Groundwater *Elusimicrobia* are metabolically diverse compared to gut microbiome *Elusimicrobia* and some have a novel nitrogenase paralog.

Raphaël Méheust[1,2], Cindy J. Castelle[1,2], Paula B. Matheus Carnevali[1,2], Ibrahim F. Farag[3], Christine He[1], Lin-Xing Chen[1,2], Yuki Amano[4], Laura A. Hug[5], and Jillian F. Banfield[1,2,*].

[1]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA 94720, USA

[2]Innovative Genomics Institute, Berkeley, CA 94720, USA

[3]School of Marine Science and Policy, University of Delaware, Lewes, DE 19968, USA

[4]Nuclear Fuel Cycle Engineering Laboratories, Japan Atomic Energy Agency, Tokai-mura, Ibaraki, Japan

[5]Department of Biology, University of Waterloo, ON, Canada

[*]Corresponding author: Email: jbanfield@berkeley.edu

Abstract

Currently described members of *Elusimicrobia*, a relatively recently defined phylum, are animal-associated and rely on fermentation. However, free-living *Elusimicrobia* have been detected in sediments, soils and groundwater, raising questions regarding their metabolic capacities and evolutionary relationship to animal-associated species. Here, we analyzed 94 draft-quality, non-redundant genomes, including 30 newly reconstructed genomes, from diverse animal-associated and natural environments. Genomes group into 12 clades, 10 of which previously lacked reference genomes. Groundwater-associated *Elusimicrobia* are predicted to be capable of heterotrophic or autotrophic lifestyles, reliant on oxygen or nitrate/nitrite-dependent respiration, or a variety of organic compounds and *Rhodobacter* nitrogen fixation-dependent (Rnf-dependent) acetogenesis with hydrogen and carbon dioxide as the substrates. Genomes from two clades of groundwater-associated *Elusimicrobia* often encode a new group of nitrogenase paralogs that co-occur with an extensive suite of radical S-Adenosylmethionine (SAM)

bioRxiv preprint doi: https://doi.org/10.1101/765248; this version posted July 16, 2020.

proteins. We identified similar genomic loci in genomes of bacteria from the Gracilibacteria phylum and the *Myxococcales* order and predict that the gene clusters reduce a tetrapyrrole, possibly to form a novel cofactor. The animal-associated *Elusimicrobia* clades nest phylogenetically within two free-living-associated clades. Thus, we propose an evolutionary trajectory in which some *Elusimicrobia* adapted to animal-associated lifestyles from free-living species via genome reduction.

Introduction

Elusimicrobia is an enigmatic bacterial phylum. The first representatives were termite gut-associated [1], although one 16S ribosomal ribonucleic acid (rRNA) gene sequence was identified in a contaminated aquifer [2]. Initially referred to as Termite Group 1 (TG1) [3], the phylum was renamed as *Elusimicrobia* after isolation of a strain from a beetle larvae gut [4]. *Elusimicrobia* is part of the *Planctomycetes-Verrucomicrobia-Chlamydia* superphylum (PVC). Currently available sequences of *Elusimicrobia* form a monophyletic group distinct from other PVC phyla which include *Planctomycetes*, *Verrucomicrobia*, *Chlamydia*, *Omnitrophica*, *Desantisbacteria* [5], *Kiritimatiellaeota* [6] and *Lentisphaerae* [7].

The current understanding of *Elusimicrobia* mostly relies on a single taxonomic class for which the name *Endomicrobia* has been proposed [8]. *Endomicrobia* comprise abundant members of the hindgut of termites, cockroaches and other insects, as well as in rumen where they occur as endosymbionts [9], or ectosymbionts [10] of various flagellated protists, or as free-living bacteria [11]. Of the three *Endomicrobia* genomes that have been described, all belong to the genus *Endomicrobium*. One is for a free-living *E. proavitum* [12] and two are *E. trichonymphae* Rs-D17 endosymbionts, genomovar Ri2008 [13] and genomovar Ti2015 [14]. The fourth *Elusimicrobia* genome available is for *Elusimicrobium minutum* strain Pei191^{T} [15], which was isolated from a beetle larva gut [4]. *E. minutum* forms a distinct monophyletic family-level lineage of gut-adapted organisms for which the name *Elusimicrobiaceae* was proposed [4]. Cultivation and genome-based studies revealed that *E. proavitum* and *E. minutum* strain Pei191^{T} are strictly anaerobic ultramicrobacteria capable of glucose fermentation [4, 16, 17].

Despite prior focus on gut-associated *Elusimicrobia*, phylogenetic analysis of environmental sequences revealed numerous novel and distinct lineages from a wide range of habitat types, including soils, sediments and groundwater [16, 18]. Moreover, several published metagenomics analyses reconstructed genomes from *Elusimicrobia* but none of these studies analyzed them in detail [5, 19, 20]. We augmented the sampling with 30 unpublished genomes from animal-associated and groundwater-associated metagenomes from sequencing efforts that targeted all bacteria and archaea in the metagenomes. Here, we present a comparative genomic analysis of 94 draft-quality and non-redundant genomes from diverse environments. We identified 12 lineages, including 10 that previously lacked genomic representatives. We predict numerous traits that constrain the roles of *Elusimicrobia* in biogeochemical cycles, identify a new nitrogenase paralog, and infer the evolutionary processes involved in adaptation to animal-associated lifestyles.

Materials and Methods

Elusimicrobia genome collection

Thirty unpublished genomes were added to the 120 genomes downloaded from the genome NCBI database in June 2018.

Six genomes were obtained from groundwater samples from Genasci Dairy Farm, located in Modesto, California (CA) and from Serpens Ridge, a private property located in Middletown, CA. Over 400 L of groundwater were filtered from monitoring well 5 on Genasci Dairy Farm, located in Modesto, CA, over a period ranging from March 2017 to June 2018. Over 400 L of groundwater were filtered from Serpens Ridge, a private property located in Middletown, CA in November 2017. Deoxyribonucleic acid (DNA) was extracted from all filters using Qiagen DNeasy PowerMax Soil kits and ~10 Gbp of 150 bp, paired end Illumina reads were obtained for each filter. Scaffolds were binned on the basis of GC content, coverage, presence of ribosomal proteins, presence/copies of single copy genes, tetranucleotide frequency, and patterns of coverage across samples. Bins were obtained using manual binning on ggKbase [21], Maxbin2 [22], CONCOCT [23], Abawaca1, and Abawaca2 (https://github.com/CK7/abawaca), with DAS Tool [24] used to choose the best set of bins from all programs. All bins were manually checked to remove incorrectly assigned scaffolds using ggKbase.

The remaining 23 genomes came from previous sequencing efforts. In brief, three genomes were obtained from sediment samples of Tibet hot springs in 2016. The samples were collected as previously described [25], and for DNA processing and sequencing methods refer to [26]. One genome was obtained from sediment samples collected from Zodletone spring, an anaerobic sulfide and sulfur-rich spring in Western Oklahoma. Detailed site descriptions and metagenomic sequencing were previously reported in [27]. Seven genomes were obtained from groundwater samples collected from the Mizunami and the Horonobe Underground Research Laboratories (URL) in Japan. For sampling refer to [28] and [29], and for DNA processing and sequencing methods refer to [29]. Six genomes were obtained from an aquifer adjacent to the Colorado River near the town of Rifle, Colorado, USA in 2011 [19] one genome from the Crystal Geyser system in Utah, USA [30]. For DNA processing and sequencing methods see [5, 19]

Finally, six genomes were assembled from mammal microbiome raw data used in previous studies following the methods described in [31] (**Suppl. Dataset 1**).

Genome completeness assessment and de-replication.

Genome completeness and contamination were estimated based on the presence of single-copy genes (SCGs) as described in [19]. Genome completeness was estimated using 51 SCGs, following [19]. Genomes with completeness > 70% and contamination < 10% (based on duplicated copies of the SCGs) were considered as draft-quality genomes. Genomes were de-replicated using dRep (version v2.0.5 with average nucleotide identity (ANI) > 99%) [32]. The most complete genome per cluster was used in downstream analyses.

16S rRNA gene phylogeny

To construct a comprehensive 16S rRNA gene tree, all gene sequences assigned to the *Elusimicrobia* in the SILVA database [33] were exported. Sequences longer than 750 bp were clustered at 97% identity using uclust, and the longest representative gene for each cluster included in the phylogenetic tree. All 16S rRNA genes associated with a binned *Elusimicrobia* genome were added to the SILVA reference set, along with an outgroup of *Planctomycetes*, *Verrucomicrobia*, *Chlamydia*, *Omnitrophica*, *Desantisbacteria*, *Kiritimatiellaeota* and *Lentisphaerae* rRNA gene sequences from organisms with published genomes (45 outgroup sequences total). The final dataset comprised 711 sequences, which were aligned together using the SILVA SINA alignment algorithm. Common gaps and positions with less than 3% information were stripped from the alignment, for a final alignment of 1,593 columns. A maximum likelihood phylogeny was inferred using RAxML [34] version 8.2.4 as implemented on the CIPRES high performance computing cluster [35], under the GTR model of evolution and the MRE-based bootstopping criterion. A total of 408 bootstrap replicates were conducted, from which 100 were randomly sampled to determine support values.

Concatenated 16 ribosomal proteins phylogeny

A maximum-likelihood tree was calculated based on the concatenation of 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19). Homologous protein sequences were aligned using Muscle [36], and alignments refined to remove regions of ambiguity by removing columns with less than 3% information and manually removing abberrent N and C-terminal extensions. The protein alignments were concatenated, with a final alignment of 147 genomes and 2,388 positions. The tree was inferred using RAxML [34] (version 8.2.10) (as implemented on the CIPRES web server [35]), under the LG plus gamma model of evolution, and with the number of bootstraps automatically determined via the MRE-based bootstopping criterion. A total of 108 bootstrap replicates were conducted, from which 100 were randomly sampled to determine support values.

GTDB taxonomic assignment

Taxonomy assignment was performed for each genome using the GTDB-Tk software (v1.0.1) (default parameters) [37].

Protein clustering

Protein clustering into families was achieved using a two-step procedure as previously described in [38]. A first protein clustering was done using the fast and sensitive protein sequence searching software MMseqs2 (version 9f493f538d28b1412a2d124614e9d6ee27a55f45) [39]. An all-vs-all search was performed using e-value: 0.001, sensitivity: 7.5 and cover: 0.5. A sequence similarity network was built based on the pairwise similarities and the greedy set cover algorithm from MMseqs2 was performed to define protein subclusters. The resulting subclusters were defined as subfamilies. In order to test for distant homology, we grouped subfamilies into protein families using an HMM-HMM comparison procedure as follows: the proteins of each subfamily with at least two protein members were aligned using the result2msa parameter of mmseqs2, and, from the multiple sequence alignments, HMM profiles were built using the HHpred suite (version 3.0.3) [40]. The subfamilies were then compared to each other using hhblits [41] from the HHpred suite (with parameters -v 0 -p 50 -z 4 -Z 32000 -B 0 -b 0). For subfamilies with probability scores of ≥ 95% and coverage ≥ 0.50, a similarity score (probability X coverage) was used as weights of the input network in the final clustering using the Markov Clustering algorithm [42], with 2.0 as the inflation parameter. These clusters were defined as the protein families.

Phylogenetic analyses of individual protein sequences.

Each individual tree was built as follows. Sequences were aligned using MAFFT (version 7.390) (--auto option) [43]. Alignment was further trimmed using Trimal (version 1.4.22) (--gappyout option) [44]. Tree reconstruction was performed using IQ-TREE (version 1.6.6) (as implemented on the CIPRES web server [35]) [45], using ModelFinder [46] to select the best model of evolution, and with 1000 ultrafast bootstrap [47].

Protein sequence annotation

Protein sequences were functionally annotated based on the accession of their best Hmmsearch match (version 3.1) (E-value cut-off 0.001) [48] against an HMM database constructed based on ortholog groups defined by the KEGG [49] (downloaded in June 10, 2015). Domains were predicted using the same Hmmsearch procedure against the Pfam database (version 31.0) [50]. The domain architecture of each protein sequence was predicted using the DAMA software (version 1.0) (default parameters) [51]. SIGNALP (version 4.1) (parameters: -f short -t gram-) [52] was used to predict the putative cellular localization of the proteins. Prediction of transmembrane helices in proteins was performed using TMHMM (version 2.0) (default parameters) [53]. The transporters were predicted using BLASTP (version 2.6.0) [54] against the TCDB database (downloaded in February 2019) (keeping the best hit, e-value cut-off 1e-20) [55].

Metabolic pathways annotation

The Embden-Meyerhof pathway (module M00001), the pentose phosphate pathway (modules M00006 and M00007), cobalamin biosynthesis (M00122) and pyruvate oxidation (module M00307) were considered present based on the completeness of their corresponding KEGG modules (complete if at least 80% of the KEGG module, partial if between 50 and 80%, absent otherwise). The capacity of synthesizing common energy-storage polysaccharides (starch or glycogen) was considered if both the starch/glycogen synthase (KEGG accessions K00703, K20812 or K13679) and 1,4-alpha-glucan branching enzyme were present (K00700 or K16149) [56]. F-type (module M00157, 8 subunits) and V/A-type ATPases (module M00159, 9 subunits) were considered complete if at least 80% of the subunits were present, partial if between 50 and 80%, absent otherwise). The tricarboxylic acid (TCA) cycle was considered complete if 70% of the key enzymes represented by their KEGG accessions were present (partial if between 50 and 70%, absent otherwise). The key enzymes of the TCA cycle include the aconitate hydratase (K01681 or K01682), the fumarate hydratase (K01676 or K01679 or K01677 and K01678), the isocitrate dehydrogenase (K00031 or K00030), the citrate synthase (K01647 or K05942), the malate dehydrogenase (K00024 or K00025 or K00026 or K00116), the succinate dehydrogenase (K00239, K00240 and K00241), the 2-oxoglutarate dehydrogenase (K00174 and K00175) and the succinyl-CoA synthetase (K01899 and K01900 or K01902 and K01903 or K08692 and K01903). The *Rhodobacter* nitrogen fixation (Rnf) complex was considered complete in a genome if at least 4 out of 6 subunits were found in operon along the genome (subunits ABCDEG represented by the KEGG accessions K03617, K03616, K03615, K03614, K03613 and K03612 respectively). The Wood–Ljungdahl (WL) pathway is divided into two branches, the methyl and the carbonyl branches. For the carbonyl branch, reference sequences of the five subunits of the carbon monoxide dehydrogenase/acetyl-coenzyme A (CoA) synthase (CODH/ ACS) were investigated. The AcsE (KPK98995.1 and KPJ61844.1), the CdhA/AcsA (KPK97464.1 and KPJ58813.1), the CdhC/AcsB (KPK97461.1 and KPJ58814.1), the CdhD/AcsD (KPK98994.1 and KPJ61843.1) and the CdhE/AcsC (KPK98991.1 and KPJ61839.1) protein sequences from *Planctomycetes bacterium* DG_23 and *Omnitrophica* WOR_2 bacterium SM23_72 [57] were searched against *Elusimicrobia* to retrieve similar sequences using BLASTP (e-value $1e^{-10}$) (version

2.6.0+) [54]. The carbonyl branch was considered complete if 4 out of 5 subunits were present (partial if 3 subunits). The methyl branch were considered complete if the alpha subunit of the formate dehydrogenase (NADP+) (K05299), the formate--tetrahydrofolate ligase (K01938), the methylenetetrahydrofolate dehydrogenase (NADP+)/methenyltetrahydrofolate cyclohydrolase (folD) (K01491), methylenetetrahydrofolate reductase (metF, MTHFR) (K00297) or if three key enzymes were detected on the same scaffold (partial if not on the same scaffold). The butyrate fermentation pathway was considered complete in a genome if the two subunits of the acetate CoA/acetoacetate CoA-transferase (K01034 and K01035), the butyrate kinase (K00929) and the butyryl-CoA dehydrogenase (K00248) were present. The acetate metabolism pathway was considered complete if both the acetate/propionate kinase (K00925 or K00932) and the phosphate acetyltransferase (K00625, K13788 or K15024) were present or if the two subunits of the acetate-CoA ligase were present (K01905 and K22224). The ethanol fermentation pathway was considered complete if both the aldehyde dehydrogenase (K00128, K00129, K14085, K00149 or K00138) and the alcohol dehydrogenase (K00114, K13951, K13980, K13952, K13953, K13954, K00001, K00121 or K18857) were present or if the multifunctional aldehyde-alcohol dehydrogenase (encoded by the *adhE* gene, K04072) that catalyzes the sequential reduction of acetyl-CoA to acetaldehyde and then to ethanol under fermentative conditions was present [58]. Lactate fermentation was considered present if the lactate dehydrogenase was present (K00016) and malate fermentation if malate dehydrogenase was present (K00024) [59]. We searched the *Elusimicrobia* genomes for evidence of uroporphyrinogen III synthesis by looking for genes encoding the porphobilinogen synthase (K01698), the hydroxymethylbilane synthase (K01749) and the uroporphyrinogen synthase (K01719). These three enzymes are part of the three-enzymatic-step core pathway from 5-aminolevulinate to uroporphyrinogen III [60]. Based on KEGG accessions, we also annotated the ammonium transporter (K03320), the transhydrogenase (NfnAB) (K00324, K00325), the electron transfer flavoprotein (EtfAB) (K03522, K03521), the ferredoxin nitrite reductase (NirK and NirS) (K00368, K15864), the cytochrome-c nitrite reductase (NfrAH) (K03385, K15876), the phosphotransferase system (PTS) glucose-specific IIC component (K02779), the PTS Nacetylglucosamine-specific IIC component (K02804), the PTS sucrose-specific IIC component (K02810), the PTS maltose/glucose-specific IIB component (K02790), the PTS fructose-specific IIA component (K02768) and the PTS mannose-fructose-sorbose family (K02793, K02794, K02795 and K02796). The succinate dehydrogenase was considered complete in a genome if at least 3 out of 4 subunits were found in operon along the genome (K00239, K00240, K00241, K00242) (KEGG module M00149). The cytochrome *bd* ubiquinol oxidase was considered complete in a genome if the two subunits were found in operon along the genome (K00425 and K00426). The cytochrome *bo*$_3$ was considered complete if two subunits out of four were present (K02297, K02298, K02299 and K02300). Hydrogenases sequences were retrieved using Hmmsearch (version 3.1) (E-value cut-off 0.001) [48] and custom HMMs built from previous studies [61, 62]. All hydrogenases were annotated based on the reconstruction of phylogenetic trees and careful inspection of the genes next to the catalytic subunits. A similar strategy was employed to annotate the formate dehydrogenase, the alternative complex III, the nitrite oxidoreductase (NxrA) and the nitrate reductases (NapA, NasA, and NarB) thanks to a phylogenetic tree of the dimethyl sulfoxide (DMSO) reductase superfamily [61, 63] and to annotate the types AAA and cbb cytochrome oxidase and quinol nitric oxide reductase (qNOR) thanks to a phylogenetic tree of the catalytic subunit I of heme-copper oxygen reductases [61].

Results and Discussion

Selection and phylogenetic analysis of a non-redundant set of genomes

As of June 2018, 120 genomes assigned to *Elusimicrobia* were available in the NCBI genome database. We augmented these genomes with 30 newly reconstructed genomes from metagenomic samples from Zodletone spring sediments (Oklahoma, USA), groundwater or aquifer sediments (Serpens Ridge, California, USA; Rifle, Colorado, USA; Genasci dairy, Modesto, California, USA; Horonobe, Japan), hot spring sediments (Tibet, China), and the arsenic impacted human gut (see Materials and Methods). Genomes with >70% completeness and <10% contamination were clustered at ≥ 99% ANI, and 113 genomes representative of the clusters were selected based on completeness and contamination metrics. Among them, 19 genomes were missing many ribosomal proteins and thus were excluded, resulting in a final dataset of 94 non-redundant *Elusimicrobia* genomes of good quality (median completeness of 92%, 69 genomes were >90% completeness, **Suppl. Dataset 1**). Of the 94 genomes, 23 are from intestinal habitats and 71 come from other habitats, mostly groundwater-associated (**Suppl. Dataset 1**). Although some groundwater genomes are publicly available [19], none have been analyzed in detail.

The *Elusimicrobia* genomes were further classified into lineages, consistent with the previous classification based on 16S rRNA gene sequences [4, 16]. In order to improve the phylogenetic tree, we also added 16S rRNA gene sequences from genomes that were not originally chosen as representative genomes. The 16S rRNA genes were binned for 41 of the 94 genomes, 31 out of 41 belong to lineages I (*Endomicrobia*), IIa, IIc, III (*Elusimicrobiaceae*), IV, V and VI. Thus, the 16S rRNA gene sequences investigated here span 7 of the 9 previously defined lineages (**Figure S1**) [4, 16]. The remaining 10 sequences do not belong to the well-defined lineages [4, 16] but instead cluster in groups defined by Yarza *et al* [64]. In total, 9 sequences were considered as new phyla by Yarza *et al* whereas the remaining 32 sequences cluster into one single phylum. We compared the classification based on the 16S rRNA sequences with the recent classification proposed by the Genome Taxonomy Database (GTDB) based on concatenated protein phylogeny [65] and found that the 94 genomes are classified into three potentially phylum-level lineages. Of these, 89 genomes genuinely belong to the *Elusimicrobia* phylum (**Suppl. Dataset 1**).

To clarify the taxonomy, we reconstructed a phylogenetic tree based on 16 ribosomal proteins (RPs) and mapped the results of the classifications based on the 16S rRNA and the GTDB (**Figure 1, Figure S2 and Suppl. Dataset 1**). The 94 genomes form a supported monophyletic group (**Figure 1A**, bootstrap: 95) with *Desantisbacteria* and *Omnitrophica* as sibling phyla (**Figure 1A**, bootstrap: 89). The bootstrap support (75) was insufficient to confirm whether *Desantisbacteria* or *Omnitrophica* is the most closely related phylum. All of the gut-associated genomes are clustered into the two known lineages, i.e., *Endomicrobia* (lineage I) and *Elusimicrobiaceae* (lineage III). The *Endomicrobia* lineage formed a distinct clade (n=6, bootstrap: 100) that includes the three earliest described *Endomicrobium* genomes and a new genome from the rumen of sheep [66]. *Elusimicrobiaceae*, which includes genomes related to *Elusimicrobium minutum*, contains genomes from animal habitats, with the exception of one genome from palm oil mill effluent. Of note, two newly reported *Elusimicrobiaceae* genomes in this group are from the gut microbiomes of humans living in Bangladesh [31] and in Tanzania [67]. This is consistent with the recent finding that *Elusimicrobia* genomes are present in the gut microbiomes of non-westernized populations [68]. As previously suggested [4], the lineage III clade (n = 19) is nested within three clades comprising bacteria from groundwater environments (lineages IV, V and VI, n = 38). Based on extrapolated genome sizes, *Elusimicrobiaceae* genomes were significantly smaller than those of sibling lineages IV, V and VI (**Figure S3**). The overall placement and short branch lengths within the *Elusimicrobiaceae* raise the possibility that the *Elusimicrobiaceae* adapted and diversified recently from ancestral groundwater organisms from lineages IV, V and VI. Eleven genomes clustered into lineages IIa (n=2, bootstrap: 41) and IIc (n=9, bootstrap: 100). The lack of bootstrap support does not allow us to confirm the relationship between the two lineages. None of the 94 genomes were classified as lineages IIb and IId, due either to a true absence of IIb and IId genome sequences or because lineages IIb and IId genomes in the dataset lack 16S rRNA gene sequences.

The 22 genomes that were not assigned to the previously defined *Elusimicrobia* lineages comprise five clades, of which four are basal to previously reported *Elusimicrobia* lineages. Sixteen of these 22 genomes are from groundwater, and the other 6 genomes were from sediment, oil sand and peat metagenomes. We named three of the five *Elusimicrobia* lineages VII (3 genomes), VIII (8 genomes) and IX (6 genomes), following the current naming procedure. The two other lineages are basal, and are considered as potentially different phyla by both the GTDB and the 16S rRNA gene classification (Figures S1 and S2). These are represented by just five genomes. We assigned these genomes to *Elusimicrobia*-related lineages 1 (ERL1, 2 genomes) and 2 (ERL2, 3 genomes), but the low level of sampling precludes their definitive classification as phyla (**Figure 1 and Figure S2**).

Our analysis captured most of the currently known phylogenetic diversity based on the position in the 16S rRNA gene tree (7 out of the 9 lineages have now a genome). We also discovered 5 new clades based on the ribosomal proteins tree including two phylum-level lineages (**Figure 1**). Many 16S rRNA gene sequences come from soil, especially those from lineages IIa, IIb, IIc, IId and IV (**Figure S1**) [16]. However, the current dataset only includes three genomes from peat environments [20]. Reconstructing genomes from soil is notoriously difficult, as most soils have extremely high microbial diversity [69, 70] and might explain the small number of genomes recovered from this biome.

<u>A novel nitrogenase paralog possibly involved in the biosynthesis of a novel cofactor</u>

A previous study reported the nitrogen fixing ability of *Endomicrobium proavitum* by an unusual nitrogenase belonging to group IV [17] (**Figure 2**). Dinitrogenase reductase, encoded by *nifH*, donates reducing equivalents to a complex encoded by *nifD* and *nifK*. Canonically, N_2 fixing NifH proteins belong to groups I, II and III [71]. Analysis of the amino acids coordinating the P-cluster and FeMoCo ligands in NifD indicates that the homologs in the subcluster IVa are functional reductases. Together with the documentation of diazotrophy in *E. proavitum*, there is no doubt that some NifH paralogs of group IV comprise functional nitrogenases [17]. Other group IV NifH proteins, non affiliated to the subcluster IVa, are implicated in the biosynthesis of cofactor F430, the prosthetic group of methyl coenzyme M reductase, which catalyzes methane release in the final step of methanogenesis [72, 73]. Another NifH paralog, phylogenetically defined as group V, is involved in chlorophyll biosynthesis [74]. In this case, protochlorophyllide is converted to chlorophyllide via the BchLNB complex in which BchL is the NifH paralog and BchN and BchB are the NifD and NifK paralogs, respectively.

We searched the *Elusimicrobia* genomes for *nifH*, *nifD* and *nifK* and identified homologues in 24 *Elusimicrobia* genomes from lineages I, IV and V. We also searched for accessory genes required for the assembly of the M- and P- metallo clusters required by the nitrogenase complex. Two genomes from lineage I, RIFOXYA2_FULL_50_26 and the previously reported *E. proavitum*, contain both the nitrogenase subunits and the accessory proteins. The NifH protein from RIFOXYA2_FULL_50_26 places phylogenetically into NifH group II, and thus is plausibly associated with N_2 fixation [71].

The other NifH proteins all belong to lineages IV and V of *Elusimicrobia* and place phylogenetically outside of all previously described groups of NifH (**Figure 2**). Instead, the NifH sequences from 22 *Elusimicrobia* genomes form a new group that is sibling to group V (**Figure 2**). Consistent with the new placement, these 22 *nifH*-encoding genomes lack the accessory genes required for the assembly of the M- and P- metallo clusters of nitrogenase. We designate these genes as a new group, group VI, and infer that these paralogous proteins may have a distinct and perhaps previously undescribed biological role. The group VI NifH proteins contain the GXGXXG consensus motif for the binding of MgATP and two cysteine (Cys) residues (Cys^{97} and Cys^{132}) that bridge the two subunits through a $[Fe_4S_4]$ cluster (**Figure S4**). However, the associated *nifD-* and *nifK-like* genes are highly divergent from true nitrogenase genes (25.51 and 22.74 amino acid percent identity on average respectively) (**Figure S5**). Importantly, they lack some conserved cysteine motifs that are involved in the attachment of the P-clusters [71] (**Figure S5**).

Interestingly, nitrogenase paralogs complexes from group IV and V each modify a tetrapyrrole molecule by reducing a carbon-carbon double bound [72, 74, 75]. Biosynthesis of cofactor F430 involves the sirohydrochlorin precursor and biosynthesis of chlorophyll involves a protoporphyrin precursor, both of which derive from uroporphyrinogen III (also a building block for cobalamin). In a subset of genomes with the novel group IV *nifH* genes (five genomes of lineage IV), we identified the capacity to produce uroporphyrinogen III (for example, see *Elusimicrobia* bacterium GWA2_69_24). The absence of precursor biosynthesis pathways in other genomes of *Elusimicrobia* predicted to have the capacity to make nitrogenase clusters does not rule out an analogous function, as many bacteria scavenge such molecules (e.g., cobalamin; [76] or haem [77]).

We examined the genomic neighborhoods to get insights regarding the function of the novel group VI paralogs. The lineage IV genomes encode several copies of *nifK*, *nifD* and *nifH* whereas most genomes of lineage V genomes have only one copy of each subunit (**Figure S6**). Strikingly, many adjacent genes encode radical S-adenosylmethionine (SAM) proteins. Radical SAM proteins have many functions, including catalysis of methylation, isomerization, sulfur insertion, ring formation, anaerobic oxidation, and protein radical formation, and also in the biosynthesis of cofactors, DNA and antibiotics [78, 79]. The copy number of radical SAM genes varies greatly across the genomes, from no radical SAM genes to 13 copies in close proximity to the nitrogenase paralogs in the genome of GWC2_Elusimicrobia_56_31 (**Figure S6**). Several radical SAM genes are fused with B12-binding domains (SR-2 Biohub_180515 Elusimicrobia_69_71, GWC2_Elusimicrobia_56_31 and *Elusimicrobia bacterium* GWF2_62_30) and/or HemN_C domain (*Elusimicrobia bacterium* GWA2 _69_24) (**Figure S6**). The B12-binding domain is involved in binding cobalamin while the HemN_C domain has been suggested to bind the substrate of coproporphyrinogen III oxidase [80]. Both substrates have tetrapyrrole structures, which is consistent with the phylogenetic position of the NifH near the groups IV and V NifH clades (**Figure 2**). Finally, we constructed a Hidden Markov Model using the new group VI *Elusimicrobia* NifH sequences and searched for homologous sequences in genomes from other phyla. Interestingly, we found homologs of the group VI NifH in two genomes of Gracilibacteria, a phylum within the bacterial Candidate Phyla Radiation [81], and in fourteen *myxococcales* genomes from aquatic environments (**Figure 2 and Suppl. Dataset 2**). Together, these observations suggest that the enzymes likely do not perform nitrogen fixation, but have an alternative function that may be related to the biosynthesis of a tetrapyrrole cofactor similar to chlorophyll or F430 cofactors.

Overall metabolic potential in *Elusimicrobia* and related genomes

The availability of good quality genomes allowed us to compare their gene contents. We clustered the protein sequences for all genomes to generate groups of homologous proteins (protein families) in a two-step procedure (see Materials and Methods) (**Figure S7**) [38]. The objective was to compare the proteomes across the genomes. Our approach is agnostic and unbiased by preconceptions about the importance of genes, and allows us to cluster protein sequences with no homology within annotation databases. This resulted in 6,608 protein families that were present in at least three distinct genomes. We constructed an array of the genomes versus the protein families, and sorted the families based on profiles of genome presence/absence. Numerous protein families group together due to co-existence in multiple genomes (**Figure S7**). Several blocks of families are fairly lineage specific. This is particularly apparent for the gut-associated *Elusimicrobiaceae* lineage (lineage III) which lacks numerous families abundant in other lineages but also contains 161 families abundant in *Elusimicrobiaceae* but rare or absent in other lineages (red boxes in **Figure S7**). Unlike lineage III, lineage V encodes extended sets of protein families (blue box in **Figure S7**) consistent with their larger genome size (**Figure S3**). Other lineages also have enriched groups of families, although it is less apparent than for lineages III and V (**Figure S7**). Overall, the patterns of presence/absence of protein families are consistent with the lineages defined by the ribosomal protein phylogeny (e.g., genomes from the same lineage tend to have a similar protein families set) and may reflect different metabolic strategies.

Previously represented by gut-associated lineages, the expansion of *Elusimicrobia* identifies a broad variety of energetic strategies within the phylum. Non-gut-associated *Elusimicrobia* display substantial metabolic versatility, with the genomic potential for both autotrophic and heterotrophic-based lifestyles, though not usually both within the same genome or lineage (**Figure 3**). Most of the non-gut-associated genomes are predicted to have the capacity for sugar fermentation to acetate, malate, butyrate and/or ethanol, generating ATP via substrate-level phosphorylation (**Suppl. Dataset 3**). We also identified a large variety of membrane-bound and cytoplasmic [NiFe] and [FeFe] hydrogenases (**Suppl.**

Dataset 3), some of which may be electron bifurcating. Other electron bifurcating complexes, such as the transhydrogenase (NfnAB) or the Bcd/EtfAB complex are also present in these genomes (**Suppl. Dataset 3**), suggesting the ability to minimize free energy waste and to optimize energy conservation [82]. This large repertoire of metabolic capacities differs from their animal habitat-residing counterparts (lineages I and III), which rely solely on fermentation for energy [12, 13, 15, 17] (**Figure 3**).

Elusimicrobia-related lineages 1 and 2

Analyses of the five genomes assigned to ERL1 and ERL2 (**Figure 1**) suggest that they are likely obligate fermenters. All of the genomes lack a complete tricarboxylic acid cycle, NADH dehydrogenase (complex I), and most other complexes from the oxidative electron transport phosphorylation chain (e.g., complex II, complex III, complex IV, and quinones). Interestingly, all but one genome has partial ATP synthase, so the functionality of this complex V remains uncertain. All genomes included in this study have complete or near-complete glycolysis and/or pentose phosphate pathway(s) and are predicted to have the capacity to produce acetate, lactate, and/or hydrogen as byproducts of fermentation (**Suppl. Dataset 3**). Acetate kinase and phosphotransacetylase have been identified and are likely involved in acetate and ATP production. Thus, we cannot establish that they have a functional complete ATPase, they should be able to produce ATP via substrate level phosphorylation. These organisms may also be capable of synthesizing common energy-storage polysaccharides, as we identified several genes encoding enzymes for starch or glycogen metabolism (**Suppl. Dataset 3**). Overall, we predict widespread fermentation-based metabolism in the nearest neighbour lineages of *Elusimicrobia*.

Diverse respiratory strategies in groundwater-associated *Elusimicrobia*

Oxidative phosphorylation and the tricarboxylic acid cycle are commonly encoded in genomes from lineages IIa, IIc, IV, V and VI (**Figure 4A**). Lineage V genomes encode a canonical NADH dehydrogenase (complex I with 14 subunits) and succinate dehydrogenase (complex II) that link electron transport to oxygen as a terminal electron acceptor via the high-affinity oxygen cytochrome *bd* oxidase (complex IV). Unlike genomes encoding electron transport chains, lineage V genomes consistently lack a complex III (bc_1-complex, cytochrome c reductase). Organisms that carry one or more bd-type oxidases usually also possess at least one heme-copper oxygen reductase [83]. For instance, *Escherichia coli* also lacks complex III and/or alternative complex III but carries a cytochrome bo_3, a proton-pumping oxidoreductase [83, 84]. However, we did not detect genes encoding cytochrome bo_3 in genomes of lineage V (**Suppl. Dataset 3**). The absence of any heme–copper oxygen reductase was also observed in organisms such as *Lactobacillus plantarum* [85] and *Zymomonas mobilis* [86]. As all of these organisms are capable of oxygen-based respiration, we conclude that there may be a variety of ways to circumvent the lack of complex III (or alternative complex III). Thus, we cannot rule out the possibility that these *Elusimicrobia* have a functional aerobic electron transport chain.

Several genomes from lineages IIa, IIc, IV, VI and VI have the genomic potential for respiring a variety of organic compounds (including ribose, galactose, glucose, acetate and possibly propionate and butyrate) as energy and carbon sources. Further, the genomes indicate the capacity for utilization of fatty acids via the ß-oxidation pathway. Examination of the genomes revealed the presence of glycosyl hydrolases that would support growth by utilization of externally derived sugar polymers. Indeed, the organisms from lineages IIa, IIc, IV, VI and VI may also be capable of synthesizing and/or utilizing common energy-storage polysaccharides, as we identified several genes encoding enzymes for starch, glycogen, dextrin metabolism (**Figure 4A**).

Lineage V *Elusimicrobia* typically have a suite of hydrogenases, where hydrogenases are sparsely distributed in other lineages. Some groundwater-associated lineage V genomes have trimeric group A [FeFe] hydrogenases directly downstream from a monomeric group C hydrogenase, related to those seen in *Ignavibacterium album* and *Caldithrix abyssi* [62] (**Figure S8**). In general, [FeFe] hydrogenases can either use or produce H_2 whereas group C hydrogenases are co-transcribed with regulatory elements and are predicted to sense H_2 [87]. In most group C hydrogenases this sensing occurs via a Per-ARNT-Sim (PAS) domain [62, 88]. However, the *Elusimicrobia* hydrogenase seems to be fused to a histidine kinase HATPase domain, probably fulfilling this sensory function. Additionally, the genomic neighborhood of these hydrogenases includes several regulatory genes encoding proteins such as Rex, which is known to sense NADH to NAD^+ ratios for transcriptional regulation [89]. Groundwater-associated genomes of lineage V also encode genes for different types of [NiFe] hydrogenases. Seven genomes encode cytoplasmic group 3b (NADP-reducing) [NiFe] hydrogenases that are likely bidirectional (also known as sulfhydrogenase). Four other genomes also have cytoplasmic group 3c (methyl viologen-reducing) [NiFe] hydrogenases (seen in one other genome in this clade), probably involved in H_2 utilization (**Figure S9**) [90, 91]. Additionally, most genomes in lineage V encode membrane-bound group 4 [NiFe] hydrogenases of the Mbh-Mrp type, likely involved in H_2 production (**Figure S10**). Membrane-bound hydrogenases are known to oxidize reduced ferredoxin, and the presence of antiporter-like subunits suggests that they may be involved in ion translocation across the membrane and the generation of a membrane potential [92]. These genomes also encode hydrogenase-related complexes (Ehr), the role of which is still unknown, although it has been suggested that they may interact with oxidoreductases and quinones and may be involved in ion translocation across the membrane [61, 93]. Taken together, this indicates lineage V *Elusimicrobia* sense H_2 and NADH levels and regulate their metabolisms in response. The repertoire of functions indicated by their suite of hydrogenases suggests they are important players in the hydrogen economy of their ecosystem.

Groundwater and peat-associated lineages IIa, IIc, IV, and VI have somewhat distinct respiratory capacities compared to the groundwater-associated lineage V *Elusimicrobia*. All have a complex I lacking the diaphorase N-module (*nuoEFG* genes), a complex which is hypothesized to use reduced ferredoxin instead of NADH [94]. Lineage IV genomes also have a canonical complex I with an N-module. All genomes from these lineages have complex II (succinate dehydrogenase) and complex IV (cytochrome c oxidase, type A; **Figure S11**). Lineage IV is further different in that it encodes an alternative complex III and a few lineage IV genomes contain a canonical complex III (cytochrome c reductase). Unlike lineage V, which has a consistent set of hydrogenases, some of these genomes have hydrogenases and others do not (**Suppl. Dataset 3**).

The genomes of some members of the groundwater and peat-associated lineages IIa, IIc, IV, V and VI encode nitrogen cycling capacities that are rare in other lineages. Specifically, seven genomes encode a nitrite oxidoreductase (NxrA, **Figure S12**) and/or nitrate reductase (**Figure S12**), indicating that these organisms can respire using nitrate. We also identified other enzymes involved in the nitrogen cycle, such as quinol-dependent nitric-oxide reductase (qNOR, **Figure S11**), cytochrome-c nitrite reductase (NfrAH), and copper and/or cytochrome cd1 nitrite reductase(s) (NirK and NirS) (**Suppl. Dataset 3**). These findings expand the known lifestyles in the phylum *Elusimicrobia* from exclusively fermentative to include several respiratory strategies, using oxygen or nitrogen compounds as terminal electron acceptors for energy conservation.

RNF-dependent acetogenesis in groundwater-associated *Elusimicrobia*

Other groundwater-associated *Elusimicrobia* in lineages VII, VIII and IX lack the capacity to reduce oxygen or nitrate. Instead, lineages VII, VIII and IX encode the Wood-Ljungdahl pathway (WLP) for the reduction of carbon dioxide to acetyl-coenzyme A with concomitant energy conservation (**Figure 4B and Suppl. Dataset 3**). The WLP is often coupled with cytochromes and quinones to generate a membrane potential for ATP synthesis via an ATP synthase [95]. However, cytochromes and quinones are missing in these lineages. Instead, these genomes encode the Rnf complex, a sodium-motive ferredoxin:NAD oxidoreductase. The Rnf complex generates a sodium ion potential across the cell membrane in *Acetobacterium woodii* [96]. These observations suggest that lineages VII, VIII and IX are capable of autotrophic growth with molecular hydrogen and carbon dioxide as substrates. Members of lineages VII, VIII, IX, ERL2, and a few genomes in lineage I have membrane-bound group 4 [NiFe] hydrogenases of the Mbh-Mrp-oxidoreductase type (**Figure S10 and Suppl. Dataset 3**). The function of this oxidoreductase is unknown, though it has been seen before in other organisms such as the *Thermococcales* [92] and *Saganbacteria* [61]. Both Rnf complexes and group 4 NiFe membrane-bound hydrogenases could extrude ions across the membrane and contribute to the creation of a membrane potential. These *Elusimicrobia* lineages are predicted acetogens, another lifestyle not previously associated with the phylum.

Gut-associated *Elusimicrobia*

The genomes and general metabolic characteristics of gut-related *Elusimicrobia* (lineages I and III) have been well described previously [12–15]. However, here we extend prior work by detailing the hydrogenases of these bacteria (**Figure 3 and Suppl. Dataset 3**). In brief, prior work indicates that gut-associated *Elusimicrobia* rely on fermentation [4, 12, 13, 16, 17]. Most gut-related *Elusimicrobia* have 'ancestral' group B and group A [FeFe] hydrogenases. The 'ancestral' group B is typically found in anaerobic bacteria (*e.g.*, *Clostridia*, *Actinobacteria*, and *Bacteroidetes*) that inhabit gastrointestinal tracts and anoxic soils and sediments [62, 88] (**Figure S8**). Group B [FeFe] hydrogenases remain largely uncharacterized. It has been suggested that they could couple ferredoxin oxidation to H_2 evolution, because fermentative organisms rely on ferredoxin as a key redox equivalent [62]. In general, group B enzymes are monomeric. Many of the gut-associated *Elusimicrobia* genomes also harbor a trimeric group A [FeFe] hydrogenase that may be involved in fermentative metabolism. Group A [FeFe] hydrogenases contain NADH-binding domains in the beta subunit and are known to be involved in electron bifurcation or the reverse reaction, electron confurcation (**Figure S8**). In the case of electron bifurcation, [FeFe] hydrogenases oxidize H_2 by coupling the exergonic reduction of NAD^+ with the endergonic reduction of oxidized ferredoxin. In electron confurcation, these [FeFe] hydrogenases produce H_2 by using the exergonic oxidation of ferredoxin with protons (H^+) to drive the endergonic oxidation of NADH with H^+ [97–99]. We conclude that many gut-associated *Elusimicrobia* can potentially both produce and use H_2, but using different hydrogenase types than those of other *Elusimicrobia* lineages.

Distinct from these, two gut-associated *Elusimicrobia* have additional mechanisms to produce H_2. One genome (genome 277) harbors a monomeric (as opposed to trimeric, as described above) group A [FeFe] hydrogenase in addition to a group B hydrogenase (**Figure S8**). Monomeric group A hydrogenases are thought to be involved in H_2 production from ferredoxin and flavodoxin [88, 100]. Another gut-associated genome (*Endomicrobium proavitum*) has a tetrameric group A [FeFe] hydrogenase, which is also known to be involved in electron bifurcation and likely H_2 production [88].

The *Elusimicrobium minutum* Pei191 genome is unusual because it harbors four different kinds of [FeFe] hydrogenases (**Figure S8**), three of which have been described before, a monomeric group B [FeFe] hydrogenase (HydA2), downstream from a dimeric 'sensory' group C (HydS) [FeFe] hydrogenase [89] and a trimeric group A [FeFe] hydrogenase [15]. The fourth [FeFe] hydrogenase is another monomeric group B hydrogenase like others found in gut-associated *Elusimicrobia*, with a slightly different domain composition of the catalytic subunit than the other group B hydrogenase (HydA2) in this genome. A membrane-bound group 4 [NiFe] hydrogenase of the Mbh-Mrp type also occurs in this genome [15] (**Figure S10**). This is striking, as the rest of the gut-associated *Elusimicrobia* completely lack [NiFe] hydrogenases (**Suppl. Dataset 3**).

Conclusions

Previously, our understanding of the *Elusimicrobia* phylum mostly relied on gut-associated habitats although 16S rRNA gene sequences indicated their wider environmental distribution and some genomes from non-gut habitats were available in genome databases. We expanded the genomic representation of this phylum with thirty new genome sequences, mainly from groundwater. However, six of our draft genomes came from humans and pigs, thus expand on the finding of *Elusimicrobia* of lineage III in the gut microbiome of non-Westernized humans [68]. Our results reshape our view of the *Elusimicrobia* phylum, as most of these genomes likely derive from free-living organisms that are predicted to have greater metabolic flexibility than the strictly fermentative prior representatives of the phylum. Our phylogenetic results reveal five new lineages including two potential phylum-level lineages and suggest that animal gut-associated *Elusimicrobia* of lineages I and III evolved from two distinct groups of free-living ancestors, lineages IV, V and group VII, respectively. Based on extant members of groundwater groups, the ancestors of gut-associated *Elusimicrobia* likely underwent substantial genome reduction following the transition from nutrient-poor (groundwater or soil) to nutrient-rich (gut) environments. Genome reduction following formation of a long-term association with another organism has been well documented for many endosymbionts (for example of insects; [101] or in lineage I; [12]). However, for *Elusimicrobia*, genome reduction of free-living bacteria may have involved genome streamlining rather than genetic drift enabled by small population sizes in the intracellular environment.

Perhaps the most remarkable finding is the presence of a novel group of nitrogenase paralogs that are phylogenetically distinct from the five known groups. The new paralogs branch next to the group IV and V and co-occur with an extensive suite of radical SAM-based proteins in groundwater-associated *Elusimicrobia* from lineages IV and V. Based on gene context and protein phylogeny, it is unlikely the new paralog performs nitrogen fixation. Instead, we predict that the nifH proteins reduce a tetrapyrrole. Given the functions of related clades of genes, the nifH-like protein may synthesize a novel cofactor, but further investigations are needed to determine the product and its function. We anticipate exciting and perhaps unprecedented chemistry associated with this enzyme, given the versatility of nitrogenases and nitrogenase homologs in substrate reduction [75].

Acknowledgements

Support was provided by grants from the Lawrence Berkeley National Laboratory's Genomes-to-Watershed Scientific Focus Area. The U.S. Department of Energy (DOE), Office of Science, and Office of Biological and Environmental Research funded the work under contract DE-AC02-05CH11231 and the DOE carbon cycling program DOE-SC10010566, the Innovative Genomics Institute at Berkeley and the Chan Zuckerberg Biohub. The Ministry of Economy, Trade and Industry of Japan funded a part of the work as "The project for validating assessment methodology in geological disposal system". Teruki Iwatsuki, Kazuki Hayashida, Toshihiro Kato, Mitsuru Kubota, Kazuya Miyakawa, and Akihito Mochizuki assisted with groundwater sampling at Mizunami and Horonobe Underground Research Laboratories, Japan Atomic Energy Agency (JAEA). C.H. acknowledges the Camille and Henry Dreyfus Foundation Postdoctoral Program in Environmental Chemistry for a Fellowship. L.A.H is supported by a Tier II Canada Research Chair.

Conflict of interests

J.F.B. is a founder of Metagenomi. The other authors declare no competing interests.

Author contributions

J.F.B., L.A.H and R.M. conceived the study. C.J.C., P.M.C., L.A.H. and R.M. analysed the genomic data. R.M. and L.A.H. performed the phylogenetic analyses. I.F. performed the Cazy analysis. L-X.C., Y.A., C.H., I.F. and J.F.B. provided the new genomes. R.M., J.F.B, C.J.C and P.M.C. wrote the manuscript with input from all authors. All documents were edited and approved by all authors.

Data and Materials availability

The genomes of the herein analysed *Elusimicrobia* have been made publicly available on NCBI and on the ggkbase database (https://ggkbase.berkeley.edu/non_redundant_elusimicrobia_2018/organisms). Detailed annotations of the metabolic repertoire are provided in **Suppl. Dataset 3** accompanying this paper. Raw data files (phylogenetic trees, fasta sequences and the list of KEGG annotations) are made available via figshare under the following link: https://figshare.com/articles/Elusimicrobia_analysis/8939678. Correspondence and material requests should be addressed to jbanfield@berkeley.edu.

References

1. Ohkuma M, Kudo T. Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. *Appl Environ Microbiol.* 1996; **62**: 461–468.
2. Dojka MA, Hugenholtz P, Haack SK, Pace NR. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl Environ Microbiol.* 1998; **64**: 3869–3877.
3. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol.* 1998; **180**: 4765–4774.
4. Geissinger O, Herlemann DPR, Mörschel E, Maier UG, Brune A. The ultramicrobacterium '*Elusimicrobium minutum*' gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. *Appl Environ Microbiol.* 2009; **75**: 2831–2840.
5. Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, *et al.* Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. *Environ Microbiol.* 2017; **19**: 459–474.
6. Spring S, Bunk B, Spröer C, Schumann P, Rohde M, Tindall BJ, *et al*. Characterization of the first cultured representative of *Verrucomicrobia* subdivision 5 indicates the proposal of a novel phylum. *ISME J.* 2016; **10**: 2801–2816.
7. Cho J-C, Vergin KL, Morris RM, Giovannoni SJ. *Lentisphaera araneosa* gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, *Lentisphaerae*. *Environ Microbiol.* 2004; **6**: 611–621.
8. Stingl U, Radek R, Yang H, Brune A. '*Endomicrobia*': Cytoplasmic Symbionts of Termite Gut Protozoa Form a Separate Phylum of Prokaryotes. *Appl Environ Microbiol.* 2005; **71**: 1473–1479.
9. Ohkuma M, Sato T, Noda S, Ui S, Kudo T, Hongoh Y. The candidate phylum Termite Group 1 of bacteria: phylogenetic diversity, distribution, and endosymbiont members of various gut flagellated protists. *FEMS Microbiology Ecology*. 2007; **60**: 467–476

10. Izawa K, Kuwahara H, Sugaya K, Lo N, Ohkuma M, Hongoh Y. Discovery of ectosymbiotic *Endomicrobium* lineages associated with protists in the gut of stolotermitid termites. *Environ Microbiol Rep*. 2017; **9**: 411–418.

11. Mikaelyan A, Thompson CL, Meuser K, Zheng H, Rani P, Plarre R, *et al*. High-resolution phylogenetic analysis of *Endomicrobia* reveals multiple acquisitions of endosymbiotic lineages by termite gut flagellates. *Environmental Microbiology Reports*. 2017; **9**: 477–483.

12. Zheng H, Dietrich C, Brune A. Genome Analysis of *Endomicrobium proavitum* Suggests Loss and Gain of Relevant Functions during the Evolution of Intracellular Symbionts. *Appl Environ Microbiol.* 2017; **83**: e00656-17.

13. Hongoh Y, Sharma VK, Prakash T, Noda S, Taylor TD, Kudo T, *et al*. Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc Natl Acad Sci USA.* 2008; **105**: 5555–5560.

14. Izawa K, Kuwahara H, Kihara K, Yuki M, Lo N, Itoh T, *et al*. Comparison of Intracellular '*Ca*. Endomicrobium Trichonymphae' Genomovars Illuminates the Requirement and Decay of Defense Systems against Foreign DNA. *Genome Biol Evol.* 2016; **8**: 3099–3107.

15. Herlemann DPR, Geissinger O, Ikeda-Ohtsubo W, Kunin V, Sun H, Lapidus A, *et al*. Genomic Analysis of '*Elusimicrobium minutum*', the First Cultivated Representative of the Phylum '*Elusimicrobia*' (Formerly Termite Group 1). *Applied and Environmental Microbiology*. 2009; **75**: 2841–2849.

16. Herlemann DPR, Geissinger O, Brune A. The termite group I phylum is highly diverse and widespread in the environment. *Appl Environ Microbiol.* 2007; **73**: 6682–6685.

17. Zheng H, Dietrich C, Radek R, Brune A. *Endomicrobium proavitum*, the first isolate of *Endomicrobia* class. nov. (phylum *Elusimicrobia*) - an ultramicrobacterium with an unusual cell cycle that fixes nitrogen with a Group IV nitrogenase. *Environmental Microbiology*. 2016; **18**: 191–204.

18. Brune A. *Elusimicrobia*. In: W.B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun, P. DeVos,

B. Hedlund and S. Dedysh (eds). Bergey's Manual of Systematics of Archaea and Bacteria (2020).

19. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, *et al*. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun.* 2016; **7**: 13219.
20. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, *et al*. Genome-centric view of carbon processing in thawing permafrost. *Nature.* 2018; **560**: 49–54.
21. Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, *et al*. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife.* 2015; **4**: e05477.
22. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics.* 2016; **32**: 605–607.
23. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, *et al*. Binning metagenomic contigs by coverage and composition. *Nat Methods.* 2014; **11**: 1144–1146.
24. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, *et al*. Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. *Nat Microbiol*. 2018; **3**: 836–843.
25. Song Z-Q, Wang F-P, Zhi X-Y, Chen J-Q, Zhou E-M, Liang F, *et al.* Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. *Environmental Microbiology*. 2013; **15**: 1160–1175.
26. Chen L-X, Al-Shayeb B, Méheust R, Li W-J, Doudna JA, Banfield JF. Candidate Phyla Radiation Roizmanbacteria From Hot Springs Have Novel and Unexpectedly Abundant CRISPR-Cas Systems. *Front Microbiol.* 2019; **10**: 928.
27. Youssef NH, Farag IF, Ryan Hahn C, Premathilake H, Fry E, Hart M, *et al*. *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. *Systematic and Applied Microbiology*. 2019; **42**: 85–93.

28. Ino K, Hernsdorf AW, Konno U, Kouduka M, Yanagawa K, Kato S, *et al*. Ecological and genomic profiling of anaerobic methane-oxidizing archaea in a deep granitic environment. *ISME J*. 2018; **12**: 31–47.
29. Hernsdorf AW, Amano Y, Miyakawa K, Ise K, Suzuki Y, Anantharaman K, *et al*. Potential for microbial H_2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. *ISME J*. 2017; **11**: 1915–1929.
30. Probst AJ, Weinmaier T, Raymann K, Perras A, Emerson JB, Rattei T, *et al*. Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat Commun*. 2014; **5**: 5497.
31. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, *et al*. Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat Microbiol*. 2019; **4**: 693–700.
32. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017; **11**: 2864–2868.
33. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013; **41**: 590–6.
34. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014; **30**: 1312–1313.
35. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop (GCE)*. 2010; pp 1–8.
36. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; **32**: 1792–1797.
37. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*. 2019; **36**: 1925–1927.

38. Méheust R, Burstein D, Castelle CJ, Banfield JF. The distinction of CPR bacteria from other bacteria based on protein family content. *Nat Commun.* 2019; **10**: 4173.

39. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* 2017; **35**: 1026–1028.

40. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005; **21**: 951–960.

41. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods.* 2011; **9**: 173–175.

42. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002; **30**: 1575–1584.

43. Katoh K, Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics.* 2016; **32**: 1933–1942.

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

45. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015; **32**: 268–274.

46. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 2017; **14**: 587–589.

47. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol.* 2018; **35**: 518–522.

48. Eddy SR. Profile hidden Markov models. *Bioinformatics.* 1998; **14**: 755–763.

49. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016; **44**: 457–62.

50. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, *et al*. The Pfam protein families database. *Nucleic Acids Res.* 2012; **40**: 290–301.

51. Bernardes JS, Vieira FRJ, Zaverucha G, Carbone A. A multi-objective optimization approach

accurately resolves protein domain architectures. *Bioinformatics*. 2016; **32**: 345–353.

52. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011; **8**: 785–786.
53. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001; **305**: 567–580.
54. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; **25**: 3389–3402.
55. Saier MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res*. 2016; **44**: 372–D379.
56. Wilson WA, Roach PJ, Montero M, Baroja-Fernández E, Muñoz FJ, Eydallin G, *et al*. Regulation of glycogen metabolism in yeast and bacteria. *FEMS Microbiology Reviews*. 2010; **34**: 952–985.
57. Adam PS, Borrel G, Gribaldo S. Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc Natl Acad Sci USA*. 2018; **115**: 1166–1173.
58. Membrillo-Hernandez J, Echave P, Cabiscol E, Tamarit J, Ros J, Lin EC. Evolution of the *adhE* gene product of *Escherichia coli* from a functional reductase to a dehydrogenase. Genetic and biochemical studies of the mutant proteins. *J Biol Chem*. 2000; **275**: 33869–33875.
59. Madern D. Molecular evolution within the L-malate and L-lactate dehydrogenase super-family. *J Mol Evol*. 2002; **54**: 825–840.
60. Dailey HA, Dailey TA, Gerdes S, Jahn D, Jahn M, O'Brian MR, *et al*. Prokaryotic Heme Biosynthesis: Multiple Pathways to a Common Essential Product. *Microbiol Mol Biol Rev*. 2017; **81**: e00048-16.
61. Matheus Carnevali PB, Schulz F, Castelle CJ, Kantor RS, Shih PM, Sharon I, *et al*. Hydrogen-based metabolism as an ancestral trait in lineages sibling to the *Cyanobacteria*. *Nat Commun*. 2019; **10**: 463.

62. Greening C, Biswas A, Carere CR, Jackson CJ, Taylor MC, Stott MB, *et al*. Genomic and metagenomic surveys of hydrogenase distribution indicate H_2 is a widely utilised energy source for microbial growth and survival. *ISME J.* 2016; **10**: 761–777.

63. Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D, *et al*. Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Commun.* 2013; **4**: 2120.

64. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol.* 2014; **12**: 635–645.

65. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol.* 2018; **36**: 996–1004.

66. Shi W, Moon CD, Leahy SC, Kang D, Froula J, Kittelmann S, et al. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res.* 2014; **24**: 1517–1525.

67. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, *et al*. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr Biol.* 2015; **25**: 1682–1693.

68. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell.* 2019; **176**: 649–662.e20.

69. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol.* 2017; **15**: 579–590.

70. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci USA.* 2014; **111**: 4904–4909.

71. Raymond J, Siefert JL, Staples CR, Blankenship RE. The natural history of nitrogen fixation. *Mol Biol Evol.* 2004; **21**: 541–554.

72. Moore SJ, Sowa ST, Schuchardt C, Deery E, Lawrence AD, Ramos JV, *et al*. Elucidation of the

biosynthesis of the methane catalyst coenzyme F430. *Nature.* 2017; **543**: 78–82.

73. Zheng K, Ngo PD, Owens VL, Yang X-P, Mansoorabadi SO. The biosynthetic pathway of coenzyme F430 in methanogenic and methanotrophic archaea. *Science.* 2016; **354**: 339–342.
74. Burke DH, Hearst JE, Sidow A. Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. *Proc Natl Acad Sci USA.* 1993; **90**: 7134–7138.
75. Hu Y, Ribbe MW. Nitrogenase and homologs. *J Biol Inorg Chem.* 2015; **20**: 435–445.
76. Rosnow JJ, Hwang S, Killinger BJ, Kim Y-M, Moore RJ, Lindemann SR, *et al*. A Cobalamin Activity-Based Probe Enables Microbial Cell Growth and Finds New Cobalamin-Protein Interactions across Domains. *Appl Environ Microbiol.* 2018; **84**: e00955-18.
77. Naoe Y, Nakamura N, Doi A, Sawabe M, Nakamura H, Shiro Y, *et al.* Crystal structure of bacterial haem importer complex in the inward-facing conformation. *Nat Commun.* 2016; **7**: 13411.
78. Sofia HJ, Chen G, Hetzler BG, Reyes-Spindola JF, Miller NE. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. *Nucleic Acids Res.* 2001; **29**: 1097–1106.
79. Mehta AP, Abdelwahed SH, Mahanta N, Fedoseyenko D, Philmus B, Cooper LE, *et al*. Radical S-adenosylmethionine (SAM) enzymes in cofactor biosynthesis: a treasure trove of complex organic radical rearrangement reactions. *J Biol Chem.* 2015; **290**: 3980–3986.
80. Layer G, Moser J, Heinz DW, Jahn D, Schubert W-D. Crystal structure of coproporphyrinogen III oxidase reveals cofactor geometry of Radical SAM enzymes. *EMBO J.* 2003; **22**: 6214–6224.
81. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, *et al*. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature.* 2015; **523**: 208–211.
82. Peters JW, Miller A-F, Jones AK, King PW, Adams MW. Electron bifurcation. *Curr Opin Chem Biol.* 2016; **31**: 146–152.
83. Borisov VB, Gennis RB, Hemp J, Verkhovsky MI. The cytochrome bd respiratory oxygen reductases. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* . 2011; **1807**: 1398–1413.

84. Kracke F, Vassilev I, Krömer JO. Microbial electron transport and energy conservation - the foundation for optimizing bioelectrochemical systems. *Front Microbiol.* 2015; **6**: 575.

85. Brooijmans RJW, de Vos WM, Hugenholtz J. *Lactobacillus plantarum* WCFS1 electron transport chains. *Appl Environ Microbiol.* 2009; **75**: 3580–3585.

86. Sootsuwan K, Lertwattanasakul N, Thanonkeo P, Matsushita K, Yamada M. Analysis of the respiratory chain in Ethanologenic *Zymomonas mobilis* with a cyanide-resistant *bd*-type ubiquinol oxidase as the only terminal oxidase and its possible physiological roles. *J Mol Microbiol Biotechnol.* 2008; **14**: 163–175.

87. Shaw AJ, Hogsett DA, Lynd LR. Identification of the [FeFe]-hydrogenase responsible for hydrogen generation in *Thermoanaerobacterium saccharolyticum* and demonstration of increased ethanol yield via hydrogenase knockout. *J Bacteriol.* 2009; **191**: 6457–6464.

88. Calusinska M, Happe T, Joris B, Wilmotte A. The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. *Microbiology.* 2010; **156**: 1575–1588.

89. Zheng Y, Kahnt J, Kwon IH, Mackie RI, Thauer RK. Hydrogen Formation and Its Regulation in *Ruminococcus albus*: Involvement of an Electron-Bifurcating [FeFe]-Hydrogenase, of a Non-Electron-Bifurcating [FeFe]-Hydrogenase, and of a Putative Hydrogen-Sensing [FeFe]-Hydrogenase. *Journal of Bacteriology*. 2014; **196**: 3840–3852.

90. Orsi WD, Vuillemin A, Rodriguez P, Coskun ÖK, Gomez-Saez GV, Lavik G, *et al*. Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nat Microbiol.* 2019. **5**: 248–255.

91. Vignais PM, Billoud B. Occurrence, Classification, and Biological Function of Hydrogenases: An Overview. *Chemical Reviews*. 2007; **107**: 4206–4272.

92. Schut GJ, Boyd ES, Peters JW, Adams MWW. The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol Rev.* 2013; **37**: 182–203.

93. Marreiros BC, Batista AP, Duarte AMS, Pereira MM. A missing link between complex I and group

4 membrane-bound [NiFe] hydrogenases. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. 2013; **1827**: 198–209.

94. Battchikova N, Eisenhut M, Aro E-M. Cyanobacterial NDH-1 complexes: novel insights and remaining puzzles. *Biochim Biophys Acta*. 2011; **1807**: 935–944.

95. Schuchmann K, Müller V. Energetics and Application of Heterotrophy in Acetogenic Bacteria. *Appl Environ Microbiol*. 2016; **82**: 4056–4069.

96. Poehlein A, Schmidt S, Kaster A-K, Goenrich M, Vollmers J, Thürmer A, *et al*. An ancient pathway combining carbon dioxide fixation with the generation and utilization of a sodium ion gradient for ATP synthesis. *PLoS One*. 2012; **7**: e33439.

97. Schuchmann K, Chowdhury NP, Müller V. Complex Multimeric [FeFe] Hydrogenases: Biochemistry, Physiology and New Opportunities for the Hydrogen Economy. *Frontiers in Microbiology*. 2018; **9**

98. Müller V, Chowdhury NP, Basen M. Electron Bifurcation: A Long-Hidden Energy-Coupling Mechanism. *Annu Rev Microbiol* 2018; **72**: 331–353.

99. Buckel W, Thauer RK. Flavin-Based Electron Bifurcation, Ferredoxin, Flavodoxin, and Anaerobic Respiration With Protons (Ech) or NAD (Rnf) as Electron Acceptors: A Historical Review. *Frontiers in Microbiology*. 2018; **9**: 2911.

100. Demuez M, Cournac L, Guerrini O, Soucaille P, Girbal L. Complete activity profile of *Clostridium acetobutylicum* [FeFe]-hydrogenase and kinetic parameters for endogenous redox partners. *FEMS Microbiology Letters*. 2007; **275**: 113–121.

101. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. 2011; **10**: 13–26.

Figure legends

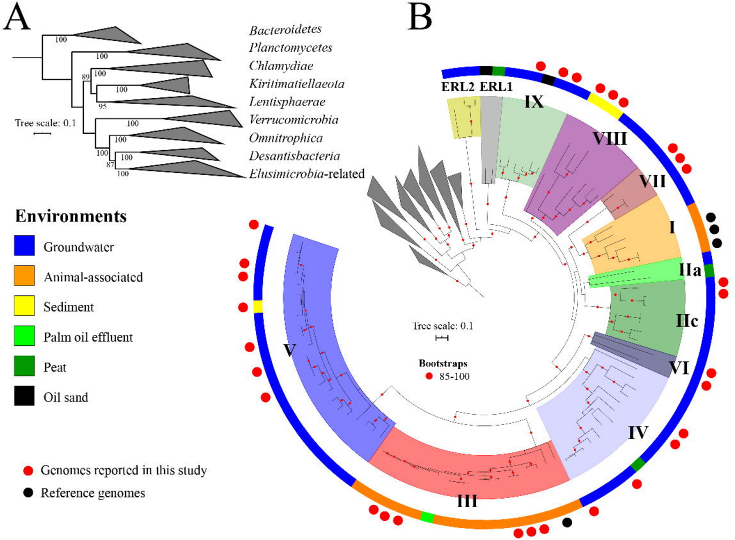
Figure 1. Phylogenetic placement of newly reconstructed genomes. A. Relationships between the phyla of the *PVC* superphylum. **B.** Placement of the 94 genomes related to the *Elusimicrobia* phylum. The maximum likelihood tree was constructed based on a concatenated alignment of 16 ribosomal proteins under an LG+gamma model of evolution. The outside color bar indicates the environment of origin. Black circles represent the four genomes described in previous studies while red circles represent the 30 new genomes added in this study. The remaining genomes come from the NCBI genome database. Scale bars indicate the mean number of substitutions per site. The lineages of genomes are indicated by colored ranges and roman numbers. The bootstraps are indicated by red circles when ≥ 85. The complete ribosomal protein tree is available in rectangular format in Figure S2.

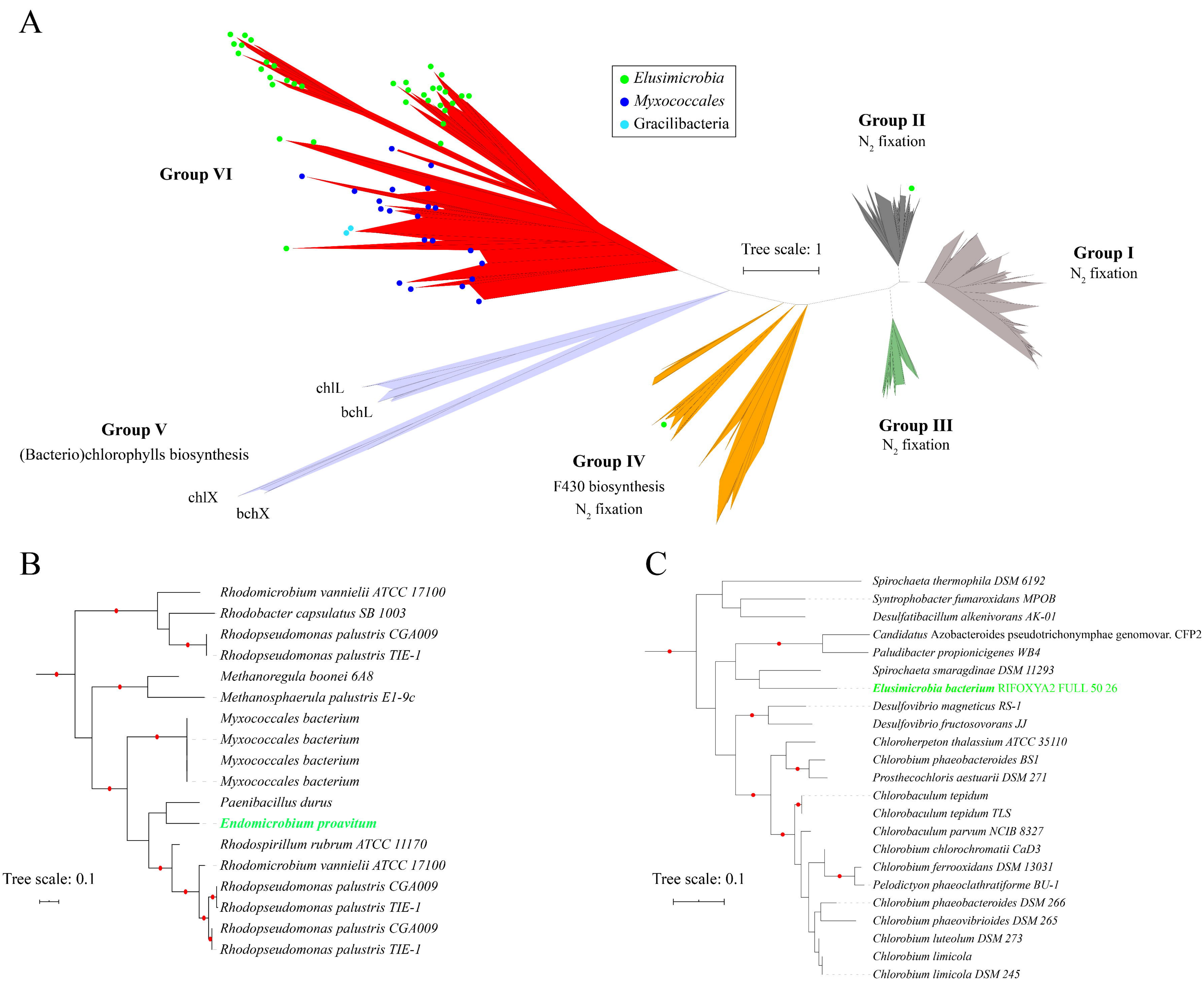
Figure 2. Maximum likelihood phylogeny of the nitrogenase subunit NifH. A. Full phylogeny of the NifH subunit. **B**. Detailed phylogeny of the group IV. **C**. Detailed phylogeny of the group II. The tree was inferred using an LG+I+G4 model of evolution. The tree was made using reference sequences from [71]. The green circles highlight the sequences present in *Elusimicrobia* genomes. Scale bar indicates the mean number of substitutions per site. The sequences and tree are available with full bootstrap values in Fasta and Newick format in the Supplementary Data. Branches with bootstrap values ≥ 95 are indicated by red circles.

Figure 3. Schematic representation of the relationship of the *Elusimicrobia*-related lineages and the distribution of selected metabolic features across the different members of the phylum. Grey scale of the filled circles represents the prevalence of a function within a given lineage, with black = 100% of genomes and white = 0% of genomes. See **Suppl. Dataset 3** for a fully detailed version of each genome.

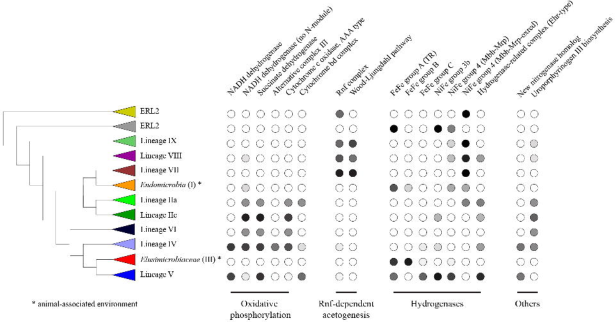
Figure 4. Overview of the metabolic capabilities of non-gut-associated *Elusimicrobia*. A. Lineage IV and V, (B). Lineage IX. Abbreviations not defined in the text: GHS, glycoside hydrolases; TCA, tricarboxylic acid cycle; 2-oxo, 2-oxoglutarate; Suc-CoA, succinyl-CoA; DH, dehydrogenase; HYD,

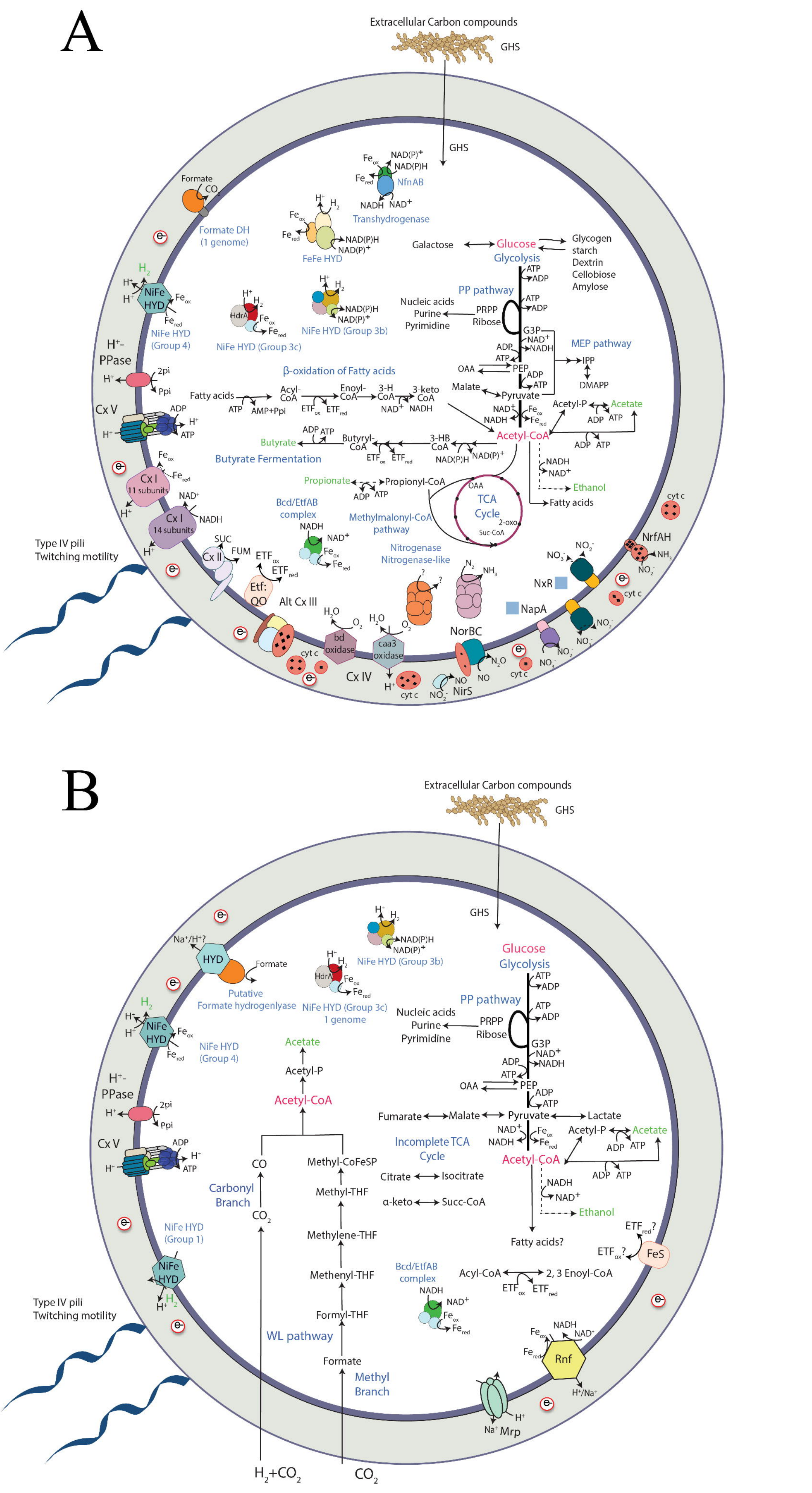
hydrogenase; PPi, pyrophosphate; Pi, inorganic phosphate; PRPP, Phosphoribosyl pyrophosphate; G3P, glyceraldehyde-3-phosphate; PEP, phosphoenolpyruvate; OAA, oxaloacetate; MEP pathway, 2-*C*-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate pathway; IPP, isopentenyl pyrophosphate; DMAPP, dimethylallylpyrophosphate; FeS, FeS oxidoreductase; ETF, electron transfer flavoprotein; ETF:QO, Electron-transferring flavoprotein dehydrogenase/ubiquinone oxidoreductase; BCD/ETFAB, butyryl-CoA dehydrogenase/electron-transfer flavoprotein complex; 3HB-CoA, 3-hydroxybutyryl-CoA dehydrogenase; 3-H-CoA, 3-hydroxyacyl-CoA; 3-keto-CoA, 3-ketoacyl-CoA; H+ -PPase, proton-translocating pyrophosphatase; Cx V, ATP synthase; Cx I, NADH dehydrogenase; Cx II, succinate dehydrogenase/fumarate reductase; Alt Cx III, alternative complex III; Cx IV, cytochrome c oxidase or cytochrome bd oxidase; NorBC, nitric-oxide reductase; NirS, cytochrome cd1-type nitrite reductase; cyt c, c-type cytochrome; NapA, periplasmic nitrate reductase; NxR, nitrite/nitrate oxidoreductase; NrfAH, nitrite reductase; Mrp, multi-subunit Na^+/H^+ antiporter; WL pathway, Wood–Ljungdahl pathway; THF, tetrahydrofolate; Rnf, Ferredoxin:NAD+-Oxidoreductase.


A
Bacteroidetes
Planctomycetes
Chlamydiae
Kiritimatiellaeota
Lentisphaerae
Verrucomicrobia
Omnitrophica
Desantisbacteria
Elusimicrobia-related
Tree scale: 0.1
100
100
100
89
100
95
100
100
100
87
100
100
B
ERL2
ERL1
IX
VIII
VII
I
IIa
IIc
VI
IV
III
V
Tree scale: 0.1
Bootstraps
85-100
Environments
Groundwater
Animal-associated
Sediment
Palm oil effluent
Peat
Oil sand
Genomes reported in this study
Reference genomes



A
Elusimicrobia
Myxococcales
Gracilibacteria
Group VI
Group II
N_2 fixation
Group I
N_2 fixation
Tree scale: 1
Group III
N_2 fixation
Group IV
F430 biosynthesis
N_2 fixation
Group V
(Bacterio)chlorophylls biosynthesis
chlL
bchL
chlX
bchX
B
Rhodomicrobium vannielii ATCC 17100
Rhodobacter capsulatus SB 1003
Rhodopseudomonas palustris CGA009
Rhodopseudomonas palustris TIE-1
Methanoregula boonei 6A8
Methanosphaerula palustris E1-9c
Myxococcales bacterium
Myxococcales bacterium
Myxococcales bacterium
Myxococcales bacterium
Paenibacillus durus
Endomicrobium proavitum
Rhodospirillum rubrum ATCC 11170
Rhodomicrobium vannielii ATCC 17100
Rhodopseudomonas palustris CGA009
Rhodopseudomonas palustris TIE-1
Rhodopseudomonas palustris CGA009
Rhodopseudomonas palustris TIE-1
Tree scale: 0.1
C
Spirochaeta thermophila DSM 6192
Syntrophobacter fumaroxidans MPOB
Desulfatibacillum alkenivorans AK-01
Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2
Paludibacter propionicigenes WB4
Spirochaeta smaragdinae DSM 11293
Elusimicrobia bacterium RIFOXYA2 FULL 50 26
Desulfovibrio magneticus RS-1
Desulfovibrio fructosovorans JJ
Chloroherpeton thalassium ATCC 35110
Chlorobium phaeobacteroides BS1
Prosthecochloris aestuarii DSM 271
Chlorobaculum tepidum
Chlorobaculum tepidum TLS
Chlorobaculum parvum NCIB 8327
Chlorobium chlorochromatii CaD3
Chlorobium ferrooxidans DSM 13031
Pelodictyon phaeoclathratiforme BU-1
Chlorobium phaeobacteroides DSM 266
Chlorobium phaeovibrioides DSM 265
Chlorobium luteolum DSM 273
Chlorobium limicola
Chlorobium limicola DSM 245
Tree scale: 0.1


NADH dehydrogenase
NADH dehydrogenase (no N-module)
Succinate dehydrogenase
Alternative complex III
Cytochrome c oxidase, AAA type
Cytochrome bd complex
Rnf complex
Wood-Ljungdahl pathway
FeFe group A (TR)
FeFe group B
FeFe group C
NiFe group 3b
NiFe group 4 (Mbh-Mrp)
NiFe group 4 (Mbh-Mrp-oxred)
Hydrogenase-related complex (Ehr-type)
New nitrogenase homolog
Uroporphyrinogen III biosynthesis
ERL2
ERL2
Lineage IX
Lineage VIII
Lineage VII
Endomicrobia (I) *
Lineage IIa
Lineage IIc
Lineage VI
Lineage IV
Elusimicrobiaceae (III) *
Lineage V
* animal-associated environment
Oxidative phosphorylation
Rnf-dependent acetogenesis
Hydrogenases
Others


A
Extracellular Carbon compounds
GHS
GHS
Formate
CO_2
Formate DH
(1 genome)
Fe_{ox}
NAD(P)+
NAD(P)H
Fe_{red}
NfnAB
NADH
NAD+
Transhydrogenase
H^+
H_2
Fe_{ox}
Fe_{red}
NAD(P)H
NAD(P)+
FeFe HYD
Galactose
Glucose
Glycolysis
Glycogen
starch
Dextrin
Cellobiose
Amylose
H_2
NiFe HYD
Fe_{ox}
Fe_{red}
NiFe HYD
(Group 4)
HdrA
NiFe HYD (Group 3c)
NiFe HYD (Group 3b)
PP pathway
Nucleic acids
Purine
Pyrimidine
PRPP
Ribose
ATP
ADP
G3P
NAD+
NADH
MEP pathway
IPP
DMAPP
OAA
PEP
Malate
Pyruvate
H+-
PPase
2pi
Ppi
β-oxidation of Fatty acids
Fatty acids
Acyl-CoA
Enoyl-CoA
3-H CoA
3-keto CoA
AMP+Ppi
ETF_{ox}
ETF_{red}
Cx V
Acetyl-P
Acetate
Acetyl-CoA
Butyrate
Butyryl-CoA
3-HB CoA
NAD(P)H
NAD(P)+
Butyrate Fermentation
Cx I
11 subunits
Cx I
14 subunits
NADH
Ethanol
Fatty acids
Propionate
Propionyl-CoA
TCA
Cycle
2-oxo
Suc-CoA
Bcd/EtfAB
complex
Methylmalonyl-CoA
pathway
cyt c
NrfAH
NH_3
NO_2^-
NO_3^-
Type IV pili
Twitching motility
SUC
FUM
Cx II
Etf:
QO
Alt Cx III
Nitrogenase
Nitrogenase-like
N_2
NxR
NapA
H_2O
O_2
bd
oxidase
caa3
oxidase
Cx IV
NorBC
N_2O
NO
NirS
B
Extracellular Carbon compounds
GHS
Na+/H+?
HYD
Formate
Putative
Formate hydrogenlyase
NiFe HYD (Group 3c)
1 genome
NiFe HYD (Group 3b)
Glucose
Glycolysis
PP pathway
Acetate
Acetyl-P
Acetyl-CoA
NiFe HYD
(Group 4)
Fumarate
Malate
Pyruvate
Lactate
Incomplete TCA
Cycle
CO
Methyl-CoFeSP
Carbonyl
Branch
CO_2
Methyl-THF
Citrate
Isocitrate
α-keto
Succ-CoA
Methylene-THF
Fatty acids?
ETF_{red}?
ETF_{ox}?
FeS
NiFe HYD
(Group 1)
Methenyl-THF
Bcd/EtfAB
complex
Acyl-CoA
2, 3 Enoyl-CoA
Formyl-THF
WL pathway
Formate
Methyl
Branch
Rnf
H+/Na+
Na+ Mrp
H_2+CO_2
CO_2