/10.1038/s41564-018-0225-4.
41. Robbins SJ, Singleton CM, Chan CX, Messer LF, Geers AU, Ying H, Baker A, Bell SC, Morrow KM, Ragan MA, Miller DJ, Forêt S, ReFuGe2020 Consortium, Voolstra CR, Tyson GW, Bourne DG. 2019. A genomic view of the reef-building coral Porites lutea and its microbial symbionts. Nat Microbiol 4:2090–2100. https://doi.org/10.1038/s41564-019-0532-4.
42. Martínez Arbas S, Narayanasamy S, Herold M, Lebrun LA, Hoopmann MR, Li S, Lam TJ, Kunath BJ, Hicks ND, Liu CM, Price LB, Laczny CC, Gillece JD, Schupp JM, Keim PS, Moritz RL, Faust K, Tang H, Ye Y, Skupin A, May P, Muller EEL, Wilmes P. 2021. Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. Nat Microbiol 6:123–135. https://doi.org/10.1038/s41564-020-00794-8.
43. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, Hoelzle RD, Lamberton TO, McCalley CK, Hodgkins SB, Wilson RM, Purvine SO, Nicora CD, Li C, Frolking S, Chanton JP, Crill PM, Saleska SR, Rich VI, Tyson GW. 2018. Genome-centric view of carbon processing in thawing permafrost. Nature 560:49–54. https://doi.org/10.1038/s41586-018-0338-1.
44. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. Nature 568:505–510. https://doi.org/10.1038/s41586-019-1058-x.
45. Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, Heaton M, Joshi S, Klingeman D, Leys E, Yang Z, Parks JM, Podar M. 2019. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol 37:1314–1321. https://doi.org/10.1038/s41587-019-0260-6.
46. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39:105–114. https://doi.org/10.1038/s41587-020-0603-3.
47. Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, Huang B, Arodz TJ, Edupuganti L, Glascock AL, Xu J, Jimenez NR, Vivadelli SC, Fong SS, Sheth NU, Jean S, Lee V, Bokhari YA, Lara AM, Mistry SD, Duckworth RA, III, Bradley SP, Koparde VN, Orenda XV, Milton SH, Rozycki SK, Matveyev AV, Wright ML, Huzurbazar SV, Jackson EM, Smirnova E, Korlach J, Tsai Y-C, Dickinson MR, Brooks JL, Drake JI, Chaffin DO, Sexton AL, Gravett MG, Rubens CE, Wijesooriya NR, Hendricks-Muñoz KD, Jefferson KK, Strauss JF, III, Buck GA. 2019. The vaginal microbiome and preterm birth. Nat Med 25:1012–1021. https://doi.org/10.1038/s41591-019-0450-2.
48. Bandla A, Pavagadhi S, Sridhar Sudarshan A, Poh MCH, Swarup S. 2020. 910 metagenome-assembled genomes from the phytobiomes of three urban-farmed leafy Asian greens. Sci Data 7:278. https://doi.org/10.1038/s41597-020-00617-9.
49. Gibson KM, Nguyen BN, Neumann LM, Miller M, Buss P, Daniels S, Ahn MJ, Crandall KA, Pukazhenthi B. 2019. Gut microbiome differences between wild and captive black rhinoceros - implications for rhino health. Sci Rep 9:7570. https://doi.org/10.1038/s41598-019-43875-3.
50. Breister AM, Imam MA, Zhou Z, Ahsan MA, Noveron JC, Anantharaman K, Prabhakar P. 2020. Soil microbiomes mediate degradation of vinyl ester-based polymer composites. Commun Mater 1:101. https://doi.org/10.1038/s43246-020-00102-1.
51. Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:170203. https://doi.org/10.1038/sdata.2017.203.
52. Clayton JB, Vangay P, Huang H, Ward T, Hillmann BM, Al-Ghalith GA, Travis DA, Long HT, Van Tuan B, Van Minh V, Cabana F, Nadler T, Toddes B, Murphy T, Glander KE, Johnson TJ, Knights D. 2016. Captivity humanizes the primate microbiome. Proc Natl Acad Sci U S A 113:10376–10381. https://doi.org/10.1073/pnas.1521835113.
53. Hu P, Dubinsky EA, Probst AJ, Wang J, Sieber CMK, Tom LM, Gardinali PR, Banfield JF, Atlas RM, Andersen GL. 2017. Simulation of Deepwater Horizon oil plume reveals substrate specialization within a complex community of hydrocarbon degraders. Proc Natl Acad Sci U S A 114:7432–7437. https://doi.org/10.1073/pnas.1703424114.
54. Schulze-Makuch D, Wagner D, Kounaves SP, Mangelsdorf K, Devine KG, de Vera J-P, Schmitt-Kopplin P, Grossart H-P, Parro V, Kaupenjohann M, Galy A, Schneider B, Airo A, Frösler J, Davila AF, Arens FL, Cáceres L, Cornejo FS, Carrizo D, Dartnell L, DiRuggiero J, Flury M, Ganzert L, Gessner MO, Grathwohl P, Guan L, Heinz J, Hess M, Keppler F, Maus D, McKay CP, Meckenstock RU, Montgomery W, Oberlin EA, Probst AJ, Sáenz JS, Sattler T, Schirmack J, Sephton MA, Schloter M, Uhl J, Valenzuela B, Vestergaard G, Wörmer L, Zamorano P. 2018. Transitory microbial habitat in the hyperarid Atacama Desert. Proc Natl Acad Sci U S A 115:2670–2675. https://doi.org/10.1073/pnas.1714341115.
55. Munk P, Andersen VD, de Knegt L, Jensen MS, Knudsen BE, Lukjancenko O, Mordhorst H, Clasen J, Agersø Y, Folkesson A, Pamp SJ, Vigre H, Aarestrup FM. 2017. A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. J Antimicrob Chemother 72:385–392. https://doi.org/10.1093/jac/dkw415.
56. Huddy RJ, Sachdeva R, Kadzinga F, Kantor R, Harrison STL, Banfield JF. 2021. Thiocyanate and organic carbon inputs drive convergent selection for specific autotrophic Afipia and Thiobacillus strains within complex

microbiomes. Front Microbiol 12:643368. https://doi.org/10.3389/fmicb.2021.643368.
57. Kantor RS, van Zyl AW, van Hille RP, Thomas BC, Harrison STL, Banfield JF. 2015. Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. Environ Microbiol 17:4929–4941. https://doi.org/10.1111/1462-2920.12936.
58. Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC, Banfield JF. 2017. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. Environ Microbiol 19:459–474. https://doi.org/10.1111/1462-2920.13362.
59. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S-I. 2019. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. Environ Microbiol 21:4740–4754. https://doi.org/10.1111/1462-2920.14816.
60. Lemos LN, Medeiros JD, Dini-Andreote F, Fernandes GR, Varani AM, Oliveira G, Pylro VS. 2019. Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle. Mol Ecol 28:4259–4271. https://doi.org/10.1111/mec.15208.
61. Sharrar AM, Crits-Christoph A, Méheust R, Diamond S, Starr EP, Banfield JF. 2020. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. mBio 11:e00416-20. https://doi.org/10.1128/mBio.00416-20.
62. Zeng Y, Chen X, Madsen AM, Zervas A, Nielsen TK, Andrei A-S, Lund-Hansen LC, Liu Y, Hansen LH. 2020. Potential rhodopsin- and bacteriochlorophyll-based dual phototrophy in a high arctic glacier. mBio 11:e02641-20. https://doi.org/10.1128/mBio.02641-20.
63. Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, Goudeau D, Eichorst SA, Malmstrom RR, Bowers RM, Katz LA, Blanchard JL, Woyke T. 2020. Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. mSystems 5:e00768-19. https://doi.org/10.1128/mSystems.00768-19.
64. Vavourakis CD, Mehrshad M, Balkema C, van Hall R, Andrei A-S, Ghai R, Sorokin DY, Muyzer G. 2019. Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. BMC Biol 17:69. https://doi.org/10.1186/s12915-019-0688-7.
65. Campanaro S, Treu L, Rodriguez-R LM, Kovalovszki A, Ziels RM, Maus I, Zhu X, Kougias PG, Basile A, Luo G, Schlüter A, Konstantinidis KT, Angelidaki I. 2020. New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. Biotechnol Biofuels 13:25. https://doi.org/10.1186/s13068-020-01679-y.
66. Vavourakis CD, Andrei A-S, Mehrshad M, Ghai R, Sorokin DY, Muyzer G. 2018. A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. Microbiome 6:168. https://doi.org/10.1186/s40168-018-0548-7.
67. Wang W, Hu H, Zijlstra RT, Zheng J, Gänzle MG. 2019. Metagenomic reconstructions of gut microbial metabolism in weanling pigs. Microbiome 7:48. https://doi.org/10.1186/s40168-019-0662-1.
68. Tian R, Ning D, He Z, Zhang P, Spencer SJ, Gao S, Shi W, Wu L, Zhang Y, Yang Y, Adams BG, Rocha AM, Detienne BL, Lowe KA, Joyner DC, Klingeman DM, Arkin AP, Fields MW, Hazen TC, Stahl DA, Alm EJ, Zhou J. 2020. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. Microbiome 8:51. https://doi.org/10.1186/s40168-020-00825-w.
69. Keren R, Lawrence JE, Zhuang W, Jenkins D, Banfield JF, Alvarez-Cohen L, Zhou L, Yu K. 2020. Increased replication of dissimilatory nitrate-reducing bacteria leads to decreased anammox bioreactor performance. Microbiome 8:7. https://doi.org/10.1186/s40168-020-0786-3.
70. Cao Y, Xu H, Li R, Gao S, Chen N, Luo J, Jiang Y. 2019. Genetic basis of phenotypic differences between Chinese Yunling black goats and Nubian goats revealed by allele-specific expression in their F1 hybrids. Front Genet 10:145. https://doi.org/10.3389/fgene.2019.00145.
71. Finstad KM, Probst AJ, Thomas BC, Andersen GL, Demergasso C, Echeverría A, Amundson RG, Banfield JF. 2017. Microbial community structure and the persistence of cyanobacterial populations in salt crusts of the hyperarid Atacama Desert from genome-resolved metagenomics. Front Microbiol 8:1435. https://doi.org/10.3389/fmicb.2017.01435.
72. Kantor RS, Miller SE, Nelson KL. 2019. The water microbiome through a pilot scale advanced treatment facility for direct potable reuse. Front Microbiol 10:993. https://doi.org/10.3389/fmicb.2019.00993.
73. Beam JP, Becraft ED, Brown JM, Schulz F, Jarett JK, Bezuidt O, Poulton NJ, Clark K, Dunfield PF, Ravin NV, Spear JR, Hedlund BP, Kormas KA, Sievert SM, Elshahed MS, Barton HA, Stott MB, Eisen JA, Moser DP, Onstott TC, Woyke T, Stepanauskas R. 2020. Ancestral absence of electron transport chains in Patescibacteria and DPANN. Front Microbiol 11:1848. https://doi.org/10.3389/fmicb.2020.01848.
74. Tung J, Barreiro LB, Burns MB, Grenier J-C, Lynch J, Grieneisen LE, Altmann J, Alberts SC, Blekhman R, Archie EA. 2015. Social networks predict gut microbiome composition in wild baboons. Elife 4:e05224. https://doi.org/10.7554/eLife.05224.
75. Hervé V, Liu P, Dietrich C, Sillam-Dussès D, Stiblik P, Šobotník J, Brune A. 2020. Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. PeerJ 8:e8614. https://doi.org/10.7717/peerj.8614.
76. Skennerton C. 2013. Phage-host evolution in a model ecosystem. PhD thesis. School of Chemical Engineering, The University of Queensland, Brisbane, Australia.
77. Espinoza JL, Harkins DM, Torralba M, Gomez A, Highlander SK, Jones MB, Leong P, Saffery R, Bockmann M, Kuelbs C, Inman JM, Hughes T, Craig JM, Nelson KE, Dupont CL. 2018. Supragingival plaque microbiome ecology and functional potential in the context of health and disease. mBio 9:e01631-18. https://doi.org/10.1128/mBio.01631-18.
78. Stamps BW, Spear JR. 2020. Identification of metagenome-assembled genomes containing antimicrobial resistance genes, isolated from an advanced water treatment facility. Microbiol Resour Announc 9:e00003-20. https://doi.org/10.1128/MRA.00003-20.
79. Zhou Z, Liu Y, Xu W, Pan J, Luo Z-H, Li M. 2020. Genome- and community-level interaction insights into carbon utilization and element cycling functions of *Hydrothermarchaeota* in hydrothermal sediment. mSystems 5:e00795-19. https://doi.org/10.1128/mSystems.00795-19.
80. Mehrshad M, Lopez-Fernandez M, Sundh J, Bell E, Simone D, Buck M, Bernier-Latmani R, Bertilsson S, Dopson M. 2020. Energy efficiency and biological interactions define the core microbiome of deep oligotrophic groundwater. bioRxiv https://www.biorxiv.org/content/10.1101/2020.05.24.111179v1.abstract.
81. Ortiz M, Leung PM, Shelley G, Van Goethem MW, Bay SK, Jordaan K, Vikram S, Hogg ID, Makhalanyane TP, Chown SL, Grinter R, Cowan DA, Greening C. 2020. A genome compendium reveals diverse metabolic adaptations of Antarctic soil microorganisms. bioRxiv https://www.biorxiv.org/content/10.1101/2020.08.06.239558v1.abstract.
82. Buck M, Garcia SL, Fernandez L, Martin G, Martinez-Rodriguez GA, Saarenheimo J, Zopfi J, Bertilsson S, Peura S. 2021. Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. Sci Data 8:131. https://doi.org/10.1038/s41597-021-00910-1.
83. Shaiber A, Eren AM. 2019. Composite metagenome-assembled genomes reduce the quality of public genome repositories. mBio 10:e00725-19. https://doi.org/10.1128/mBio.00725-19.
84. Soro V, Dutton LC, Sprague SV, Nobbs AH, Ireland AJ, Sandy JR, Jepson MA, Micaroni M, Splatt PR, Dymock D, Jenkinson HF. 2014. Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria. Appl Environ Microbiol 80:6480–6489. https://doi.org/10.1128/AEM.01827-14.
85. Bor B, Collins AJ, Murugkar PP, Balasubramanian S, To TT, Hendrickson EL, Bedree JK, Bidlack FB, Johnston CD, Shi W, McLean JS, He X, Dewhirst FE. 2020. Insights obtained by culturing saccharibacteria with their bacterial hosts. J Dent Res 99:685–694. https://doi.org/10.1177/0022034520905792.
86. Moreira D, Zivanovic Y, López-Archilla AI, Iniesto M, López-García P. 2021. Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. Nat Commun 12:2454. https://doi.org/10.1038/s41467-021-22762-4.
87. Yamamoto T, Iino H, Kim K, Kuramitsu S, Fukui K. 2011. Evidence for ATP-dependent structural rearrangement of nuclease catalytic site in DNA mismatch repair endonuclease MutL. J Biol Chem 286:42337–42348. https://doi.org/10.1074/jbc.M111.277335.
88. Cardenas JP, Quatrini R, Holmes DS. 2016. Aerobic lineage of the oxidative stress response protein rubrerythrin emerged in an ancient microaerobic, (hyper)thermophilic environment. Front Microbiol 7:1822. https://doi.org/10.3389/fmicb.2016.01822.
89. Rissanen AJ, Saarela T, Jäntti H, Buck M, Peura S, Aalto SL, Ojala A, Pumpanen J, Tiirola M, Elvert M, Nykänen H. 2021. Vertical stratification patterns of methanotrophs and their genetic controllers in water columns of oxygen-stratified boreal lakes. FEMS Microbiol Ecol 97:fiaa252. https://doi.org/10.1093/femsec/fiaa252.
90. Maliar N, Okhrimenko IS, Petrovskaya LE, Alekseev AA, Kovalev KV, Soloviov DV, Popov PA, Rokitskaya TI, Antonenko YN, Zabelskii DV, Dolgikh DA, Kirpichnikov MP, Gordeliy VI. 2020. Novel pH-sensitive

microbial rhodopsin from Sphingomonas paucimobilis. Dokl Biochem Biophys 495:342–346. https://doi.org/10.1134/S1607672920060162.
91. Béjà O, Lanyi JK. 2014. Nature's toolkit for microbial rhodopsin ion pumps. Proc Natl Acad Sci U S A 111:6538–6539. https://doi.org/10.1073/pnas.1405093111.
92. Sheridan PO, Raguideau S, Quince C, Holden J, Zhang L, Thames Consortium, Williams TA, Gubry-Rangin C. 2020. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat Commun 11:5494. https://doi.org/10.1038/s41467-020-19132-x.
93. Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. Syst Biol 62:901–912. https://doi.org/10.1093/sysbio/syt054.
94. Bik EM, Costello EK, Switzer AD, Callahan BJ, Holmes SP, Wells RS, Carlin KP, Jensen ED, Venn-Watson S, Relman DA. 2016. Marine mammals harbor unique microbiotas shaped by and yet distinct from the sea. Nat Commun 7:10516. https://doi.org/10.1038/ncomms10516.
95. Bor B, McLean JS, Foster KR, Cen L, To TT, Serrato-Guillen A, Dewhirst FE, Shi W, He X. 2018. Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. Proc Natl Acad Sci U S A 115:12277–12282. https://doi.org/10.1073/pnas.1810625115.
96. Batinovic S, Rose JJA, Ratcliffe J, Seviour RJ, Petrovski S. 2021. Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams. Nat Microbiol 6:703–711. https://doi.org/10.1038/s41564-021-00892-1.
97. Méheust R, Castelle CJ, Matheus Carnevali PB, Farag IF, He C, Chen L-X, Amano Y, Hug LA, Banfield JF. 2020. Groundwater Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have a novel nitrogenase paralog. ISME J 14:2907–2922. https://doi.org/10.1038/s41396-020-0716-1.
98. McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26. https://doi.org/10.1038/nrmicro2670.
99. Martijn J, Schön ME, Lind AE, Vosseberg J, Williams TA, Spang A, Ettema TJG. 2020. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. Nat Commun 11:5490. https://doi.org/10.1038/s41467-020-19200-2.
100. Abby SS, Kerou M, Schleper C. 2020. Ancestral reconstructions decipher major adaptations of ammonia-oxidizing archaea upon radiation into moderate terrestrial and marine environments. mBio 11:e02371-20. https://doi.org/10.1128/mBio.02371-20.
101. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. Genome Res 30:315–333. https://doi.org/10.1101/gr.258640.119.
102. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
103. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126.
104. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36:2251–2252. https://doi.org/10.1093/bioinformatics/btz859.
105. Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210. https://doi.org/10.1186/1471-2148-10-210.
106. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
107. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290.
108. Boyd JA, Woodcroft BJ, Tyson GW. 2018. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. Nucleic Acids Res 46:e59. https://doi.org/10.1093/nar/gky174.
109. Crits-Christoph A, Diamond S, Al-Shayeb B, Valentin-Alvarado L, Banfield JF. 2021. A widely distributed genus of soil Acidobacteria genomically enriched in biosynthetic gene clusters. bioRxiv https://www.biorxiv.org/content/10.1101/2021.05.10.443473v2.abstract.
110. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
111. Pushkarev A, Inoue K, Larom S, Flores-Uribe J, Singh M, Konno M, Tomida S, Ito S, Nakamura R, Tsunoda SP, Philosof A, Sharon I, Yutin N, Koonin EV, Kandori H, Béjà O. 2018. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. Nature 558:595–599. https://doi.org/10.1038/s41586-018-0225-9.
112. Hasegawa M, Hosaka T, Kojima K, Nishimura Y, Nakajima Y, Kimura-Someya T, Shirouzu M, Sudo Y, Yoshizawa S. 2020. A unique clade of light-driven proton-pumping rhodopsins evolved in the cyanobacterial lineage. Sci Rep 10:16752. https://doi.org/10.1038/s41598-020-73606-y.

**SUPPLEMENTARY FIGURES:**

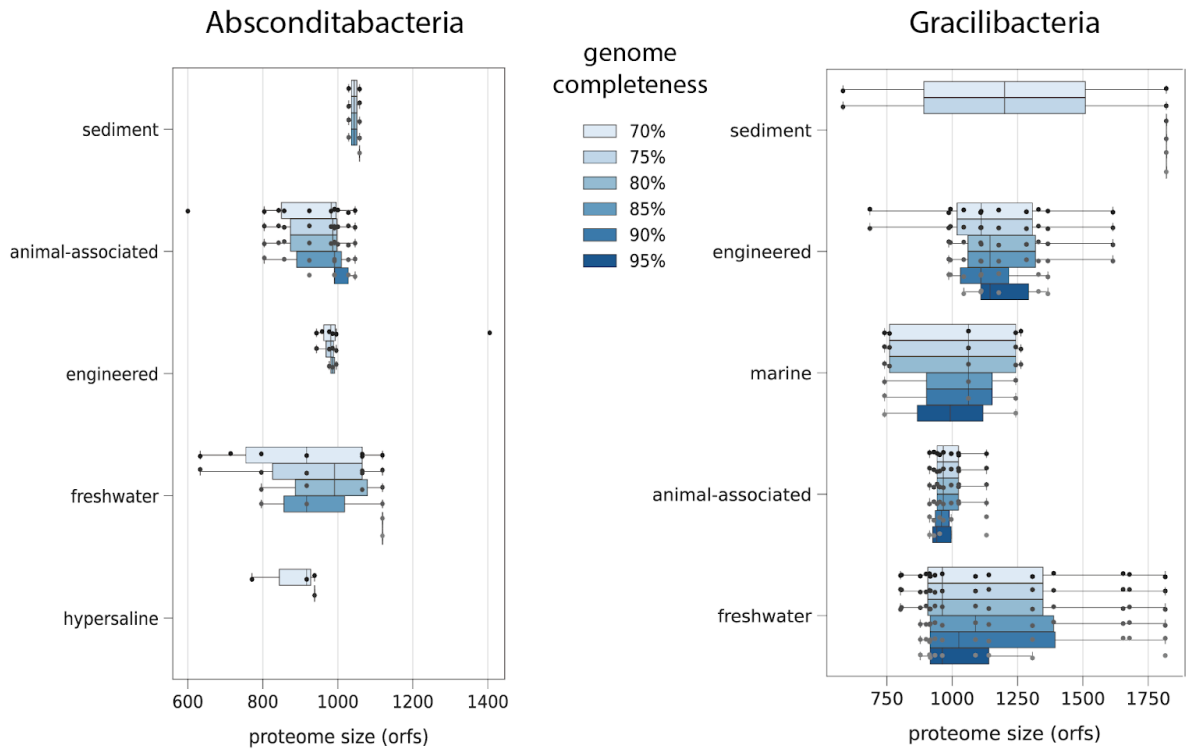


**Fig. S1.** Proteome size as a function of genome completeness and habitat of origin for Absconditabacteria and Gracilibacteria.

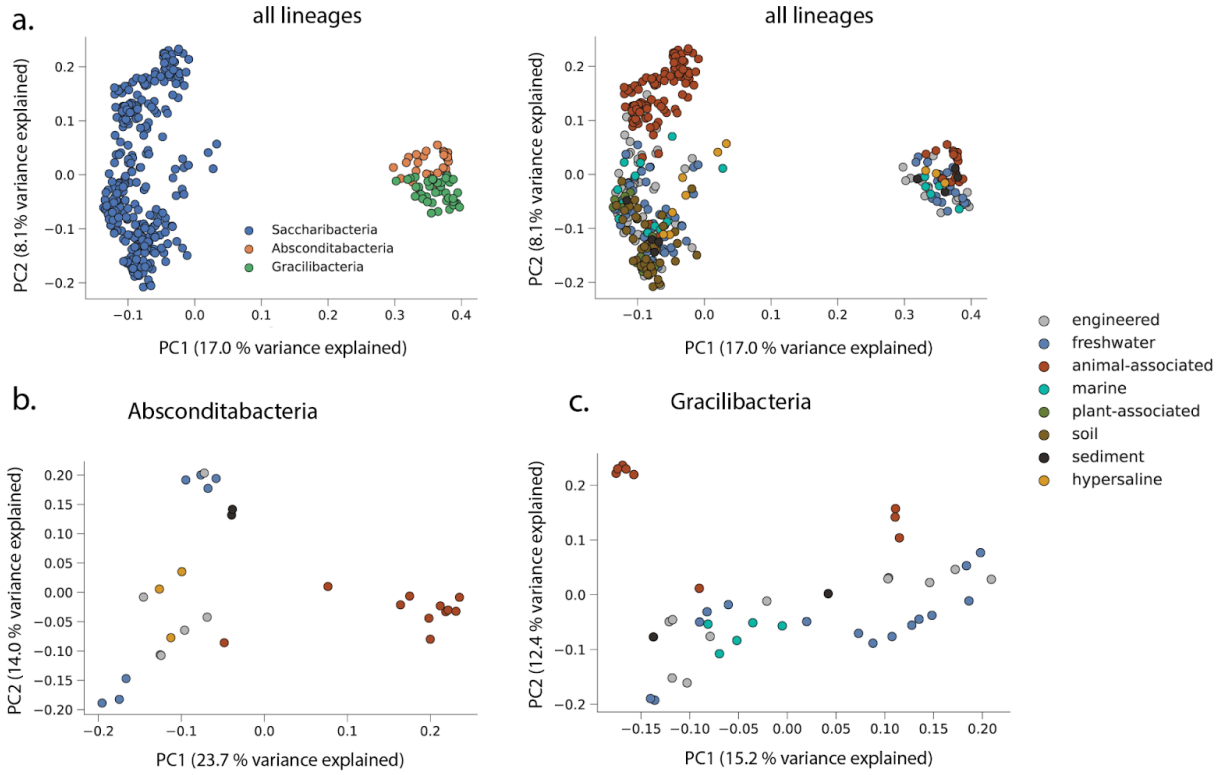


**Fig. S2.** Principal coordinates analysis based on all protein families with 5 or more members among **a)** all lineages, **b)** Absconditabacteria, and **c)** Gracilibacteria.

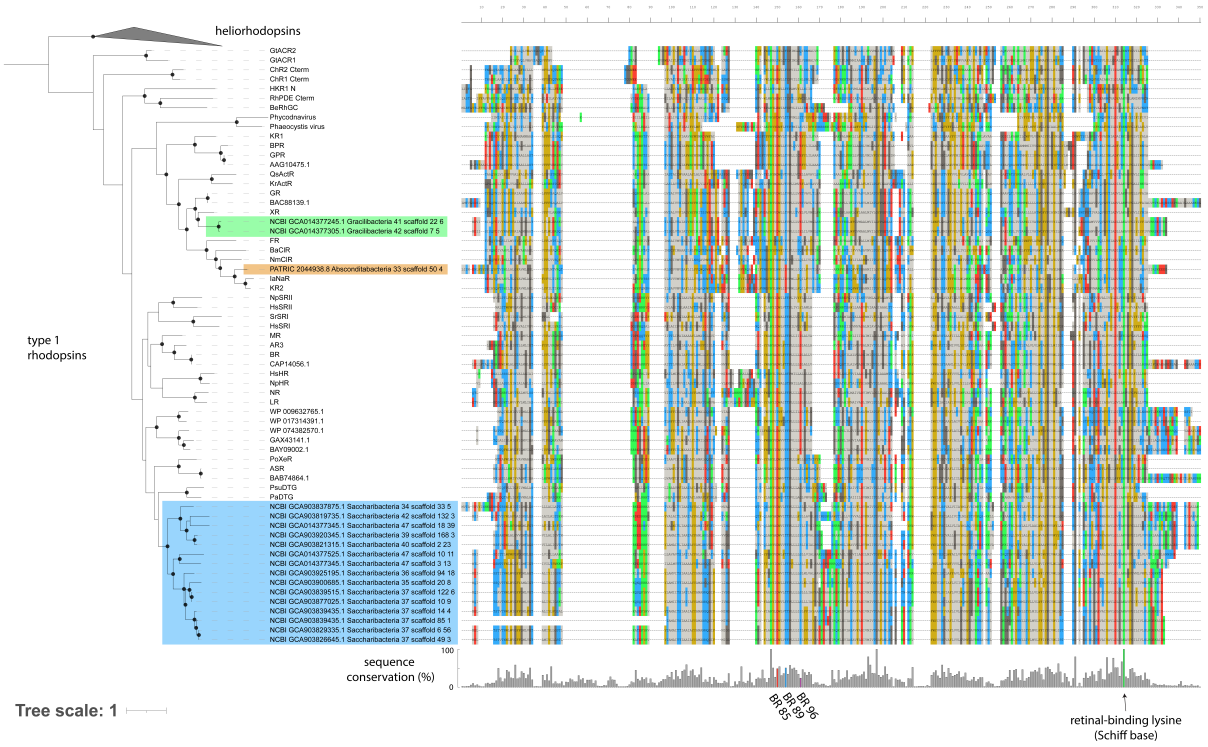


**Fig. S3.** Phylogenetic relationships and trimmed protein alignment among Type 1 rhodopsins and CPR rhodopsin homologs. Black dots indicate tree nodes with ≥ 95% ultrafast bootstrap support. CPR sequences from the Absconditabacteria (NDQ motif at the three labelled bacteriorhodopsin (BR) reference sites), Gracilibacteria (DTE motif), and Saccharibacteria (DTS motif) are highlighted in the tree. Sequence conservation at each aligned site, the location of bacteriorhodopsin (BR) site 85, 89, 96, and the location of retinal-binding lysine (Schiff Base linkage) are also indicated.