Title

Ecological and genomic analyses of candidate phylum WPS-2 bacteria in an unvegetated soil

Journal

Environmental Microbiology, 22(8)

ISSN

1462-2912

Authors

Sheremet, Andriy

Jones, Gareth M

Jarett, Jessica

et al.

Publication Date

2020-08-01

DOI

10.1111/1462-2920.15054

Peer reviewed

Ecological and genomic analyses of candidate phylum WPS-2 bacteria in an unvegetated soil

Andriy Sheremet[†], Gareth M. Jones[†], Jessica Jarett[2], Robert M. Bowers[2], Isaac Bedard[1], Cassandra Culham[1], Emiley A. Eloe-Fadrosh[2], Natalia Ivanova[2], Rex R. Malmstrom[2], Stephen E. Grasby[3], Tanja Woyke[2], and Peter F. Dunfield*[1]

[1] Department of Biological Sciences, University of Calgary, 2500 University Dr. NW Calgary AB Canada T2N 1N4

[2] Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA.

[3] Geological Survey of Canada, Calgary, Alberta T2L 2A7, Canada

*Corresponding author: Peter F. Dunfield, Department of Biological Sciences, University of Calgary, 2500 University Dr. NW Calgary AB Canada T2N 1N4, pfdunfie@ucalgary.ca

Running title: Organoheterotrophic WPS-2 bacteria

Keywords: WPS-2, Eremiobacteraeota, acidophile, uncultured bacteria

Supplementary File 1: Supplementary Tables and Figures

Supplementary File 2: Reference genomes and genome predictions

Originality-Significance Statement

The manuscript describes the characterization of bacteria belonging to an uncultured phylum (WPS-2 or Eremiobacteria), using a combination of methods including community analysis, quantitative PCR of environmental samples, single cell genomics and metagenomics. We qualitatively and quantitatively analysed communities in a large sample set taken across different microhabitats of an area affected by iron-sulfur springs. The data revealed a preference of WPS-2 and its associated community network to bare rather than vegetated soils. We then assembled genomes using both SAG and MAG techniques. Cross-referencing of SAGs and MAGs increased the reliability of genomic inferences. We propose that most WPS-2 bacteria in our site are efficient heterotrophic scavengers. These bacteria are therefore physiologically distinct from previously described members of the WPS-2 phylum, which have been proposed to be lithoautotrophs and photoautotrophs. In combination with previous studies, these data suggest that phylum WPS-2 includes bacteria with diverse metabolic capabilities. Our combination of ecogeographic community analysis and genome inference provides a detailed analysis for a member of this uncultured bacterial phylum.

Summary

Members of the bacterial candidate phylum WPS-2 (or Eremiobacteraeota) are abundant in several dry, bare soil environments. In a bare soil deposited by an extinct iron-sulfur spring, we found that WPS-2 comprised up to 24% of the bacterial community and up to 10^8 cells per g of soil based on 16S rRNA gene sequencing and quantification. A single genus-level cluster (*Ca.* Rubrimentiphilum) predominated in bare soils, but was less abundant in adjacent forest soils. Nearly complete genomes of *Ca.* Rubrimentiphilum were recovered as single amplified genomes (SAGs) and metagenome-assembled genomes (MAGs). Surprisingly, given the abundance of WPS-2 in bare soils, the genomes did not indicate any capacity for autotrophy, phototrophy, or trace gas metabolism. Genomic analysis instead suggesting a predominantly aerobic organoheterotrophic lifestyle, perhaps based on parasitizing or scavenging amino acids, nucleotides, and complex oligopeptides, along with lithotrophic capacity on thiosulfate. Other notable features included many genes encoding resistance to antimicrobial compounds. Network analyses of the entire community showed that some species of *Chloroflexi*, *Actinobacteria*, and candidate phylum AD3 (or Dormibacterota) co-occurred strongly with *Ca.* Rubrimentiphilum, and may represent ecological or metabolic partners. We propose that *Ca.* Rubrimentiphilum act as efficient heterotrophic scavengers in the site. In combination with previous studies, these data suggest that the phylum WPS-2 includes bacteria with diverse metabolic capabilities.

Introduction

There are between 112 and 1500 main lineages or phyla within the Domain *Bacteria*, depending on the criteria used for definition (Yarza *et al.*, 2014; Parks *et al.*, 2018). However all estimates agree that most bacterial phyla have no cultured members. Recent advances in high-throughput sequencing, single-cell sorting, and bioinformatics have facilitated genomic reconstructions of individual bacteria from many of these uncultivated phyla (Rinke *et al.*, 2013; Eloe-Fadrosh *et al.*, 2016; Hug *et al.*, 2016; Parks *et al.*, 2017). There are two principal approaches to recovering individual genomes of uncultured bacteria. Individual cells can be sorted out of complex microbial communities and then subjected to whole genome amplification and shotgun sequencing yielding single amplified genomes (SAGs); or individual genomes can be computationally separated from a complex metagenome using binning approaches based on composition, coverage, and other contig properties, yielding metagenome assembled genomes (MAGs). These techniques have already proven useful in describing uncultured phyla. The Genomic Encyclopedia of Bacteria and Archaea Microbial Dark Matter project provided substantial SAG data for over 20 candidate phyla, and in doing so identified a novel purine synthesis pathway and variants in the genetic code (Rinke *et al.*, 2013). MAG studies have been utilized to describe candidate phyla from contaminated aquifers (Wrighton *et al.*, 2012; Hug *et al.*, 2016), deep terrestrial biospheres (Wu *et al.*, 2016), permafrost soils (Taş *et al.*, 2014), and other sites (Parks *et al.*, 2017; Holland-Moritz *et al.*, 2018). Drawbacks of these approaches include incomplete genomes, genome contamination, and the difficulty of predicting phenotypes based on annotation alone. Predicting phenotypes is especially problematic with candidate phyla, which are often evolutionarily deeply rooted groups that often contain many hypothetical genes

and/or genes with low sequence homology to characterized reference genes. Nevertheless, in combination with ecological data these are powerful approaches to characterizing new microbial groups.

The Paint Pots, located in British Columbia, Canada, is a system of naturally occurring acidic iron-sulfur springs with high concentrations of heavy metals (Grasby *et al.*, 2013). It is an important site to the indigenous Ktunaxa people, and a tourist destination in Kootenay National Park. Geological studies have shown that as the spring precipitates iron oxide at its outlet, the hydraulic pressure increases, eventually causing the spring source to migrate and leaving a relic spring feature behind. Although the site has geochemical similarities to acid mine drainage (AMD), the microbial community is not typical of AMD (Grasby *et al.*, 2013). Instead, the community is more similar to communities in natural acidic soils, with one peculiarity being an elevated abundance of the candidate phylum WPS-2 in some areas.

WPS-2 is recognizable as a phylum-level grouping in most bacterial taxonomy databases and classification systems (Parks *et al.*, 2018). Ji *et al.* (2017) have suggested the alternate name Eremiobacteraeota for WPS-2. Since neither name has validity in official nomenclature, we will primarily use “WPS-2” here due do its precedence. WPS-2 bacteria were originally detected in polychlorinated biphenyl (PCB) polluted soil (Nogales *et al.*, 2001), and later in other soils (Costello *et al.*, 2009; Lin *et al.*, 2012; Grasby *et al.*, 2013; Pascual *et al.*, 2016; Hermans *et al.*, 2017), various temperate to Arctic peatlands (Bragina *et al.*, 2015; Holland-Moritz *et al.*, 2018; Woodcroft *et al.*, 2018), gas-producing shale (Trexler *et al.*, 2014), the

canine oral microbiome (Dewhirst *et al.*, 2012; Camanocha and Dewhirst, 2014). They are abundant in several organic-poor soil environments such as bare Antarctic soils (Ji *et al.*, 2016), Arctic cryoconite (Stibal *et al.*, 2015), extremely dry volcanic soils (Costello *et al.*, 2009), and bare metal-contaminated soils (Grasby *et al.*, 2013). WPS-2 MAGs were recently described from bare soils in Antarctica and the potential ecological role of these bacteria in autotrophic CO_2 fixation and scavenging of atmospheric H_2 was proposed (Ji *et al.*, 2016). MAGs of WPS-2 bacteria associated with boreal mosses also contained genes encoding the Calvin Cycle, as well as key genes for bacteriochlorophyll-based anoxygenic photosynthesis (Holland-Moritz *et al.*, 2018; Ward *et al.*, 2019).

The unusually high relative abundance of WPS-2 at the Paint Pots location, higher than reported anywhere else, provides an ideal situation for studying members of this candidate phylum, and simplifies genome recovery. In this study we performed a detailed environmental survey of microbial communities, and WPS-2 specifically, throughout the site. Forested sites adjacent to the bare soils served as a control for assessing the role of vegetation and soil organic matter in shaping the bacterial communities. MAG and SAGs were constructed to predict the physiology of the WPS-2 bacteria.

Results

Environmental distribution of candidate phylum WPS-2

Based on 16S rRNA gene amplicon sequencing, microbial communities in vegetated soils of the experimental area generally had low relative abundances of WPS-2 (<2.5 % of total reads), while bare ochre-coloured soils supported higher relative

abundances of the phylum (Figure 1; Supplementary Table 1). In transects, WPS-2 relative abundance increased from the forest edge outwards into the bare Mound soil, reaching maxima of 15%-24% (Figure 1, Supplementary Figure 1).

Multiple 16S rRNA gene-based OTUs of WPS-2 were detected, which phylogenetically fell into two distinct clusters (Supplementary Figure 2). The primary cluster we denote as *Ca.* Rubrimentiphilales, and its most abundant OTU as *Ca.* Rubrimentiphilum (from L. neut. n. rubrimentum, red ink; N.L. neut. adj. philum [from Gr. neut. adj. philon], friend, loving). The *Ca.* Rubrimentiphilales bacteria comprised 72% of the WPS-2 16S rRNA gene reads, on average, over all samples collected (Supplementary Table 1). The other WPS-2 cluster we refer to as AS-11 based on Silva (Supplementary Figure 2).

The absolute abundance of bacteria belonging to the *Ca.* Rubrimentiphilales cluster, based on a specific qPCR assay, was significantly higher in bare soils than in vegetated soils (Figure 2A). This higher absolute abundance was measured in bare soils despite the fact that significantly less DNA was recoverable, indicating a lower overall bacterial load (Figure 2A). The relative abundance of *Ca.* Rubrimentiphilales (based on its share of reads in 16S rRNA amplicons) was positively correlated with absolute abundance (based on qPCR), and negatively correlated with the total community DNA (Figure 2B). Its predominance in bare soil was therefore a combined effect of higher *Ca.* Rubrimentiphilales populations and lower populations of other bacteria.

Compared to forested soils, bare soils had higher relative abundances of the phyla *Chloroflexi* and candidate phylum AD3 along with WPS-2, while *Proteobacteria, Acidobacteria*, and *Verrucomicrobia* were lower (Supplementary Figure 1). An OTU co-occurrence analysis (Figure 3) similarly showed a small network of OTUs connected to *Ca.* Rubrimentiphilum, including *Thermogemmatisporaceae* (*Chloroflexi*), candidate phylum AD3, *Conexibacter* (*Actinobacteria*), *Beijerinckiaceae* (*Proteobacteria*) and AS11 (WPS-2).

WPS-2 SAGs and MAGs

Binning of several metagenomes made from DNA extracts of bare Mound soils (Supplementary Table 3) failed to produce high-quality MAGs of WPS-2. We therefore applied a differential centrifugation procedure to preferentially recover small cells from the soil. This procedure depleted communities of *Actinobacteria, Acidobacteria,* and *Chloroflexi*, while WPS-2 and TM7 became more predominant (Supplementary Figure 3). The overall diversity was also lowered by the centrifugation procedure, as shown by the Shannon Index (Supplementary Figure 3). This simplified community dominated by WPS-2 was metagenome sequenced in an attempt to obtain high-quality MAGs. MetaBAT binning yielded 12 bins (Supplementary Figure 4), two of which (bins 6 and 7) contained 16S rRNA gene fragments (<120 bp) identified as WPS-2. Bin 6 was particularly well assembled, with only 48 contigs and an N_{50} of 185 kb. However, MAG Bin 6 had two copies of nearly every single-copy marker gene based on CheckM. The most likely explanation was the presence of two similarly abundant strains of WPS-2 in bin 6. As this bin was comprised of only 48 contigs, we applied a manual curation procedure to divide it into 2 sub-bins (6A and 6B). Details of the procedure and results are

given in Supplementary Table 4. The resulting sub-bins were 2.34-2.58 Mbp and estimated to be 95% complete (Table 1).

Eight WPS-2 SAGs were also generated. These ranged from 0.58 bp to 1.66 Mbp, with estimated completeness from 12% to 54% (Table 1). The SAGs were used primarily to verify the MAG data. Local BLAST and ANI searches against the combined SAG data demonstrated that all 48 contigs in MAG bin 6 could be recruited to these SAGs at high (>75%) nucleotide identity (Supplementary Figure 5), suggesting that the WPS-2 MAGs contained no contaminating DNA sequences from other bacteria.

Pairwise average nucleotide identity (ANI) and average amino acid identity (AAI) comparisons between the 8 SAGs and 2 MAGs were performed using IMG/M (Supplementary Table 5). ANIs were never <78%, and AAIs never <76%, which represent roughly genus level thresholds (Konstantinidis and Tiedje, 2005; Rodriguez-R and Konstantinidis, 2014; Rodriguez-R *et al.*, 2018).We conclude that all SAGs and MAGs belong to a single genus (Rodriguez-R *et al.*, 2018), for which we suggest the name *Candidatus* Rubrimentiphilum. However, 5 separate putative species of this genus were identified based on a ANI threshold of 95% (Richter and Rosselló-Móra, 2009).

Phylogenetic analyses verified the close relationship of the SAGs, MAGs, and sub-OTUs from the amplicon analyses. One SAG (H17) contained a full-length 16S rRNA gene sequence, which matched perfectly to a 16S rRNA gene fragment in MAG bin 6 and to a sub-OTU (rubr5) from the V3-V4 amplicon analyses (Supplementary Figure

2). This sequence was closely related to the most predominant sub-OTU (rubr1). Highly resolved phylogenies were also created for the SAGs and the MAGs based on a set of 56 conserved genomic markers (Figure 4), verifying that all SAGs and MAGs from this study were closely related and belonged to phylum WPS-2.

Phylogenetic and functional comparisons of WPS-2 genomes

A phylogenetic reconstruction using 56 conserved genomic marker genes, including all available WPS-2 genomes along with selected reference genomes, placed the SAGs and MAGs from our study into a monophyletic cluster denoted as *Ca.* Rubrimentiphilum (Figure 4). Several WPS-2 MAGs from a large-scale study of Arctic peatlands (Woodcroft *et al.* 2018) fell into this group as well. Potential hydrogenase-encoding genes and key genes encoding ribulose bisphosphate carboxylase are widespread among WPS-2 genomes, but were not found in any of the *Ca.* Rubrimentiphilum genomes.

The abundance profiles of Pfams in the two *Ca.* Rubrimentiphilum MAGs 6A and 6B were compared to a taxonomically balanced reference database consisting of 2363 genomes of cultured bacteria. This analysis aimed to identify genomes with a similar functional gene complement (i.e. a similar niche) to *Ca.* Rubrimentiphilum. The most similar genomes to WPS-2 based on this analysis belonged to various soil organoheterotrophs, mostly within the phyla *Acidobacteria, Armatimonadetes, Actinobacteria,* and *Chloroflexi* (Figure 5; Supplementary Figure 6). All of the most closely related bacteria were aerobic organoheterotrophs (Zarilla and Perry, 1984; Sako *et al.*, 2003; Urios *et al.*, 2006; Barabote *et al.*, 2009; Johnson *et al.*, 2009; Ward *et al.*, 2009; Wu *et al.*, 2009; Pati *et al.*, 2010; Dunfield *et al.*, 2012; Kielak *et*

al., 2016), and none were autotrophic. Most were also thermoacidophiles, which may indicate an abundance of polyextremophilic tolerance mechanisms in common with *Ca.* Rubrimentiphilum.

Compared to the 2363 reference genomes, the *Ca.* Rubrimentiphilum genomes had a large proportion of genes dedicated to amino acid and protein metabolism (Figure 6 and Supplementary Figure 7). Additionally, genes for “Cell wall, membrane, envelope biogenesis” comprised more than 9% of *Ca.* Rubrimentiphilum genomes, comparable only to *Acidobacteria* and significantly higher than the average in the other selected lineages (Figure 6). Another notable property of *Ca.* Rubrimentiphilum was the low number of genes responsible for inorganic ion complexing and import, which could reflect the metal rich environment.

Metabolic potential of Ca. Rubrimentiphilum

Metabolic predictions for *Ca.* Rubrimentiphilum were made based on the nearly complete MAG bins 6A and 6B, and the different SAGs (Figure 7). The genomic data indicate rod-shape, flagella, P and type IV pili, and an outer membrane with a typical lipid A core. A chromosomal cluster for biosynthesis of peptidoglycan includes the usual enzymes for biosynthesis, translocation and cross-linking of a disaccharide-pentapeptide monomeric unit, but in addition encodes a protein with low similarity to the CofE enzyme, which catalyzes GTP-dependent glutamylation of coenzyme F420 precursor. The presence of this enzyme may indicate unusual peptidoglycan structure with peptide units modified with amino acids or polyamines. Biosynthetic pathways for lysine, arginine, cysteine, branched-chain and aromatic amino acids, folate, riboflavin, thiamin, biotin, pyridoxine, heme, and CoA are

complete. While cobalamin-dependent methionine synthase and ribonucleoside diphosphate reductase are encoded, the genes for de novo cobalamin biosynthesis appear to be lacking, suggesting the possibility of cross-feeding with other populations in the community.

The genomes encode proteins for glycolysis, the oxidative pentose-phosphate shunt, a complete TCA cycle, and an electron transport chain including an NADH dehydrogenase complex, quinol-cytochrome c reductase (cytochrome bc1 complex), aa3-type cytochrome c oxidase, and an F_0-F_1-type ATPase (Figure 7). The respiratory chain could utilize both ubiquinone and menaquinone, the latter synthesized via a futalosine pathway. The respiratory chain is likely to be even more complex due to the presence of periplasmic and membrane-associated multi-copper oxidase enzymes and other proteins of unknown function with cupredoxin domains. The presence of all these genes suggests aerobic metabolism with the capability to adapt to varying oxygen concentrations.

Some functions previously proposed to be important for WPS-2 bacteria, i.e. autotrophy, phototrophy, and trace gas metabolosm, were not evident in *Ca.* Rubrimentiphilum (Figure 4; Supplementary Table 6). Homologues of genes encoding the large subunit of ribulose bisphosphate carboxylase, although common in other WPS-2 bacteria, were not found in any *Ca.* Rubrimentiphilum SAG or MAG, including those detected in Arctic peats (Figure 4), ruling out the Calvin Benson Bassham Cycle. Key steps for other inorganic carbon fixation pathways were also missing. There was also no evidence of phototrophy in *Ca.* Rubrimentiphilum. Indeed, there was no evidence for chlorophyll or bacteriochlorophyll-based

phototrophy in the entire soil metagenome (although there are predicted bacteriorhodopsin-encoding genes), suggesting that phototrophy is not driving the community as a whole. Trace gas metabolism-encoding genes were abundant in the soil metagenome, over 800 genes were annotated as hydrogenase components. However, none were associated with *Ca.* Rubrimentiphilum. Only a weak homologue to a gene for aerobic CO/xanthine dehydrogenase (<40% identity) was detected.

These bacteria appear to prefer nucleotides and amino acids as growth substrates. We predict that it can degrade glycine via glycine dehydrogenase complex, branched-chain amino acids via branched-chain oxoacid dehydrogenase complex, threonine via threonine-3-dehydrogenase and threonine aldolase; histidine via histidine ammonia-lyase; tyrosine via homogentisate pathway; tryptophan via kynurenine, and lysine via lysine aminomutase pathway. A high proportion of protein processing pfams/COGs was noted in comparison to other bacteria (Figure 6). Particularly abundant individual COGs identified via analysis of heatmaps in IMG/M included: peptide/Ni ABC transporter substrate binding proteins; dipeptidyl aminopeptidase/acylaminoacyl peptidase; amino acid transporters; and TonB C-terminal (Supplementary Table 6). TonB_C interacts with outer membrane transporters to facilitate the transport of large molecules like siderophores, vitamin B_{12}, Ni, or large polymers such as oligopeptides (Schauer *et al.,* 2008).

A few enzymes for metabolism of complex carbohydrates are encoded, including a putative beta-hexosaminidase of the glycosyl hydrolase (GH) 20 family, putative GH family 5 and GH43/DUF377 proteins, a secreted protein with low similarity to 1,3-beta-glucanase, and an unusual protein with fused GH family 1 and a periplasmic

substrate-binding protein. Similar fusion proteins are also present in several Cyanobacteria, but their enzymatic activity and physiological function remain unexplored. N-acetyl-D-glucosamine kinase and N-acetylglucosamine-6-phosphate deacetylase for peptidoglycan degradation are encoded. A third gene required for this pathway, encoding glucosamine-6-phosphate deaminase, was not annotated, but predicted copper amine oxidase-encoding genes are located near the N-acetylglucosamine-6-phosphate deacetylase genes in each MAG that may serve this function. Other potential functions for degradation of polymers include a predicted peptidoglycan/xylan/chitin deacetylase, an N-acylglucosamine-6-phosphate 2-epimerase involved in the N-acetylmannosamine (ManNAc) utilisation pathway found in pathogenic bacteria, and β-N-acetylhexosaminidase (Supplementary Table 6). β-*N*-acetylhexosaminidases cleave and transfer diverse substrates including the β-1,4 bond between *N*-acetylglucosamine and anhydro-*N*-acetylmuramic acid (Slámová *et al.*, 2010).

Other energy sources may include formate and thiosulfate, the latter being oxidized to tetrathionate by thiosulfate dehydrogenase. Genes expected in an iron oxidising organism, i.e. those encoding ferric reductase or rusticyanin (Hedrich *et al.*, 2011) were not predicted.

Many gene products were classified within beta lactamase pfam categories (Supplementary Table 6). Indeed, the peptidoglycan degradation mechanisms described above may be involved in the recycling of peptidoglycan units after the action of beta lactams. In addition, four genes in each *Ca.* Rubrimentiphilum MAG (and up to 5 in each SAG) were annotated as virginiamycin B lyase or streptogramin

lyase, enzymes active against actinomycetal antibiotics. Remarkably, this COG (COG4257) was found in only 489 of 6992 finished bacterial genomes on IMG, and never at more than 2 copies per genome. Each of the 4 copies in the MAGs are different, with as little as 38% amino acid identity to each other.

Enrichment and cultivation efforts

The *Ca.* Rubrimentiphilum-specific qPCR assay was used for rapid screening of potential enrichments (Supplementary Table 7). Incubations failed to yield conclusive growth under any of the conditions tested (data not shown).

Discussion

An extensive 16S rRNA-gene based survey of microbial communities in the Paint Pots area revealed an abundance of bacteria belonging to the candidate phylum WPS-2. The most predominant WPS-2 OTUs belonged to a cluster we designated as *Ca.* Rubrimentiphilum. These were particularly abundant in bare soils rich in iron oxides deposited by former springs. The acidity, high metal content, and extremely compact nature of these soils prevent vegetation from establishing (Grasby *et al.*, 2013). Recent studies have reported high relative abundances of WPS-2 in other organic-poor soils, indicating that these may be a preferred habitat (Costello *et al.*, 2009; Stibal *et al.*, 2015; Ji *et al.*, 2016). Here we quantified this observation by comparisons of adjacent bare and vegetated areas. Bacterial communities in vegetated soil not only supported lower relative abundances of *Ca.* Rubrimentiphilum, but also lower absolute abundances based on a qPCR assay, indicating that these bacteria do indeed prefer bare soils.

There are similarities between the bacterial communities in the bare Paint Pots soils and communities recently described in bare Antarctic soils by Ji *et al.* (2017). Both show predominant WPS-2 and AD3 candidate phyla, along with *Thermogemmatisporaceae* (*Chloroflexi)* and *Actinobacteria*. These assemblages are likely adapted to stresses such as metal toxicity, low pH, extreme dryness, high radiation, and very limited organic matter. Ji *et al*. (2017) present convincing evidence that the Antarctic soil community is supported largely by trace gas metabolism and autotrophy, and analysis of one WPS-2 MAG (bin 22) verified the presence of genes encoding these functions (Ji *et al.*, 2017). Additionally, these genes are found in multiple WPS-2 genomes reported by Woodcroft *et al.* (2018) as summarized in Figure 4. Photoautotrophy has also been predicted in WPS-2 MAGs recovered from *Sphagnum* wetlands (Holland-Moritz *et al.*, 2018; Ward *et al.*, 2019). These MAGs contain 16S rRNA gene fragments too short to be unambiguously placed on SSU phylogeny. However, genome-based phylogenies place them in distant clusters of WPS-2 compared to the *Ca*. Rubrimentiphilum we analysed in this study.

Genes for autotrophy and trace gas metabolism (but not phototrophy) were conspicuous in our bare soil metagenome, and these processes likely provide much of the primary productivity. For example, relatives of the *Thermogemmatisporaceae* co-occurred strongly with *Ca.* Rubrimentiphilum in the bare soil. These bacteria have been shown to oxidise atmospheric CO (King and King, 2014), and both publicly available genomes of this group: *Thermogemmatispora* sp. T81 (Stott *et al.*, 2008) and *Thermogemmatispora carboxidovorans* PM5 (King and King, 2014) also contain hydrogenases (but not the CBB cycle). However, although these processes

are undoubtedly critical to the soil community as a whole, they are not present in the predominant WPS-2 cluster designated as *Ca.* Rubrimentiphilum. These genomes lacked any obvious genes encoding methanotrophy, H_2 oxidation, autotrophy, or phototrophy. We note that a second WPS-2 MAG predicted by Ji *et al.* (2017) (bin 23) from Antarctic soil is closely related to our *Ca.* Rubrimentiphilum group (91% identity, Supplementary Figure 2), and also does not show any evidence of autotrophy or trace gas metabolism (Supplementary Table 6).

Instead, genome annotation of *Ca.* Rubrimentiphilum indicated an aerobic heterotroph, perhaps growing on amino acids, polypeptides, nucleotides or some complex polymers. Thiosulfate lithotrophy may provide additional energy. This niche is also indicated by comparisons of the genomic pfam content to reference cultured organisms. *Ca.* Rubrimentiphilum genomes were functionally most similar to versatile organoheterotrophic bacteria, not lithoautotrophs. The most functionally similar bacteria (Figure 5) are aerobic heterotrophs with versatile organic substrates: for example *Acidothermus cellulolyticus* grows on diverse cell wall polymers (Barabote *et al.*, 2009), while all known *Armatimonadetes (Dunfield et al.*, 2012) and most *Acidobacteria (Ward et al.*, 2009; Kielak *et al.*, 2016) grow on several complex polymeric substrates. It seems counterintuitive for such an organic-poor environment, but genomic comparisons with other bacteria, and specific inferences based on genome annotations, both indicate a primarily organoheterotrophic lifestyle for *Ca.* Rubrimentiphilum. The soil does experience some input from deadfall, and a close examination did show the incorporation of conifer needles. However, we did not detect elevated amounts of *Ca.*

Rubrimentiphilum associated with this organic material (data not shown). They may instead grow in association with other bacteria.

A key to obtaining high quality MAGs in this study was differential extraction of small cells from soil before metagenome sequencing. A similar filtration approach has proven effective in characterizing members of the Patescibacteria (Rinke *et al.*, 2013; Parks *et al.*, 2018) or Candidate Phylum Radiation (Wrighton *et al.*, 2012; Hug *et al.*, 2016). SAGs were also generated to verify the MAGs. All contigs in the MAGs could be recruited to SAG DNA with high identity, indicating that MAGs were not contaminated with DNA from other organisms. Metagenome binning always runs the risk of contamination, but the combination of single cell genomics with metagenomic binning of simplified communities in our study makes contamination very unlikely, and adds confidence to our metabolic interpretations.

The estimated genome size of *Ca.* Rubrimentiphilum (max 2.6 Mb) is small but not atypical of a free-living extremophile (Podar *et al.*, 2008). The bacterium co-occured with a simple network of OTUs belonging to the *Chloroflexi,* AD3, *Acidobacteria*, *Proteobacteria, Actinobacteria* and another group of WPS-2 (AS11). This network may reflect direct biotic relationships such as parasitism, or it may simply reflect common environmental adaptations. A remarkable feature of the *Ca.* Rubrimentiphilum genomes was the number of genes encoding resistance against antimicrobial agents, including more genes encoding for streptogramin lyase than found in any other genome. These bacteria may live in close association with antimicrobial producing bacteria, and survive in part by scavenging cell components of these other microbes. The most conspicuous partners of the *Ca.*

Rubrimentiphilum in our site are not well studied in terms of antibiotic production capabilities. However, there are indications that some *Chloroflexi* may have this capacity (Nett and König, 2007). *Thermosporothrix hazakensis* SK20-1T, a relative of one *Cloroflexi* OTU closely co-occurring with *Ca.* Rubrimentiphilum, has recently been shown to produce thiazoles as secondary metabolites (Park *et al.*, 2015).

Attempts to develop an enrichment of *Ca.* Rubrimentiphilum failed. Previous SIP based studies of the soil using cellulose and other EPS also did not demonstrate any enrichment of WPS-2 (Wang *et al.*, 2015). *Ca.* Rubrimentiphilum may not utilize any of the provided substrates, or precise conditions for growth were not met, although by performing the enrichment directly in the soil we attempted to bypass this issue. It is also possible that these are slowly growing bacteria and growth could simply not be seen at the time scales of these experiments. Finally, it is possible the substrate concentrations added (around 0.025-0.3% w/w, Supplementary Table 7) were too high. Oligotrophic bacteria can be inhibited by excessive substrate availability, particularly in the case of amino acids (Kuznetsov *et al.*, 1979). For example the acidobacteria *Edaphobacter* spp. can grow on casamino acids and peptone only when provided at very low concentrations around 0.01% w/v (Koch *et al.*, 2008).

In summary, we noted that there is a predominant genus of the candidate phylum WPS-2 (*Ca.* Rubrimentiphilum) in bare acidic soils of the Paint Pots site, and applied a combination of ecological and genomic studies to understand its niche. Based on our data, the bacteria prefer the bare, extreme soil to more organic forest soil. Despite the obvious hypothesis that they should survive via lithoautotrophy,

photoautotrophy, or trace gas metabolism in bare soils, they were instead predicted to be aerobic heterotrophs, possibly scavenging cell components from hardy autotrophic bacteria as energy substrates. As other bacteria from the WPS-2 phylum have been predicted to be primary producers (photoautotrophs or autotrophic hydrogenotrophs), our study suggests that there is considerable metabolic versatility across the WPS-2 phylum, with alternative autotrophic/heterotrophic and phototrophic/lithotrophic/organotrophic phenotypes.

We propose the following tentative species description for the bacterium *Ca.* Rubrimentiphilum (from L. neut. n. rubrimentum, red ink; N.L. neut. adj. philum [from Gr. neut. adj. philon], friend, loving). Gram negative, aerobic heterotrophic bacterium. Motile via flagella. Forms type P and type IV pili. Amino acids and nucleotides are the presumed preferred substrates. Other substrates include thiosulfate, formate, and possibly some oligopeptides or oligosaccharides. Prototrophic for most amino acids and vitamins, auxotrophic for vitamin B12.

Experimental Procedures

Sampling

Samples from the Paint Pots, Kootenay National Park, B.C (N 51.16991˚, W 116.14735˚) were collected into sterile 50-mL Falcon tubes (VWR, Mississauga, Canada) on Aug. 13, 2012, Oct. 20, 2014, July 19, 2016 and May 25, 2017. The geology and geochemistry of the area have been described elsewhere (Everdingen, 1970; Grasby *et al.*, 2013). Forested soils are Podzols (USDA: Spodosols) with a thick O horizon, a distinct E horizon, and a deep organic A horizon. However, areas covered with iron oxide depositions from active or extinct springs are generally unvegetated mineral soils with no O horizon, although various grasses do grow on

the fringes of the forest or in flooded marshy areas. Over 70 samples were taken from various locations as shown in Figure 1 and Supplementary Table 1. The Source is the present site of spring water discharge. The Relic Spring is a water-filled pool still connected to the water source but no longer discharging water. The Mound and the Ancient Mound are former spring sources that are now dry. Acidity, metal toxicity, and compactness due to the solidified iron oxides are the most likely reasons for the lack of vegetation on the Mound and Ancient Mound areas (Grasby *et al.*, 2013). Samples were transported back to the laboratory on ice, and 0.5-g subsamples immediately frozen at -80 ˚C until DNA extractions were performed. Soil samples for enrichments were stored at 4 ˚C in the dark until use.

Microbial community analyses

Community analyses were performed as described previously (Ruhl *et al.*, 2018). Briefly, DNA was extracted from 500 mg of soil with the FastDNA® SPIN Kit for Soil (MP Biomedicals, Santa Ana, CA, USA), with an additional 5.5 M guanidine isothiocyanate wash to remove humic acids (Knief *et al.*, 2003). 16S rRNA gene amplicon libraries were prepared as per the Illumina (San Diego, CA, USA) standard library preparation protocol: "16S Metagenomic Sequencing Library Preparation" (Part # 15044223 Rev. B), except that Taq polymerase was used in the second PCR reaction. Primers targeted the V3-V4 region (341fw and 785r). Amplicon libraries were quantified using the Qubit HS kit (Invitrogen, Carlsbad, CA, USA), diluted to 4 nM, pooled, and prepared for sequencing on a MiSeq instrument as per Illumina's standard protocol "Preparing Libraries for Sequencing on the MiSeq" protocol (Part # 15039740 Rev. D). Libraries were sequenced using the MiSeq Reagent Kit v3, 600 cycles (Illumina part number MS-102-3003). The Qiime2 software 2019.4 was used

to analyse 16S rRNA sequence data (Bolyen *et al.*, 2019). Raw reads were quality controlled and denoised sub-OTUs were formed using the deblur plugin (Amir *et al.*, 2017). Taxonomic assignment was performed with the feature-classifier plugin (Bokulich *et al.*, 2018) employing a naïve Bayes classifier approach. The taxonomy classifier for the analysis was trained on thte Silva database, release 132 (Quast *et al.*, 2013) for the V3-V4 region, after manual edits to to delineate *Ca.* Rubrimentiphilales and AS-11 lineages of phylum WPS-2 (Supplementary Figure 2).

Co-occurrence network analysis

The sub-OTUs detected from the total of 67 samples collected in 2014, 2016 and 2017 (Supplementary Table 1) were used to construct a co-occurrence network. An inclusion threshold of average abundance of 0.25% was applied to select the major lineages (See Supplementary Table 2 for full taxon strings of the sub-OTUs). The network topology was constructed using igraph from the Jaccard similarity adjacency matrix using the picante package in R (Csardi and Nepusz, 2006; Hardy, 2008). The network was visually rendered with Gephi software using the ForseAtlas2 algorithm for its layout (Jacomy *et al.*, 2014). Network modularity was computed using the Louvain algorithm for community detection (Blondel *et al.*, 2008).

Quantitative PCR-based estimation of Ca. Rubrimentiphilum abundance

The probe design function of the ARB software (Ludwig *et al.*, 2004) and ARB-Silva 111 database were used to identify paired oligonucleotide primer sequences targeting *Ca.* Rubrimentiphilales and excluding the AS-11 group. A primer pair specifically targeting a 242 bp fragment was designed: WPS-2_for (5’-

GCACTCACTCGAGACCGCCGTT – 3’) and WPS-2_rev (5’ – GGGAACGTATTCACCGCAGCGT -3’). These were quality controlled against dimer and hairpin formation via OligoCalc (Kibbe, 2007) and searches against the ARB-Silva 111 database showed a minimum of 4 mismatches to non-target sequences (including AS11). Primers were obtained from Invitrogen (Waltham, MA, USA). Illumina sequencing of these qPCR amplicons from 4 Paint Pots samples confirmed the primer specificity (data not shown). To create standards, PCR amplification was performed on soil DNA and the resulting amplicon cloned into a pJET 3.0 plasmid (ThermoFisher, Waltham, MA, USA). Sanger sequencing confirmed that the plasmid-borne sequence matched the target sequence. Standard dilutions were constructed with a purified amplicon from the plasmid. The maximum primer annealing temperature was determined using gradient PCR (Veriti 96 Well Thermal Cycler, Applied Biosystems, Waltham, MA, USA) to maximise the stringency of the qPCR assay. qPCR was performed on a Rotor-Gene 6000 (QIAGEN, Velno, Netherlands) using a SYBR Green qPCR master mix (QIAGEN, Velno, Netherlands) in 12.5-μl total reaction volumes with 1 μM of each primer included. qPCR runs were performed under the following conditions: Initial denaturation at 95 °C for 10 minutes; 40 cycles at 95 °C for 20s, 72°C for 20s and 72 °C for 20s; a pre-melt conditioning step at 72 °C for 90s; and a melt ramp from 72 °C to 95 °C increasing 0.5 °C every 5s. Amplicon melt profiles (0.5°C increments) did not show evidence of non-specific amplification.

Single cell genomics

Soil samples taken on Aug. 13, 2012 from the Mound were cryopreserved with 5% glycerol solution in TE buffer and frozen on dry ice. Single cells were isolated with

fluorescence-activated cell sorting (FACS), lysed and whole genome amplified using multiple displacement amplification (MDA), and MDA products were screened with 16S rRNA gene PCR according to standard JGI protocols (Rinke *et al.*, 2014). From 397 wells that amplified with MDA, 58 produced 16S rRNA gene PCR amplicons, of which 8 were positively identified as WPS-2 and genome sequenced (Table 1). Preparation of the libraries, sequencing and assembly procedures are summarized in Supplementary Table 4.

Metagenomic sequencing and MAG binning

Metagenomic binning of several Mound soil metagenomes (listed in Supplementary Table 3) did not yield high-quality and low-contamination genomes of *Ca.* Rubrimentiphilum. In order to obtain better MAGs, we performed cell size fractionation prior to DNA extraction. Briefly, 20 g of soil was mixed with 25 ml of cell detachment buffer (Eichorst *et al.*, 2007) supplemented with 10 mM $Na_4P_2O_7 \cdot 10H_2O$ and 1 mM dithiothreitol, and stirred for 1 hour. The soil suspension was transferred into a 50-ml conical centrifuge tube left to settle for 5 min, then decanted and used in 2-3 sequential centrifugation steps. A short spin at 1,500 × g for 2 min pelleted large debris and soil particles (Designated Fraction 0). The clear supernatant was gently transferred to new centrifuge tubes for a further 1-2 steps of centrifugation. In one (3-step) trial, centrifugation was applied at 3,000 × g for 5 min and the resulting pellet (Fraction 1) resuspended in PBS buffer; transferred supernatants were centrifuged at 19,000 × g for 30 min, and pellets (Fraction 2) resuspended in PBS buffer. Alternatively, for a 2-step procedure the middle centrifugation step was skipped, and Fraction 1 was collected after centrifugation at 19,000 × g for 30 min. Microbial communities in each fraction were analyzed as

described above. A fraction with a low-diversity community enriched in WPS-2 (NM2-5_DC) was selected for metagenome sequencing and assembly as described in Supplementary Table 4. MAG binning and quality control were performed with MetaBAT v2.12.1 (Kang *et al.*, 2015) and CheckM v1.0.9 (Parks *et al.*, 2015), respectively.

Functional comparisons of WPS-2 genomes with other genomes

MAGs and SAGs were annotated via the JGI standard pipeline (Huntemann *et al.*, 2015, 2016) and further analysed using the IMG/M platform (Chen *et al.*, 2019). The Pfam and COG profiles of WPS-2 MAGs were compared to a reference database that included 2363 bacterial genomes from IMG (available in Supplementary File 2). To minimize taxonomic bias in the database, this reference set included one bacterium (whenever possible with a finished status) from each described genus. The Pfam and COG content profiles of the MAGs were compared against the reference bacteria via the calculation of Jaccard and Bray-Curtis similarity indices (Legendre and Legendre, 2012). Two-tailed t-tests were used to calculate significance.

Phylogenetic analyses

A maximum likelihood concatenated marker gene tree was created by taking all bacterial isolate genomes in the Integrated Microbial Genomes (IMG) database, reducing the 61,619 bacterial genomes to a manageable set by clustering the RNA Polymerase beta subunit gene at 65% and using this as the reference database for tree inference. Ten WPS-2 genomes from the current study (8 SAGs and 2 MAGs) and 2 MAGs from Ji *et al.* (2017), 3 MAGs from (Holland-Moritz *et al.*, 2018) and 53 MAGs from (Woodcroft *et al.*, 2018) were added to the total set of genomes used for

the bacterial tree in Supplementary Figure 8. Briefly, trees were constructed in the following manner. Proteins were called using Prodigal v.2.6.3. Phylogenetic markers were extracted from the resulting faa files using HMMs of each of the 56 markers with HMMER v.3.1b2, then aligned using MAFFT v.7.221 and concatenated using an internal python script. Phylogenetic trees were inferred using IQ tree tree to produce maximimum likelihood trees with 1000 bootstraps (Nguyen *et al.*, 2015). Visualization was produced in R using ape and ggtree (Yu *et al.*, 2017) packages.

For the 16S rRNA gene analysis, full length sequences of WPS-2 from the Nr99 Silva database, release 132 (Quast *et al.*, 2013), were iteratively aligned to WPS-2 sub-OTUs with Muscle v.3.8.425. WPS-2 sub-OTUs of average abundance >0.25% in the entire sequencing dataset were included. Bayesian 16S rRNA gene phylogeny was constructed with MrBayes v.3.2.6 (Huelsenbeck and Ronquist, 2001). Posterior probabilities of the tree were estimated using a Markov Chain Monte Carlo of 1×10^6 cycles with first 1×10^5 states discarded and used for re-initialization (burn-in). Data was analyzed with a 4by4 nucleotide substitution model with a GTR structure. Rate variation was set to gamma-distributed with a proportion of invariable sites.

Enrichment and cultivation

Enrichments efforts for WPS-2 used 4-5 g (wet weight) amounts of soil in 15-mL Falcon tubes, with addition of various monosaccharides, polysaccharides, protein digests, and organic acids (Supplementary Table 7). Duplicates of each substrate were incubated at room temperature in the dark. 0.5-g soil samples were taken biweekly for a period of up to 2 months for DNA extraction and quantification using the WPS-2 qPCR assay.

Acknowledgements

This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (2019-06265). We are indebted to Parks Canada and the Ktunaxa Nation for allowing us to sample in Kootenay National Park. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science under Contract No. DE-AC02-05CH11231.

Conflict of Interest

The authors declare no conflict of interest.

References

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Xu, Z.Z., et al. (2017) Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**: e00191-16.

Barabote, R.D., Xie, G., Leu, D.H., Normand, P., Necsulea, A., Daubin, V., et al. (2009) Complete genome of the cellulolytic thermophile Acidothermus cellulolyticus 11B provides insights into its ecophysiological and evolutionary adaptations. *Genome Res* **19**: 1033–1043.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008) Fast unfolding of communities in large networks. *J Stat Mech* **2008**: P10008.

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., et al. (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**: 90.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852–857.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725–731.

Bragina, A., Berg, C., and Berg, G. (2015) The core microbiome bonds the Alpine bog vegetation to a transkingdom metacommunity. *Molecular Ecology* **24**: 4795–4807.

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

Chen, I.-M.A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., et al. (2019) IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **47**: D666–D677.

Costello, E.K., Halloy, S.R.P., Reed, S.C., Sowell, P., and Schmidt, S.K. (2009) Fumarole-Supported Islands of Biodiversity within a Hyperarid, High-Elevation Landscape on Socompa Volcano, Puna de Atacama, Andes. *Appl Environ Microbiol* **75**: 735–747.

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**: 1–9.

Dewhirst, F.E., Klein, E.A., Thompson, E.C., Blanton, J.M., Chen, T., Milella, L., et al. (2012) The Canine Oral Microbiome. *PLOS ONE* **7**: e36067.

Dunfield, P.F., Tamas, I., Lee, K.C., Morgan, X.C., McDonald, I.R., and Stott, M.B. (2012) Electing a candidate: a speculative history of the bacterial phylum OP10. *Environ Microbiol* **14**: 3069–3080.

Eichorst, S.A., Breznak, J.A., and Schmidt, T.M. (2007) Isolation and Characterization of Soil Bacteria That Define Terriglobus gen. nov., in the Phylum Acidobacteria. *Appl Environ Microbiol* **73**: 2708–2717.

Eloe-Fadrosh, E.A., Paez-Espino, D., Jarett, J., Dunfield, P.F., Hedlund, B.P., Dekas, A.E., et al. (2016) Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat Commun* **7**: 10476.

Everdingen, R.O. van (1970) The Paint Pots, Kootenay National Park, British Columbia—acid spring water with extreme heavy-metal content. *Can J Earth Sci* **7**: 831–852.

Grasby, S.E., Richards, B.C., Sharp, C.E., Brady, A.L., Jones, G.M., and Dunfield, P.F. (2013) The Paint Pots, Kootenay National Park, Canada — a natural acid spring analogue for Mars. *Can J Earth Sci* **50**: 94–108.

Hardy, O.J. (2008) Testing the spatial phylogenetic structure of local communities: statistical performances of different null models and test statistics on a locally neutral community. *Journal of Ecology* **96**: 914–926.

Hedrich, S., Schlömann, M., and Johnson, D.B. (2011) The iron-oxidizing proteobacteria. *Microbiology (Reading, Engl)* **157**: 1551–1564.

Hermans, S.M., Buckley, H.L., Case, B.S., Curran-Cournane, F., Taylor, M., and Lear, G. (2017) Bacteria as Emerging Indicators of Soil Condition. *Appl Environ Microbiol* **83**: e02826-16.

Holland-Moritz, H., Stuart, J., Lewis, L.R., Miller, S., Mack, M.C., McDaniel, S.F., and Fierer, N. (2018) Novel bacterial lineages associated with boreal moss species. *Environ Microbiol* **20**: 2625–2638.

Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016) A new view of the tree of life. *Nature Microbiology* 16048.

Huntemann, M., Ivanova, N.N., Mavromatis, K., Tripp, H.J., Paez-Espino, D., Palaniappan, K., et al. (2015) The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v. 4). *Standards in genomic sciences* **10**: 1.

Huntemann, M., Ivanova, N.N., Mavromatis, K., Tripp, H.J., Paez-Espino, D., Tennessen, K., et al. (2016) The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Standards in Genomic Sciences* **11**: 17.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014) ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* **9**: e98679.

Ji, M., van Dorst, J., Bissett, A., Brown, M.V., Palmer, A.S., Snape, I., et al. (2016) Microbial diversity at Mitchell Peninsula, Eastern Antarctica: a potential biodiversity “hotspot.” *Polar Biol* **39**: 237–249.

Ji, M., Greening, C., Vanwonterghem, I., Carere, C.R., Bay, S.K., Steen, J.A., et al. (2017) Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* **552**: 400–403.

Johnson, D.B., Bacelar-Nicolau, P., Okibe, N., Thomas, A., and Hallberg, K.B. (2009) Ferrimicrobium acidiphilum gen. nov., sp. nov. and Ferrithrix thermotolerans gen. nov., sp. nov.: heterotrophic, iron-oxidizing, extremely acidophilic actinobacteria. *Int J Syst Evol Microbiol* **59**: 1082–1089.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**: e1165.

Kielak, A.M., Barreto, C.C., Kowalchuk, G.A., van Veen, J.A., and Kuramae, E.E. (2016) The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Front Microbiol* **7**: 744.

King, C.E. and King, G.M. (2014) Description of Thermogemmatispora carboxidivorans sp. nov., a carbon-monoxide-oxidizing member of the class Ktedonobacteria isolated from a geothermally heated biofilm, and analysis of carbon monoxide oxidation by members of the class Ktedonobacteria. *Int J Syst Evol Microbiol* **64**: 1244–1251.

Knief, C., Lipski, A., and Dunfield, P.F. (2003) Diversity and activity of methanotrophic bacteria in different upland soils. *Appl Environ Microbiol* **69**: 6703–6714.

Koch, I.H., Gich, F., Dunfield, P.F., and Overmann, J. (2008) Edaphobacter modestus gen. nov., sp. nov., and Edaphobacter aggregans sp. nov., acidobacteria isolated from alpine and forest soils. *Int J Syst Evol Microbiol* **58**: 1114–1122.

Konstantinidis, K.T. and Tiedje, J.M. (2005) Towards a Genome-Based Taxonomy for Prokaryotes. *J Bacteriol* **187**: 6258–6264.

Kuznetsov, S.I., Dubinina, G.A., and Lapteva, N.A. (1979) Biology of oligotrophic bacteria. *Annu Rev Microbiol* **33**: 377–387.

Legendre, P. and Legendre, L. (2012) Numerical Ecology, 3rd ed. Elsevier.

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

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, null, et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Nett, M. and König, G.M. (2007) The chemistry of gliding bacteria. *Nat Prod Rep* **24**: 1245–1261.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**: 268–74.

Nogales, B., Moore, E.R., Llobet-Brossa, E., Rossello-Mora, R., Amann, R., and Timmis, K.N. (2001) Combined use of 16S ribosomal DNA and 16S rRNA

to study the bacterial community of polychlorinated biphenyl-polluted soil. *Appl Environ Microbiol* **67**: 1874–1884.

Park, J.-S., Yabe, S., Shin-ya, K., Nishiyama, M., and Kuzuyama, T. (2015) New 2-(1'H-indole-3'-carbonyl)-thiazoles derived from the thermophilic bacterium Thermosporothrix hazakensis SK20-1(T). *J Antibiot* **68**: 60–62.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**: 996–1004.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**: 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**: 1533.

Pascual, J., Blanco, S., García-López, M., García-Salamanca, A., Bursakov, S.A., Genilloud, O., et al. (2016) Assessing Bacterial Diversity in the Rhizosphere of Thymus zygis Growing in the Sierra Nevada National Park (Spain) through Culture-Dependent and Independent Approaches. *PLoS ONE* **11**: e0146558.

Pati, A., Labutti, K., Pukall, R., Nolan, M., Glavina Del Rio, T., Tice, H., et al. (2010) Complete genome sequence of Sphaerobacter thermophilus type strain (S 6022). *Stand Genomic Sci* **2**: 49–56.

Podar, M., Anderson, I., Makarova, K.S., Elkins, J.G., Ivanova, N., Wall, M.A., et al. (2008) A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. *Genome Biology* **9**: R158.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Richter, M. and Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *PNAS* **106**: 19126–19131.

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., et al. (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* **9**: 1038–1048.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Rodriguez-R, L.M., Gunturu, S., Harvey, W.T., Rosselló-Mora, R., Tiedje, J.M., Cole, J.R., and Konstantinidis, K.T. (2018) The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res* **46**: W282–W288.

Rodriguez-R, L.M. and Konstantinidis, K.T. (2014) Bypassing Cultivation To Identify Bacterial Species: Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. *Microbe Magazine* **9**: 111–118.

Ruhl, I.A., Grasby, S.E., Haupt, E.S., and Dunfield, P.F. (2018) Analysis of microbial communities in natural halite springs reveals a domain-dependent relationship of species diversity to osmotic stress: Salinity versus microbial diversity. *Environmental Microbiology Reports*.

Sako, Y., Nakagawa, S., Takai, K., and Horikoshi, K. (2003) Marinithermus hydrothermalis gen. nov., sp. nov., a strictly aerobic, thermophilic bacterium from a deep-sea hydrothermal vent chimney. *Int J Syst Evol Microbiol* **53**: 59–65.

Slámová, K., Bojarová, P., Petrásková, L., and Kren, V. (2010) β-N-acetylhexosaminidase: what's in a name…? *Biotechnol Adv* **28**: 682–693.

Stibal, M., Schostag, M., Cameron, K.A., Hansen, L.H., Chandler, D.M., Wadham, J.L., and Jacobsen, C.S. (2015) Different bulk and active bacterial communities in cryoconite from the margin and interior of the Greenland ice sheet. *Environ Microbiol Rep* **7**: 293–300.

Stott, M.B., Crowe, M.A., Mountain, B.W., Smirnova, A.V., Hou, S., Alam, M., and Dunfield, P.F. (2008) Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand. *Environ Microbiol* **10**: 2030–2041.

Taş, N., Prestat, E., McFarland, J.W., Wickland, K.P., Knight, R., Berhe, A.A., et al. (2014) Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest. *ISME J* **8**: 1904–1919.

Trexler, R., Solomon, C., Brislawn, C.J., Wright, J.R., Rosenberger, A., McClure, E.E., et al. (2014) Assessing impacts of unconventional natural gas extraction on microbial communities in headwater stream ecosystems in Northwestern Pennsylvania. *Front Microbiol* **5**: 522.

Urios, L., Agogué, H., Lesongeur, F., Stackebrandt, E., and Lebaron, P. (2006) Balneola vulgaris gen. nov., sp. nov., a member of the phylum Bacteroidetes from the north-western Mediterranean Sea. *Int J Syst Evol Microbiol* **56**: 1883–1887.

Wang, X., Sharp, C.E., Jones, G.M., Grasby, S.E., Brady, A.L., and Dunfield, P.F. (2015) Stable-Isotope Probing Identifies Uncultured Planctomycetes as Primary Degraders of a Complex Heteropolysaccharide in Soil. *Appl Environ Microbiol* **81**: 4607–4615.

Ward, L.M., Cardona, T., and Holland-Moritz, H. (2019) Evolutionary Implications of Anoxygenic Phototrophy in the Bacterial Phylum Candidatus Eremiobacterota (WPS-2). *Front Microbiol* **10**:.

Ward, N.L., Challacombe, J.F., Janssen, P.H., Henrissat, B., Coutinho, P.M., Wu, M., et al. (2009) Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. *Appl Environ Microbiol* **75**: 2046–2056.

879 Woodcroft, B.J., Singleton, C.M., Boyd, J.A., Evans, P.N., Emerson, J.B., Zayed,
880 A.A.F., et al. (2018) Genome-centric view of carbon processing in
881 thawing permafrost. *Nature* **560**: 49–54.
882 Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J.,
883 VerBerkmoes, N.C., et al. (2012) Fermentation, hydrogen, and sulfur
884 metabolism in multiple uncultivated bacterial phyla. *Science* **337**:
885 1661–1665.
886 Wu, D., Raymond, J., Wu, M., Chatterji, S., Ren, Q., Graham, J.E., et al. (2009)
887 Complete genome sequence of the aerobic CO-oxidizing thermophile
888 Thermomicrobium roseum. *PLoS ONE* **4**: e4207.
889 Wu, X., Holmfeldt, K., Hubalek, V., Lundin, D., Åström, M., Bertilsson, S., and
890 Dopson, M. (2016) Microbial metagenomes from three aquifers in the
891 Fennoscandian shield terrestrial deep biosphere reveal metabolic
892 partitioning among populations. *ISME J* **10**: 1192–1203.
893 Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H.,
894 et al. (2014) Uniting the classification of cultured and uncultured
895 bacteria and archaea using 16S rRNA gene sequences. *Nat Rev*
896 *Microbiol* **12**: 635–645.
897 Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017) ggtree: an r
898 package for visualization and annotation of phylogenetic trees with
899 their covariates and other associated data. *Methods in Ecology and*
900 *Evolution* **8**: 28–36.
901 Zarilla, K.A. and Perry, J.J. (1984) Thermoleophilum album gen. nov. and sp.
902 nov., a bacterium obligate for thermophily and n-alkane substrates.
903 *Arch Microbiol* **137**: 286–290.
904

TABLES

Table 1. Summary of WPS-2 SAGs and MAGs obtained from the Paint Pots Mound site. Genome completeness and contamination were estimated with CheckM. Genome qualities were assigned according to the standards outlined in (Bowers *et al.*, 2017).

Name	IMG ID	Assembly Size, Mb	Number of Scaffolds	N_{50}, kb	GC Content, %	Completeness estimate, %	Contamination estimate, %	Genome quality	Gene Count	Genes with Pfam	Genes with COG
MAGs from differential soil extraction metagenome [1]											
6	2781125698	4.92	48	185	60.9	100	100		5132	3671	2939
6A	2781125701	2.58	19	253	60.7	95	2.78	High-quality draft	2688	1900	1523
6B	2781125702	2.34	29	160	61.0	95	2.78	High-quality draft	2444	1771	1415
SAGs											
H17	2617270725	1.66	134	28	60.6	54	0.0	Medium-quality draft	1798	1015	728
I5	2706794755	1.47	118	23	60.5	43	0.0	Low-quality draft	1621	970	653
G15	2616644938	1.03	105	28	60.0	44	0.0	Low-quality draft	1131	680	427
C21	2706794749	1.20	94	18	60.9	29	0.0	Low-quality draft	1311	843	549
F4	2616644933	1.13	168	46	60.6	44	0.0	Low-quality draft	1306	789	501
C6	2616644939	0.59	75	19	60.5	22	0.0	Low-quality draft	678	461	271
B5	2616644827	0.58	70	16	60.1	12	0.0	Low-quality draft	648	414	269
I16	2616644826	0.74	113	30	60.5	37	0.0	Low-quality draft	863	551	354

[1] CheckM analysis suggests that bin 6 contained two closely related strains. Bin 6 was separated into bins 6A and 6B, each representing a single strain.

FIGURES

Figure 1: A: Overview map of the Paint Pots area, with orange areas representing ochre coloured soils, and white areas representing forested podzols. **B** is a blow-up of the area indicated by the dashed square in A and, **C** is blow-up of the bare Mound indicated by the dashed square in B. Percentages are 16S rRNA gene relative abundances of WPS-2 bacteria in different sampling points. Soil samples were taken from the surface 0-5 cm layer unless otherwise specified. The circle diameters in C are proportional to the WPS-2 relative abundances. Additional information on the samples is provided in Supplementary Table 1.

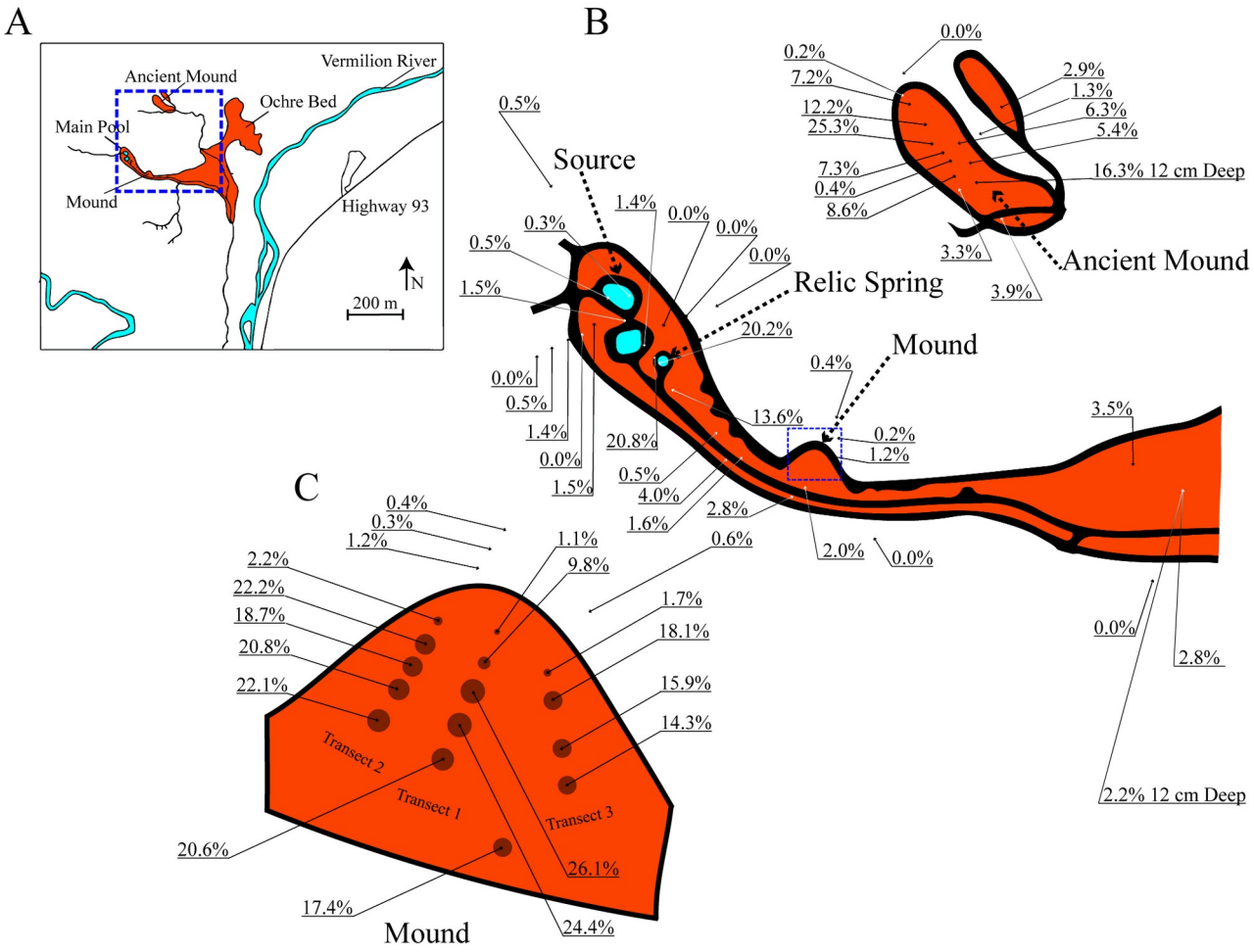

Figure 2: DNA recovery and abundance of *Ca.* Rubrimentiphilales in bare versus vegetated soils. **A:** Absolute abundance determined via a specific 16S rRNA gene targeted qPCR assay (blue) and total DNA yield (red). Unpaired t-tests were applied to determine statistical significance between bare and vegetated soils. **B:** Relationship of absolute WPS-2 abundance determined via qPCR (blue) and total DNA obtained from the samples (red), to the relative WPS-2 abundance determined via 16S rRNA gene amplicon sequencing. The slopes of both regression lines differed significantly from zero. Data are plotted as a log-log scale to ensure they are parametric.

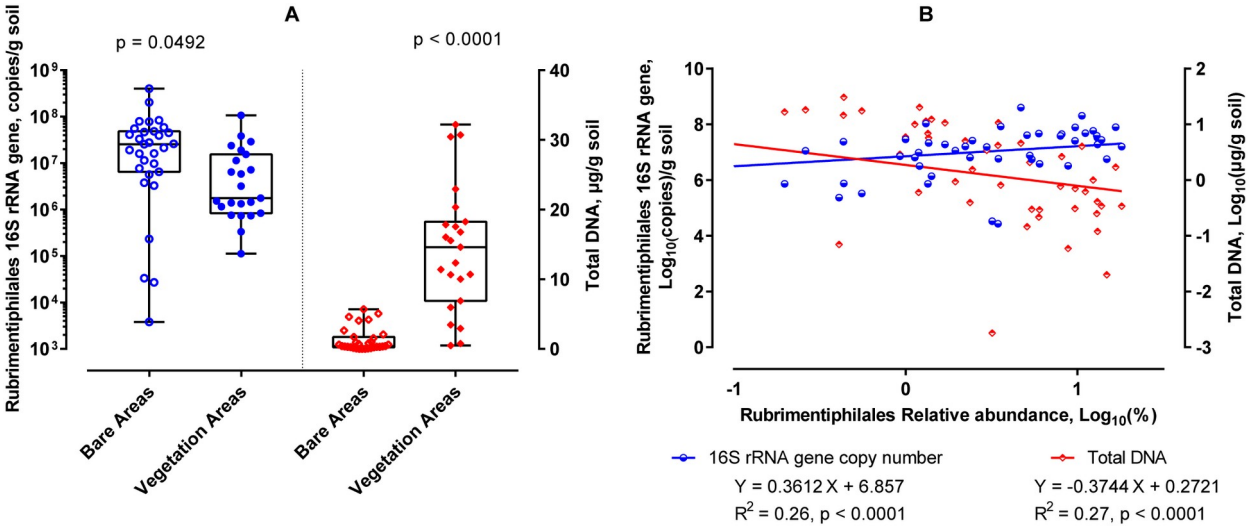

Figure 3: Co-occurrence network analysis of 16S rRNA gene sub-OTUs, based on 67 soil Paint Pots samples (Supplementary Table 1). Sub-OTUs of average relative abundance >0.25% across all samples were used to construct the network, unconnected nodes are not shown. The edges represent positive association which is proportional to line thickness. Node sizes are proportional to the average abundances of sub-OTUs. All analyzed *Chloroflexi* sub-OTUs were placed inside the same sub-network, which also contains the most abundant WPS-2 sub-OTU rubr1 (i.e. *Ca.* Rubrimentiphilum, indicated by the arrow). The strongest associations of rubr1 were formed with chl1 (*Ktedonobacteria*), chl3 (*Thermosporothrix* sp.), ad3-1 (candidate phylum AD3) and the most abundant WPS-2 sub-OTU from the AS-11 clade. Weaker associations included beij1 (*Beijerinckiaceae*), chl2 (*Ktedonobacteria)* and aci7 (*Bryobacter*). See Supplementary Table 2 for full taxon strings of the OTUs in the graph.

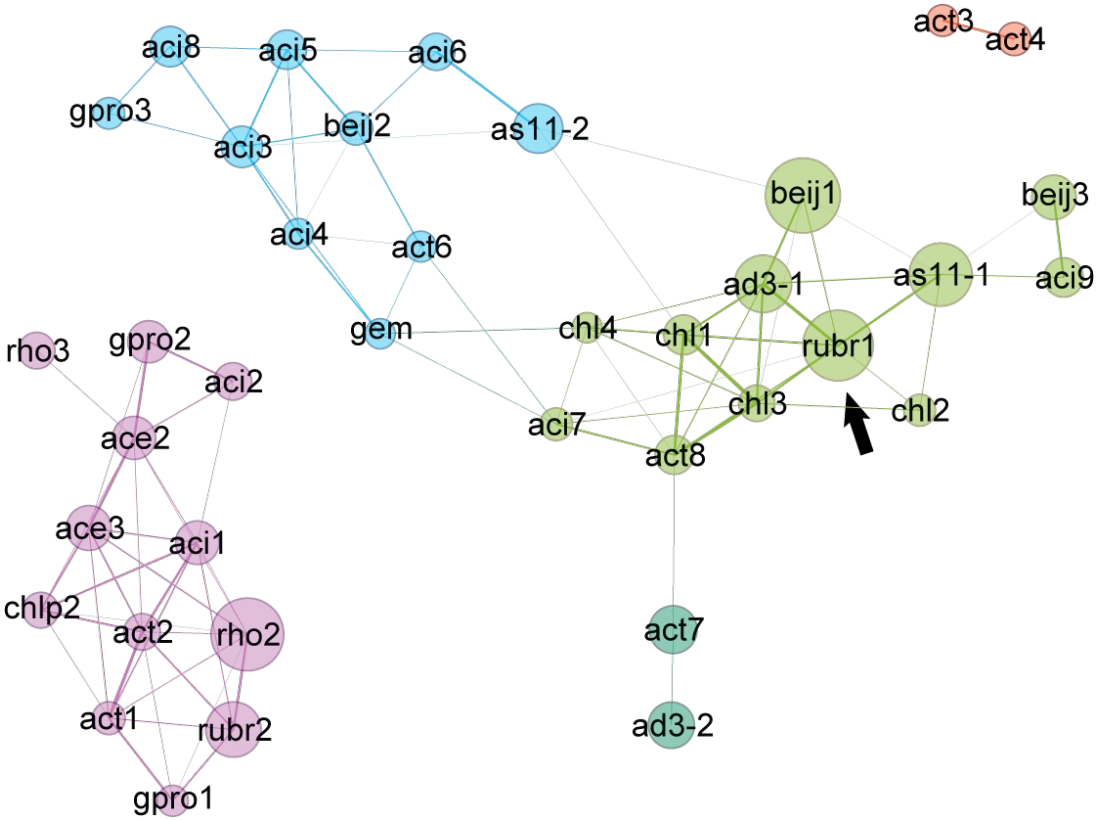

Figure 4: Phylogenetic tree based on a set of 56 conserved marker genes showing the relationships of 69 available WPS-2 genomes in relationship to reference bacteria. Phylogenetic trees were inferred using IQ tree to produce maximimum likelihood trees with 1000 bootstraps. Bootstrap support is shown as a solid blue circles for nodes with greater than 50% support with sizes proportional to the node supports. The presence of Rubisco or hydrogenases within a genome is indicated with bubbles of the sizes proportional to the copy number of a corresponding gene. Neither gene was found within any *Ca*. Rubrimentiphilum, but both are usually present in other WPS-2 genomes.

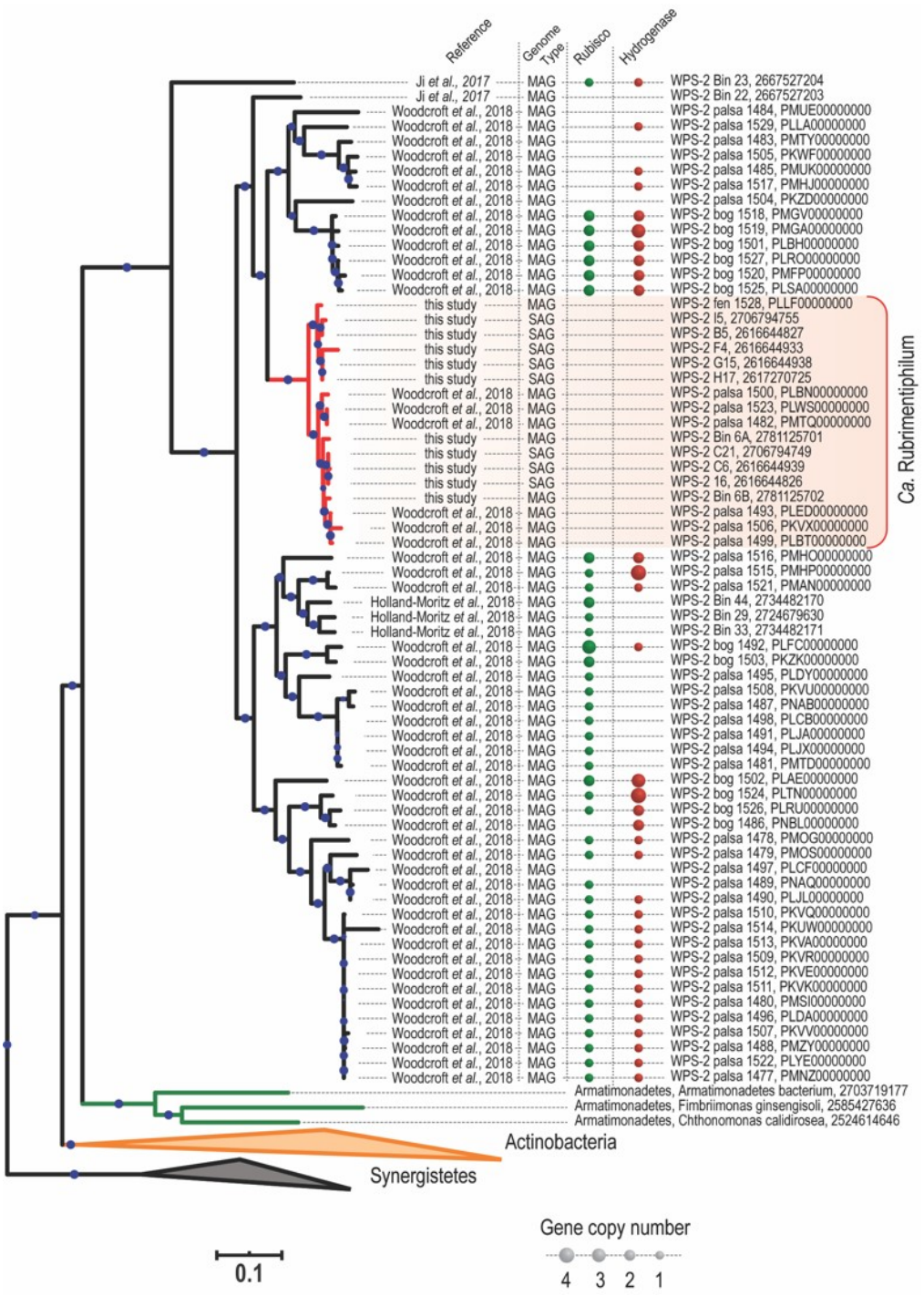
Reference
Genome Type
Rubisco
Hydrogenase
Ji et al., 2017 MAG WPS-2 Bin 23, 2667527204
Ji et al., 2017 MAG WPS-2 Bin 22, 2667527203
Woodcroft et al., 2018 MAG WPS-2 palsa 1484, PMUE00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1529, PLLA00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1483, PMTY00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1505, PKWF00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1485, PMUK00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1517, PMHJ00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1504, PKZD00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1518, PMGV00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1519, PMGA00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1501, PLBH00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1527, PLRO00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1520, PMFP00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1525, PLSA00000000
this study MAG WPS-2 fen 1528, PLLF00000000
this study SAG WPS-2 I5, 2706794755
this study SAG WPS-2 B5, 2616644827
this study SAG WPS-2 F4, 2616644933
this study SAG WPS-2 G15, 2616644938
this study SAG WPS-2 H17, 2617270725
Woodcroft et al., 2018 MAG WPS-2 palsa 1500, PLBN00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1523, PLWS00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1482, PMTQ00000000
this study MAG WPS-2 Bin 6A, 2781125701
this study SAG WPS-2 C21, 2706794749
this study SAG WPS-2 C6, 2616644939
this study SAG WPS-2 16, 2616644826
this study MAG WPS-2 Bin 6B, 2781125702
Woodcroft et al., 2018 MAG WPS-2 palsa 1493, PLED00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1506, PKVX00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1499, PLBT00000000
Ca. Rubrimentiphilum
Woodcroft et al., 2018 MAG WPS-2 palsa 1516, PMHO00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1515, PMHP00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1521, PMAN00000000
Holland-Moritz et al., 2018 MAG WPS-2 Bin 44, 2734482170
Holland-Moritz et al., 2018 MAG WPS-2 Bin 29, 2724679630
Holland-Moritz et al., 2018 MAG WPS-2 Bin 33, 2734482171
Woodcroft et al., 2018 MAG WPS-2 bog 1492, PLFC00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1503, PKZK00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1495, PLDY00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1508, PKVU00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1487, PNAB00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1498, PLCB00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1491, PLJA00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1494, PLJX00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1481, PMTD00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1502, PLAE00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1524, PLTN00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1526, PLRU00000000
Woodcroft et al., 2018 MAG WPS-2 bog 1486, PNBL00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1478, PMOG00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1479, PMOS00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1497, PLCF00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1489, PNAQ00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1490, PLJL00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1510, PKVQ00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1514, PKUW00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1513, PKVA00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1509, PKVR00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1512, PKVE00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1511, PKVK00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1480, PMSI00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1496, PLDA00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1507, PKVV00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1488, PMZY00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1522, PLYE00000000
Woodcroft et al., 2018 MAG WPS-2 palsa 1477, PMNZ00000000
Armatimonadetes, Armatimonadetes bacterium, 2703719177
Armatimonadetes, Fimbriimonas ginsengisoli, 2585427636
Armatimonadetes, Chthonomonas calidirosea, 2524614646
Actinobacteria
Synergistetes
0.1
Gene copy number
4 3 2 1

Figure 5. Bacterial genomes most similar to the WPS-2 genome (Bin 6A) based on Jaccard and Bray-Curtis similarities of Pfam profiles. Jaccard distances require only the presence or absence of a given Pfam, while Bray-Curtis distances include quantitative relationships of the Pfam profiles. Average distances over the entire dataset are indicated by red dashed lines. Extended data are shown in Supplementary Figure 6.

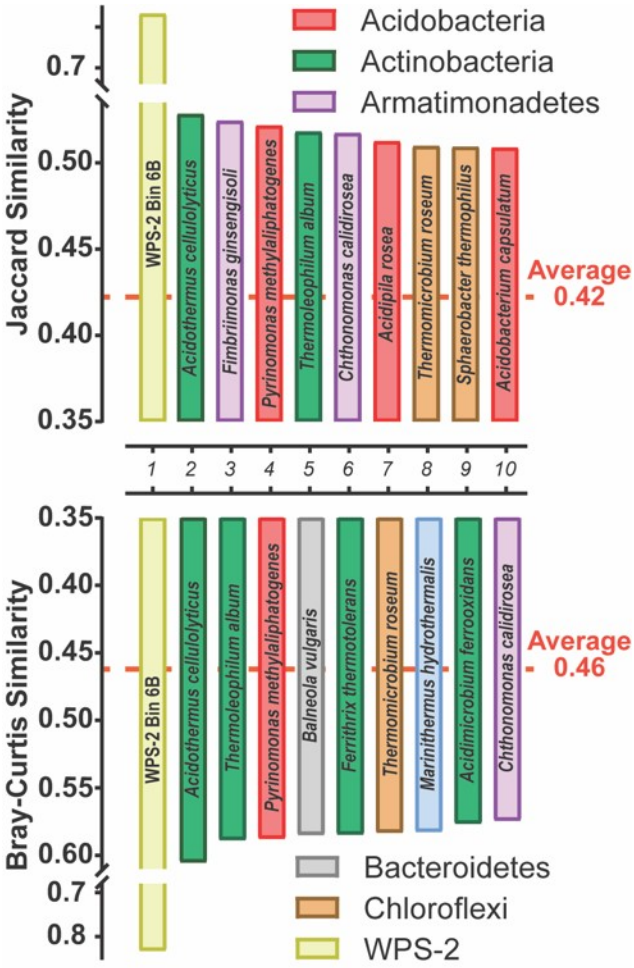

Figure 6. Comparison of gene distributions in selected Pfam categories for the *Ca.* Rubrimentiphilum MAGs versus different bacterial lineages based on a reference dataset of 2363 genomes (Supplementary File 2). Bars represent the mean percentages (±1 standard error) of the total genes that fall into a category for a taxonomic lineage. Group significance between the mean percentages within a given category were tested with one-way ANOVA. A post-hoc Bonferroni correction was used to detect significant differences between *Ca.* Rubrimentiphilum and other lineages (indicated by asterisks). Extended data are shown in Supplementary Figure 7.

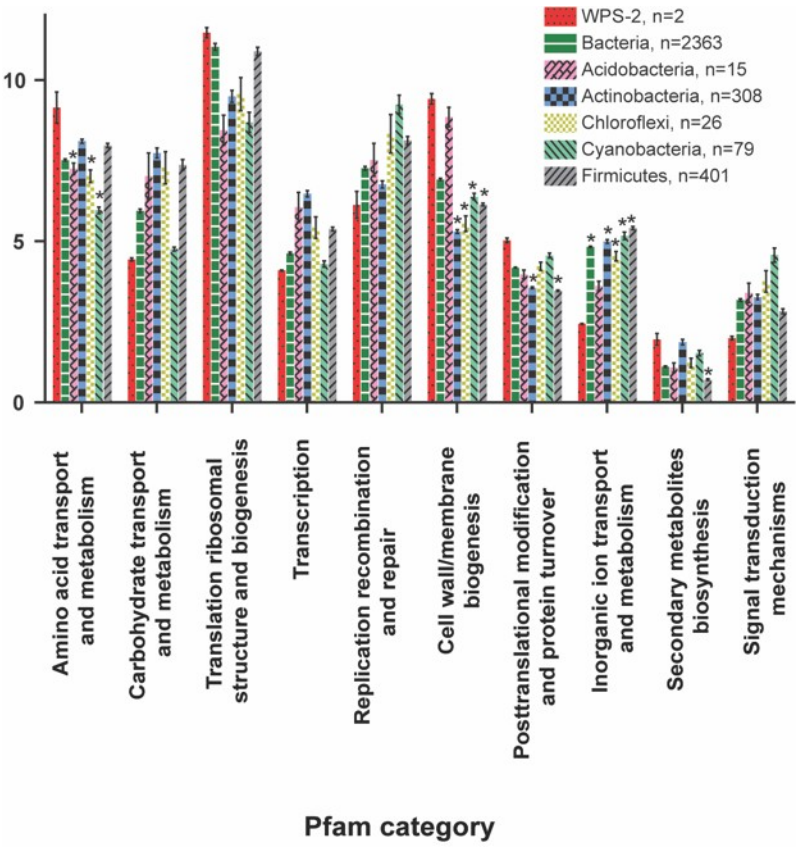

Figure 7. Predicted metabolic potential of *Ca*. Rubrimentiphilum based on the SAG and MAG data recovered in our study. The microorganisms exhibit Gram-negative cell wall structure, flagella, P and type IV pili. Transporters are scarce, and are mostly related to amino-acid and peptide transport. Amino acids are a likely growth substrate. The genomes show biosynthetic pathways for most vitamins, while crucial steps of B12 biosynthesis were missing.