Microbiome

RESEARCH

Open Access

Metaproteogenomics resolution of a high-CO_2 aquifer community reveals a complex cellular adaptation of groundwater Gracilibacteria to a host-dependent lifestyle

Perla Abigail Figueroa-Gonzalez[1†], Till L. V. Bornemann[1,2†], Tjorven Hinzke[3,4,5], Sandra Maaß[3], Anke Trautwein-Schult[3], Joern Starke[1], Carrie J. Moore[1], Sarah P. Esser[1], Julia Plewka[1], Tobias Hesse[6], Torsten C. Schmidt[2,6], Ulrich Schreiber[8], Batbileg Bor[7], Dörte Becher[3] and Alexander J. Probst[1,2*]

Abstract

Background Bacteria of the candidate phyla radiation (CPR), constituting about 25% of the bacterial biodiversity, are characterized by small cell size and patchy genomes without complete key metabolic pathways, suggesting a symbiotic lifestyle. Gracilibacteria (BD1-5), which are part of the CPR branch, possess alternate coded genomes and have not yet been cultivated. The lifestyle of Gracilibacteria, their temporal dynamics, and activity in natural ecosystems, particularly in groundwater, has remained largely unexplored. Here, we aimed to investigate Gracilibacteria activity in situ and to discern their lifestyle based on expressed genes, using the metaproteogenome of Gracilibacteria as a function of time in the cold-water geyser Wallender Born in the Volcanic Eifel region in Germany.

Results We coupled genome-resolved metagenomics and metaproteomics to investigate a cold-water geyser microbial community enriched in Gracilibacteria across a 12-day time-series. Groundwater was collected and sequentially filtered to fraction CPR and other bacteria. Based on 725 Gbps of metagenomic data, 1129 different ribosomal protein S3 marker genes, and 751 high-quality genomes (123 population genomes after dereplication), we identified dominant bacteria belonging to Gallionellales and Gracilibacteria along with keystone microbes, which were low in genomic abundance but substantially contributing to proteomic abundance. Seven high-quality Gracilibacteria genomes showed typical limitations, such as limited amino acid or nucleotide synthesis, in their central metabolism but no co-occurrence with potential hosts. The genomes of these Gracilibacteria were encoded for a high number of proteins involved in cell to cell interaction, supporting the previously surmised host-dependent lifestyle, e.g., type IV and type II secretion system subunits, transporters, and features related to cell motility, which were also detected on protein level.

Conclusions We here identified microbial keystone taxa in a high-CO_2 aquifer, and revealed microbial dynamics of Gracilibacteria. Although Gracilibacteria in this ecosystem did not appear to target specific organisms in this

[†]Perla Abigail Figueroa-Gonzalez and Till L. V. Bornemann contributed equally to this work.

*Correspondence:
Alexander J. Probst
alexander.probst@uni-due.de
Full list of author information is available at the end of the article

ecosystem due to lack of co-occurrence despite enrichment on 0.2-µm filter fraction, we provide proteomic evidence for the complex machinery behind the host-dependent lifestyle of groundwater Gracilibacteria.

Keywords CPR bacteria, BD1-5, Groundwater, Genome-resolved metagenomics, Host-dependent bacteria, Symbiotic bacteria, Proteomics

Background

Bacteria belonging to the candidate phyla radiation (CPR, also known as Patescibacteria [1]) are estimated to comprise at least one-quarter of the bacterial domain [2, 3], and are characterized by having reduced genomes as well as small cell diameters (0.1–0.4 µm, [4–7]). In addition, CPR bacteria show a distinct differentiation in protein content and ribosomal structure when compared to other bacteria [4]. Due to the very few cultivated representatives of CPR bacteria, few biochemical studies on their proteins have been conducted resulting in many proteins of unknown function [8]. However, their core metabolic genes are limited and essential pathways for free-living bacteria, such as capabilities for de novo synthesis of lipids or amino acids or the TCA cycle, are usually very incomplete in CPR bacteria. Genomes of CPR bacteria have been shown to encode a substantial amount of proteins with transmembrane helices, as well as proteins which are likely localized in the membrane or secreted [9]. With the increased use of metagenomic-based environmental surveys, CPR bacteria have been detected in a plethora of environments, including hydrothermal vents, soil, lakes, rivers, permafrost, the (deep) terrestrial subsurface and groundwater, and oral cavities of mammals [10–18]. Considering the limited metabolic potential and small genome size of CPR bacteria, which is comparable to symbiotic non-CPR bacteria [8], CPR bacteria likely depend on hosts in order to obtain essential molecular building blocks, such as amino acids, lipids, or co-factors [6, 8]. This host dependency was first shown for the phylum Saccharibacteria, and in particular for *Nanosynbacter lyticus* (also known as TM7x), which was isolated in co-culture with its host, *Schaalia odontolyticus* XH001, from the human oral cavity [15, 19]. While *N. lyticus* can, however, achieve stable growth conditions along with its host [20–22], recent reports of predatory CPR bacteria, specifically of the Absconditabacteria lineage, show that the host is killed for the CPR bacteria to survive, regardless of the environmental conditions [23, 24].

Gracilibacteria, originally named BD1-5 [7, 25], is an understudied phylum of CPR bacteria and forms a clade with the candidate phyla Abscondibacteria and Peregrinibacteria [23]. A trademark characteristic of the Gracilibacteria lineage, similar to Absconditabacteria, is their use of genetic code 25, where the codon UGA encodes for a glycine, as opposed to being a stop codon [26–28]. Like other members of CPR bacteria, Gracilibacteria have very limited metabolic capabilities (e.g., some members encode for the oxidative branch of the pentose pathway but lack steps for glycolysis and TCA). At the same time, Gracilibacteria encode for several surface proteins related to type IV pili and secretion systems [28], indicating that these organisms have the potential for cell–cell interactions.

For Absconditabacteria, two cultivated representatives exist, both having a predatory lifestyle. The first of those is *Candidatus* Absconditicoccus praedator, isolated from a hypersaline alkaline lake along with its host *Halorhodospira halophila*, an obligately anaerobic photosynthetic purple sulfur bacterium [24]. The symbiont's metabolic limitations make it completely dependent on its host, whom it kills by lysing and consuming its cytoplasmic contents. The second cultivated Absconditabacteria member is *Candidatus* Vampirococcus lugosii, discovered in an athalassic hypersaline lake in Spain. Cultures derived from the lake's microbial mat showed that *Ca.* V. lugosii is an epibiotic bacterium that feeds on *Halochromatium*, an anoxygenic photosynthetic Gammaproteobacterium [23]. Similar to *Ca.* A. praedator, the attachment of *Ca.* V. lugosii leads to compromising the cell membrane of its host, so that *Ca.* V. lugosii can consume the *Halochromatium* bacterium's cytoplasmic content, thereby killing it [23, 24]. There is no cultured Gracilibacteria representative as of yet, and the only proteomic report about Gracilibacteria focuses on their alternative genetic code and not on their protein expression, i.e., actual phenotype, related to metabolism or cell–cell interactions [7, 26]. Consequently, the current body of literature misses a thorough study on the expression profile of Gracilibacteria in their natural environment to pinpoint specific cellular adaptations to their host-dependent lifestyle.

In this study, we characterized a model ecosystem (cold-water geyser Wallender Born (WB; 50° 09′ 13.4″ N 6° 43′ 13.1″ E)) enriched in Gracilibacteria, to investigate the proteogenome of Gracilibacteria in a complex microbial community. The geyser is located in the town of Wallenborn, in the Volcanic Eifel region of Germany. The driving force for its eruptions is the accumulation of carbon dioxide (CO_2) in groundwater. CO_2 saturates the groundwater, which leads to the geyser's eruption where groundwater and CO_2 are released to the atmosphere.

These eruptions occur in approximate 40 min intervals. Taking advantage of these groundwater eruptions, we used geyser WB as a study site to characterize its microbial community on the molecular level. For this, we sampled erupted groundwater over a 12-day time series and sequentially filtered it to enrich small- and large-sized (or attached) microbes on 0.1-µm filters and 0.2-µm filters, respectively. We used genome-resolved metagenomics to obtain high-quality genomes and coupled this to metaproteomic analyses of the same time points, to generate in-depth insights into Gracilibacteria metabolism and ecology. Although co-occurrence networks were elusive and resulted in no detection of likely hosts, Gracilibacteria accumulated with the majority of other organisms on the 0.2-µm filter indicating attachment to other organisms due to their small cell size. We identified traits related to cell–cell interactions and nutrient retrieval in Gracilibacteria in the geyser, both in their metabolic potential as well as in expressed proteins, indicating that they have a complex and active machinery for retrieving nutrients from other organisms.

Material and methods

Geological context of sampling site

The cold-water geyser Wallender Born (WB) is located in the town Wallenborn, southwest of the subrecently active Quaternary volcanic field of the West Eifel region, in the Rhenish Massif (50° 09′ 13.4″ N 6° 43′ 13.1″ E). It is located between the Manderscheid anticline to the southeast and the Salmtal syncline to the northwest which formed during the Variscan orogeny event. It consists of weakly metamorphic early Devonian (Emsian) sedimentary units [29]. From a geological point of view, it belongs to the eastern edge of the old depression zone of the Eifel north–south zone (Eifeler Nord-Südzone). This zone is characterized both by the Middle and Late Devonian Eifel limestone synclines, preserved in the depression, and by numerous N-S to NNE-SSW trending fault zones [30]. The strong CO_2 outgassing is related to the magmatic activity of the West Eifel region, which together with the East Eifel volcanic field is attributed to a mantle plume underneath the Eifel (Eifel plume). Helium R/Ra values from the Wallender Born geyser show a clear mantle influx of ^{3}He, implying magmatic intrusion [31]. Please refer to Supplementary Figure S1 for a scheme of the geyser.

Groundwater sampling and geochemical analyses

Time-series sampling of biomass was done for 12 days, from the 14th to 25th of October 2020. Groundwater was collected in sterile, DNA-free containers during each eruption, for immediate filtering on-site. The groundwater was filtered sequentially on 0.2-µm (JGWP14225, Merck Millipore) and then 0.1-µm (JVWP14225, Merck Millipore) filters (see Supplementary Figure S2A for a scheme of the filtering setup). Before starting the filtration of each time point, the system was rinsed with groundwater to get rid of stagnant water left over from the previous sampled time point. Additionally, one bulk sample, filtered directly onto a 0.1-µm filter, was obtained. Approximately 100 L was filtered on each 0.2-µm filter, with the flow-throughs of two 0.2-µm filters being combined via a T-piece and filtered onto one 0.1-µm filter (please refer to Supplementary Table S1 for individual filtered volumes). Filters were then immediately transferred to a sterile falcon tube, stored on dry ice on-site, and later on transferred to − 80° C until further processing. Temperature and pH were measured in triplicates with a thermometer and pH strips, respectively.

Total iron A total of 100 µL of groundwater was mixed with 900 µL of 1 M HCl on-site. Samples were taken in triplicates and kept at 4 °C till measurement. Total iron was measured by mixing of 100 µL of aliquot with 900 µL hydroxylamine, followed by addition of 80 µL reduced aliquot to 120 µL Ferrozine, incubation for 10 min in the dark, and photometric determination by a Tecan plate reader at 560 nm [32].

Ions A total of 0.1-µm pore-size filtered water was used for ion measurements. Samples were measured in dilutions of 1:100 and 1:500 in distilled water to be able to accurately quantify both highly abundant ions and lower abundant ions. Samples were analyzed using the Dionex Aquion ion chromatography system (Thermo Scientific, USA). Anions were analyzed with a Dionex IonPac AG23-4 µm guard column, a Dionex IonPac AG23-4 µm 2×250 mm analytical column, and an AERS 500 Carbonate 2 mm suppressor, Ultimate 3000 heating element, and a DS6 heated conductivity cell detector. Cations were measured on a CS12A RFIC 2×250 mm analytical column and a CG12A RFIC detector. Measurements were done in technical triplicates.

Total organic carbon (TOC) Unfiltered groundwater samples were stored at 4 °C until measurement. Prior to measurement, the samples were filtered through a 0.45-µm filter to remove large particles. Ten milliliters of sample were augmented with 0.5 mL 1 M HCl to remove residual carbonate present in cold-water geysers, verified via pH strips. For each time point, two biological samples were measured in triplicate on a TOC-5050 Analyzer with ASL-5000 (Shimadzu) and averaged after accounting for the HCl dilution.

Epifluorescence microscopy

For each time point, groundwater was filtered on-site through a 0.2-µm syringe-filter using a sterile syringe. The flow-through was collected in a sterile, DNA-free tube and subsequently filtered through a 0.1-µm syringe-filter using a second sterile syringe. Filters were incubated with 10 µg mL^{-1} DAPI in 2% (v/v) formaldehyde for 5 min, washed with distilled water, dried for 10 min, and stored in the dark at 4 °C until use. Cells were quantified using an Axio Imager M2m epifluorescence microscope equipped with an AxioCam MRm and Zen 2 Pro software (Carl Zeiss Microscopy GmbH, Jena, Germany; version 2.0.0.0). Enumeration of the cells was done in ten horizontal and ten vertical fields, extrapolating counts to the entire filter area and normalizing through the filtered volume to calculate cells per milliliter.

DNA extraction and sequencing

Each pair of 0.2-µm filters was pooled to have a comparable sample to the respective 0.1-µm filter (see Supplementary Figure S2B for a scheme of the pooling for metagenomics). This pooling resulted in 22 samples for the 0.2-µm filtered fractions, each corresponding to the different time points obtained during the sampling campaign, with their respective 0.1-µm filtered fraction. DNA extractions of the samples were performed using the Dneasy PowerMax Soil DNA Extraction kit (Qiagen, 12,988–10) according to the manufacturer's instructions and further concentrated using ethanol precipitation with glycogen as the carrier. Of the 44 samples, 30 yielded enough biomass for library preparation via the Westburg NGS DNA Library Prep Kit (cat. No. WB 9096). Of these, 28 were successfully sequenced using Illumina NextSeq500 (2×150 bp paired-end reads). Some libraries which yielded low sequencing depth in initial sequencing runs were sequenced up to a total of three times and the reads were concatenated to reach about 20 Gbps sequencing depth per sample.

Sample preparation for metaproteomics

Sample sets analyzed for metaproteomics differed from those for metagenomics, due to the higher biomass requirements for metaproteomics as compared to metagenomics. A total of three different metaproteomics sample sets were analyzed (see Supplementary Figure S2C and Supplementary Table S3 for details): (I) We compiled a time-series consisting of individual 0.2-µm filters taken at different time points. (II) To compare the proteome of the size-fractionated microbiomes, we separately pooled 0.2-µm filters and 0.1-µm filters of the respective time points to generate biomass consistent of only microbes from the 0.2-µm fraction and of the 0.1-µm fraction, respectively. This provided sufficient biomass for the 0.1-µm filter for most protein extraction but also generated one set of 0.2-µm filter samples without the corresponding 0.1-µm sample due to too-low biomass. (III) For a whole community, we used a bulk sample filtered directly onto a 0.1-µm filter (without size fractionation).

Protein extraction

Protein extraction from the filters was performed according to previously published protocols [33] with some modifications. In brief, the filters were cut in pieces, transferred to low protein binding reaction tubes, and covered with one volume of resuspension solution 1 (50 mM Tris–HCl pH 7.5, 0.1 mg mL^{-1} chloramphenicol, 1 mM PMSF (phenylmethanesulfonyl fluoride)). After vortexing, 1.5 volume resuspension solution 2 (20 mM Tris–HCl pH 7.5, 2% (w/v) SDS) were added, and the samples were incubated at 60 °C for 10 min with vigorous shaking. After cooling down of the samples to room temperature, 5×volume of DNAse solution (1 µg mL^{-1} DNAse I) were added. Samples were lysed by ultrasonication on ice for 6 min (Sonopuls MS72 (Bandelin, Berlin, Germany), amplitude 51–60%; cycle 0.5), and subsequently incubated at 37 °C for 10 min with vigorous shaking. After centrifugation, proteins were precipitated from the supernatants by adding pre-cooled trichloroacetic acid (TCA) to reach 20% (v/v) TCA and incubating at 4 °C for 30 min in an overhead inverter. Precipitates were pelleted and washed with pre-cooled acetone. The dried pellet was resuspended in 2×SDS sampling solution (0.125 M Tris–HCl pH 6.8, 4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) ß-mercaptoethanol), and incubated in an ultrasonic bath for 15 min before heating at 95 °C for 5 min. Samples were centrifuged, supernatants were kept, and the remaining pellet was again treated with 2×SDS sampling solution. Both sample fractions, supernatant and pellet, were loaded onto pre-cast SDS Gels (Criterion TGX 4–20%, 12+2 wells, Bio-Rad) and separated by SDS-PAGE. After staining with Coomassie, proteins were in-gel digested as described previously [34]. Briefly, gel lanes were excised in ten equidistant pieces and each piece was destained, washed, and subjected to trypsin digestion (as well as mass spectrometric measurement) individually. Peptides were eluted in water using an ultrasonic bath and desalted by the use of C18 ZipTip columns (Merck) according to the manufacturer's guidelines.

Metagenome processing

Illumina paired-end reads were checked for adapters and sequencing artefacts with Bbduk v37.09 (Bushnell, https://sourceforge.net/projects/bbtools), quality-trimmed with Sickle v1.33 (https://github.com/najoshi/sickle), and deduplicated with dedup (Bushnell, https://

sourceforge.net/projects/bbtools). The diversity coverage in the samples based on raw reads was estimated using Nonpareil3 [35] in kmer mode with k = 20 (i.e., the minimal read length). For each individual metagenome, quality-controlled reads were first assembled with Metaviralspades 3.15.2 [36] to reconstruct viruses and plasmids, and paired reads where no read mate mapped to assembled viruses or plasmids (via Bowtie2 v2.3.5.1, sensitive mode, [37]) were then assembled using metaspades 3.15.2 [38]. Viral and normal assemblies were then concatenated to yield the final assembly. Unless otherwise stated, all mappings were performed with Bowtie2 [37] in sensitive mode for, e.g., determining relative scaffold abundances of the assembled scaffolds. Open reading frames were predicted on scaffolds with a minimum length of 1000 bps using Prodigal 2.6.3 [39] in meta-mode and annotated against FunTaxDB 1.3 [40] to retrieve function and taxonomic profiles.

Mass spectrometry and data processing of metaproteomics samples

Peptides were subjected to LC–MS/MS analyses on a LTQ Orbitrap Elite instrument (ThermoFisher Scientific) coupled to an EASY-nLC 1200 liquid chromatography system. Therefore, peptides were loaded on a self-packed analytical column (OD 360 µm, ID 100 µm, length 20 cm) filled with 3 µm diameter C18 particles (Maisch) and eluted by a binary nonlinear gradient starting at 5% up to 99% acetonitrile in 0.1% (v/v) acetic acid over 82 min with a flow rate of 300 nL min^{-1}. For MS analysis, a full scan in the Orbitrap with a resolution of 60,000 was followed by collision-induced dissociation (CID) of the 20 most abundant precursor ions. Database searches against a concatenated forward-reverse protein database derived from the described metagenome analyses (see "Generation of protein database from metagenomes") (15,269,672 unique entries) were performed using Mascot (Matrix Science; version 2.7.0.1), applying the following parameters: tryptic digestion, fragment ion mass tolerance 0.5 Da, and parent ion tolerance of 10 ppm, up to two missed cleavages as well as methionine oxidation as a variable modification. Scaffold v5.1.2 (Proteome Software Inc.) was used to merge the search results and to validate MS/MS-based peptide and protein identifications. During creation of the Scaffold file, an additional X! Tandem search with default settings was performed for validation. Protein and peptide-level false discovery rate (FDR) were set at 5% and protein groups required to contain at least 1 unique peptide. Protein groups consisted of proteins which were not distinguishable based on identified peptides.

For comparison of metaproteomes across samples, normalized spectral abundance factors (NSAFs) were calculated by normalizing spectral counts by the protein length, followed by total sample spectral count, as described previously [41].

Community profiling based on rpS3 marker gene

Genes encoding for the *rpS3* marker gene were identified via hmmsearch [42] against the Phylosift HMM model [43] (PhyloSift Marker: DNGNGWU00028) at an E-value of 1e-28 as well as by pulling out genes annotated as *rpS3* during data processing (see above). All identified *rpS3* gene sequences throughout all assemblies were then pooled and clustered at 99% similarity as described in Sharon et al. [44] using the usearch -uclust module [45]. Mapping on just the *rpS3* gene would result in inadequate abundance estimates, as reads mapping to the edges of sequences map poorly, causing coverage to drop off at the edges. Hence, for each *rpS3* gene sequence, the scaffold area around the gene was taken into account by extending the scaffold sequence around identified *rpS3* gene sequences by up to 1000 bp to either side [46]. The workflow and scripts to identify, cluster, and extract *rpS3* gene sequences as well as calculate their coverage and breadth are provided at https://github.com/ProbstLab/publication-associated_scripts/tree/main/Figueroa-Gonzalez_Bornemann_etal. This yielded scaffold fragments of about 3500 bp in length, which are expected to result in more accurate abundance estimates than just mapping on the gene by itself. Representative *rpS3* gene sequences were selected from each *rpS3* gene cluster by order of the following priorities: (1) centroid and can be extended by 1000 bp into both sides of the gene; (2) not centroid and can be extended into both sides by 1000 bp; and (3) longest scaffold region. The resulting *rpS3* gene sequences, ideally extended in both directions by 1000 bp, are henceforth called rpS3extended. The taxonomy of the rpS3extended sequences was determined by using usearch (ublast) against the pooled archaeal and bacterial *rpS3* gene sequences of the GTDB (E-value 0.00001). The best hits from the search were taken, and those rpS3extended sequences without a hit at this threshold were assigned as "Unclassified." The abundance of the sequences in each sample was determined by mapping of reads back to the rpS3extended sequences and their average coverage per sample was determined. Additionally, the breadth, i.e., the number of nucleotide positions having a coverage of at least 1, was determined. The coverage of an rpS3extended sequence in a given sample was set to zero if the breadth of the sequence in that sample was less than 95%. This was done to ensure that the respective sequences were present in their entirety in the respective sample, instead of just fragments of them. After breadth correction, coverages per sample were normalized by sequencing depth.

Reconstruction of metagenome-assembled genomes

Reads of all metagenomes were mapped against each assembly to yield differential coverage data used in binning. Metagenome-assembled genomes (MAGs) were reconstructed using ABAWACA v1.0.0 (https://github.com/CK7/abawaca.), CONCOCT v1.1.0 (https://github.com/BinPro/CONCOCT), and MaxBin 2.0 v2.2.7 [47]. The bin sets were consolidated with DAS tool v1.1.2 [48] with default parameters. The consolidated bin sets were curated in uBin v0.9.20 [40] and all curated MAGs of all samples were then pooled and dereplicated with dRep [49] at 99% gANI to yield all unique strains present in the dataset. CheckM1 [50] was used to assess and filter genomes based on completeness and contamination, using 70% and 10% as the cutoffs, respectively. For bacteria taxonomically classified to be Patescibacteria based on GTDB-tk (see Supplemental Table S6), a HMM marker set specific for the CPR bacteria clade was used instead of the general CheckM1 models (https://www.github.com/Ecogenomics/CheckM/wiki/Workflows#using-cpr-marker-set).

Taxonomic classification of MAGs

GTDB-tk v.2.1 [51] with the classify_wf workflow against GTDBr207 was used to determine the taxonomic origin of the recovered MAGs. Genes on these genomes were predicted with prodigal in normal mode, either with default genetic codes (for non-Gracilibacteria) or codon code 25 (for Gracilibacteria). To generate phylogenetic trees for Fig. 2 and Files S3 and S4, GTDB-tk v.2.1 with the de novo workflow, GTDBr207, and p__Firmicutes or p__Thermoplasmatota were used as outgroups for Bacteria or Archaea, respectively.

Generation of protein database from metagenomes

Open reading frames (ORFs) in amino acid format from all metagenomic assemblies, predicted by prodigal in meta-mode, were pooled. ORFs located on scaffolds that had been assigned to Gracilibacteria bins were exchanged with ORFs predicted on the entire Gracilibacteria genome and with codon code 25. Then, all ORFs were uniquified using the *usearch* command [45] to form the database. Each database entry was renamed to Protein_XXXXXXX, with X representing a sequential, unique number for each entry (this database is supplied as Supplementary File S1). Additional metadata, such as whether the respective protein or any of its cluster members were binned and if so, to which bin they belonged as well as their functional or taxonomic annotations, were stored in separate files (see Supplementary File S2).

Tracking abundances of rpS3 gene sequences and MAGs across samples

Each metagenome was mapped to the dereplicated MAGs or rpS3extended sequences to estimate their abundances across time and filter fractions. Additionally, the breadth of each MAG and rpS3extended sequence in each sample was determined. The breadth is herein defined as the percentage of the sequence length having a coverage of at least 1. This was used to filter out MAGs and rpS3extended sequences that only had coverage because a small proportion of their sequence was shared with other organisms. For genomes, the minimum required breadth to have a valid coverage was set to 30%, for rpS3extended sequences to 95%. If the breadth of a sequence in a sample was below this threshold, the coverage was set to zero.

Normalization of abundances of MAGs and rpS3 genes

MAG and rpS3extended abundances were normalized to the sequencing depth per sample using the following formula:

$$NormMAGi = MAG * (\max(SeqDepth)/SeqDepthi)$$

NormMAG_i = Normalized MAG (or rpS3extended) abundance in sample i.

max(SeqDepth) = Maximum sequencing depth in bp across all samples.

SeqDepth_i = Sequencing depth in sample i.

Shannon and Simpson diversity estimates of assemblies

Shannon and Simpson diversity estimates were calculated from rpS3extended sequence abundance tables using the diversity function of the vegan R package (https://github.com/vegandevs/vegan).

Rank abundance estimates of *phyla*

The normalized coverage of rpS3extended sequences assigned to the same phylum based on GTDB annotation were summed up by sampled time point and by each filter fraction (0.1 µm or 0.2 µm).

Non-metric multidimensional scaling

NMDS, based on Bray–Curtis dissimilarities, were calculated using the *metaMDS* function of the *vegan* package, using species tables of either MAGs or rps3extended sequences. All samples were analyzed while being pooled together as one database, as well as split into the two filter fractions. Only NMDS plots with a convergent solution (after a maximum of 400 tries) as well as sufficiently low stress are shown. To validate whether there were significant differences between 0.1-µm and 0.2-µm fractions

or whether time played a significant role the adonis2 function of the vegan package with the default 999 permutations was used.

Proportionality networks

The temporal dynamics of MAGs across time in the two filter fractions were investigated using proportionality networks implemented in the R package *propr* [52]. In contrast to conventional correlation analyses, proportionality networks are not biased when relative data is compared [53]. The proportionality measure rho (ρ) was employed and cutoff thresholds were determined to correspond to false-discovery-rate (FDR) thresholds of 1% and 5%, respectively, using the *updateCutoffs* function.

Metabolic analyses

Metabolic potential of the microbial community was determined using METABOLIC [54]. Genome annotation of Gracilibacteria was done using the Genoscope platform MAGE [55]. In-depth metabolic potential was determined by combining the annotation information from MAGE with information from MetaCyc [56], KEGG [57], and UniProt [58] (see Table S7 for automatic annotation results). Furthermore, unclassified proteins were manually searched against protein domain databases of NCBI protein domains (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and AlphaFold-EMBL database ((https://www.ebi.ac.uk/Tools/sss/fasta/), see Table S8 for manual annotation results). The information obtained in the analyses of automatic annotation was compared with the metaproteomics data to evidence expressed proteins.

Comparison of available microbial populations from cold-water geysers

All dereplicated MAGs from Crystal Geyser (UT, USA) [59], Geyser Andernach (Andernach, Germany) [60], and the Geyser Wallender Born (Wallenborn, Germany) were pooled and compared via FastANI [61] with a ≥ 50% coverage.

Visualization

Figure panels, including the NMDS, were generated with *ggplot2* [62] in R [63], and subsequently assembled and post-processed in AffinityDesigner V2 (https://affinity.serif.com/). The phylogenetic trees produced by GTDB-tk were visualized using Dendroscope v3.5.10 [64] and processed with iTol v6 [65]. The proportionality networks were generated using the *iGraph* package [66]. The number of shared genomes between the cold-water geysers was analyzed using *ggupset (*https://github.com/const-ae/ggupset*). The metabolic reconstruction scheme was generated manually in Affinity Designer V2.*

Results and discussion

Microbial community of the geyser is dominated by Gallionella and Gracilibacteria

Size fractionation of 22 groundwater samples discharged from Wallender Born resulted in 15 samples on 0.2-µm filters, 14 samples on 0.1-µm filters, and one bulk 0.1-µm filter sample. The subsequent metagenomic sequencing resulted in 725 Gbps of Illumina NextSeq data. These data covered 78% of the microbial diversity (on average) in the samples based on Nonpareil3. After individually assembling the metagenomes, we determined that about 80.0% (± 3.2% STDEV) of the reads mapped back to the scaffolds, with only slight differences between 0.1- and 0.2-µm filter fractions (78.2 ± 2.9% STDEV and 82.6 ± 1.9% STDEV, respectively (see Supplementary Table S4 for reads, assembly, and Nonpareil3 statistics). We hence conclude that the assemblies are representative of the microbial community of the geyser. The Shannon diversity index based on *rpS3* marker gene sequences was on average 3.22 for 0.1-µm size fractions, and 3.50 for 0.2-µm size fractions, while the Simpson diversity index was on average 0.90 for both size fractions with similar intra-sample variation (0.42 to 0.11, and 0.30 to 0.01, respectively; Fig. 1A). Based on *rpS3* abundances (Fig. 1B), the communities were dominated by Proteobacteria (58.0%), in particular *Gallionella*, followed by Patescibacteria (26.2%), mainly belonging to Gracilibacteria, and Bacteroidota (4.57%; please refer to Supplementary Figure S3 for the complete rank abundances and Supplementary Figure S4 for phylum-level taxa abundances across time-series samples). NMDS analyses revealed that the 0.1-µm filter fractions showed a great intra-group dissimilarity, and that 0.2-µm samples and the bulk sample clustering in close proximity (Fig. 1C). Individual NMDS analyses of 0.1-µm and 0.2-µm filter fractions confirmed this finding, with the NMDS of the 0.1-µm samples being much more dissimilar compared to their 0.2-µm counterparts (Fig. 1C). The differences between 0.1- and 0.2-µm filters were additionally confirmed via PERMANOVA, indicating that the communities on both filter fractions were significantly different (p-value < 0.001). Time did not play a significant role (p-value = 0.135).

Metaproteogenomics reveals potential key stone organisms in the geyser community

We reconstructed 751 metagenome-assembled genomes (MAGs), with completeness of at least 70% and contamination of at most 10%, based on CheckM1 estimates. These MAGs belonged to 123 strain clusters after dereplication at 99% ANI (see Supplementary Table S6 for genome information). Phylogenetic analysis based on GTDB-tk indicated that these 123 dereplicated MAGs

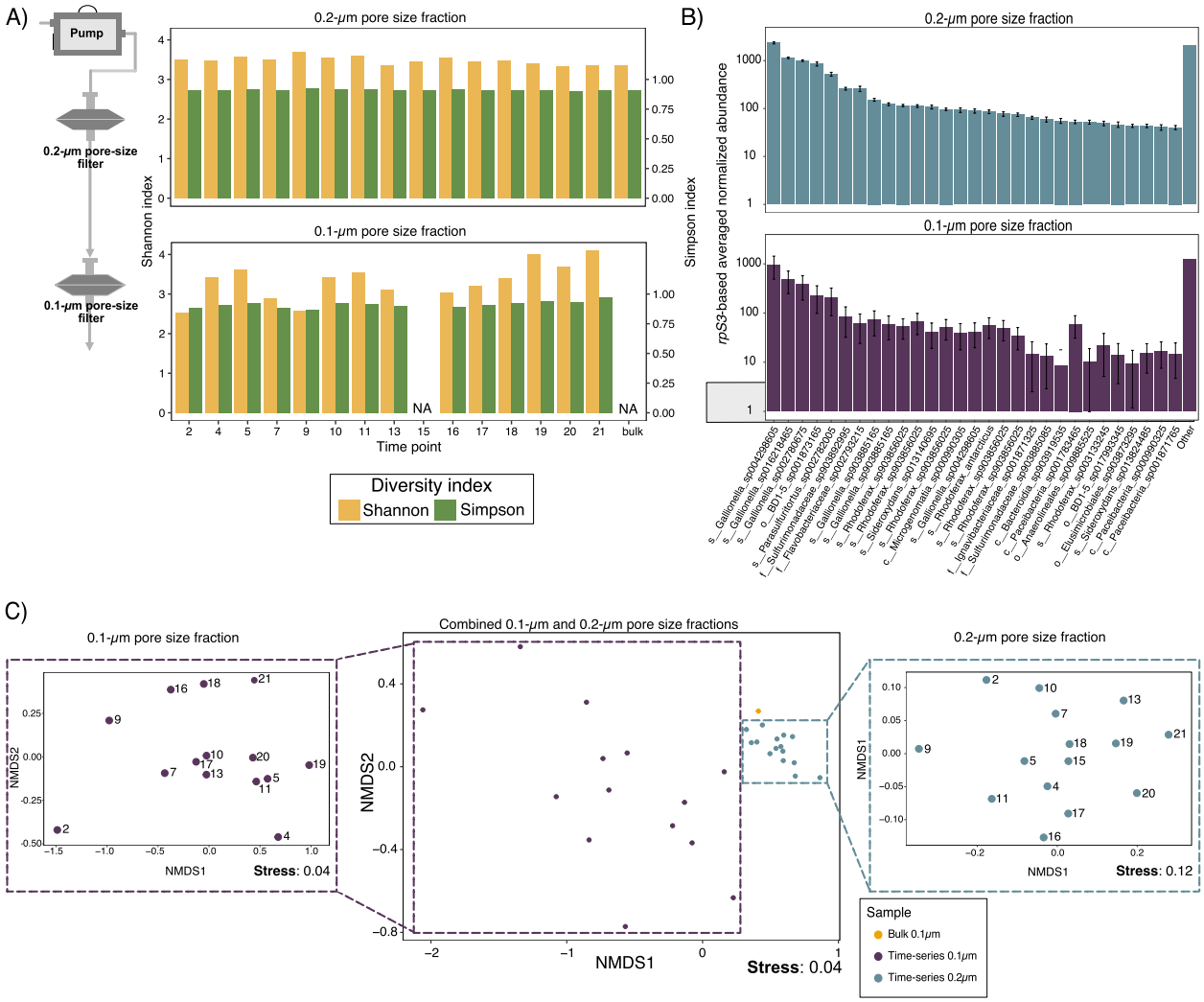

Fig. 1 Microbial community composition of the geyser Wallender Born. **A** The scheme in the left illustrates the sequential filtration we used for this study. The bar graph on the right displays the Shannon and Simpson diversity indices for the 0.2-µm and the 0.1-µm filtered fractions. NA denotes samples that we were unable to sequence due to low biomass. **B** Normalized average abundance, based on rpS3 gene, for both the 0.2-µm and the 0.1-µm filtered fractions. The taxonomy is based on GTDB-tk classification. The figure illustrates the top 29 species with the rest of the species pooled under "Other." Error bars denote standard deviation (not shown for all low abundant organisms grouped as Other). **C** NMDS, based on rpS3, of both filtered fractions, and individual NMDS for the 0.2-µm and the 0.1-µm fractions. Central panel depicts the NMDS of both filters, while side panels show separate NMDS for 0.1-µm and 0.2-µm filters, respectively

belonged to 26 different phyla, including Nitrosomonadales, Proteobacteria, Nitrospirae, and Bacteroidetes, and multiple bacteria of the Candidate Phyla Radiation (CPR), in particular Nomurabacteria and Gracilibacteria (Fig. 2). Microbial communities, in particular those of the 0.2-µm filter fraction, were stable across the sampled time series, with the filter fraction being the main discriminatory factor for microbial communities (Fig. 3A). Ordination analyses based on relative abundance of MAGs agreed with *rpS3*-based ordinations, displaying that there was a significant difference between the communities on the 0.1-µm filters compared to those of the 0.2-µm filters, which were comparatively stable (Fig. 1C). Comparison with known cold-water geyser microbial communities [59, 60] (Geyser Andernach (GA), Germany; and Crystal Geyser (CG), USA) showed that most community members were specific to their respective ecosystems, with only Altiarchaea and Hydrogenophilales being present in all three ecosystems (Supplementary Figure S5). While Altiarchaea were the dominant organisms in both GA and one aquifer in CG, they constituted a minority in WB. Out of the 123 WB genomes, 24 had related populations based on FastANI (≥ 50% genome coverage, ≥ 75% ANI, [61]) in the CG ecosystem while

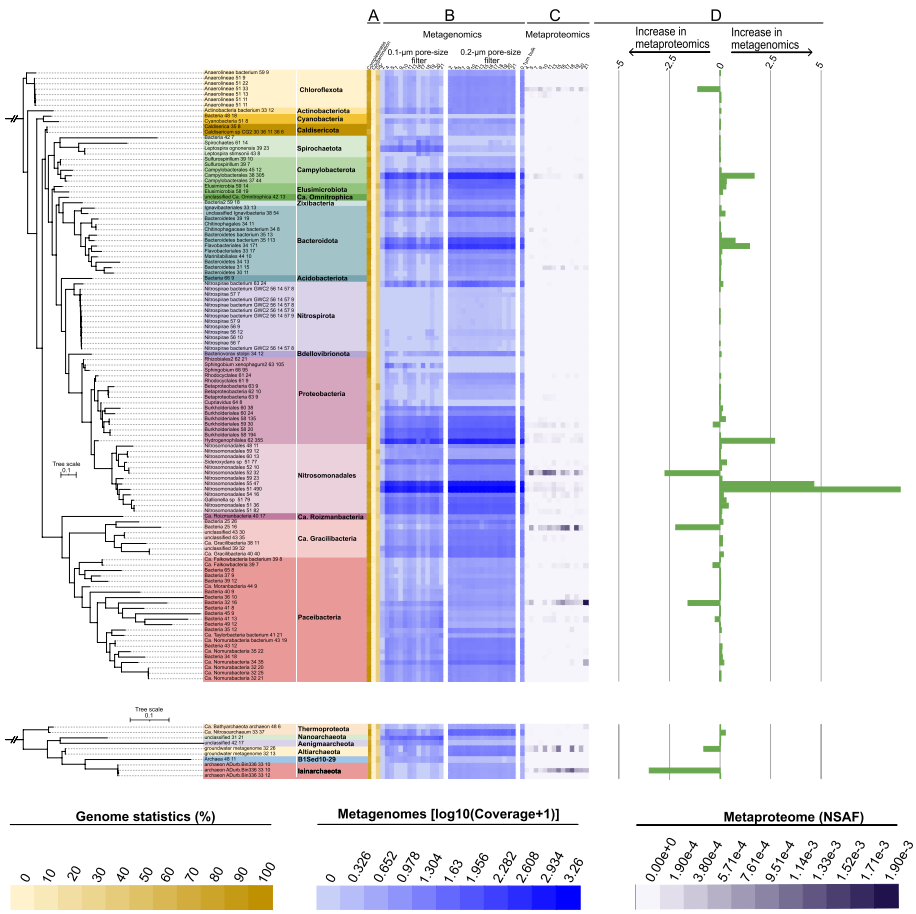

Fig. 2 Phylogeny of high-quality genomes and their prevalence in metaproteogenomics. Colored sections to the right of the phylogenetic tree denote the phylum-level taxonomic classifications of high-quality genomes. Phylogenetic relationships displayed in the trees were determined using the GTDB-tk de novo workflow, using either Firmicutes or Thermoplasmatota as the outgroups for Archaea and Bacteria, respectively. **A** Yellow-colored heatmap indicates the genome statistics (completeness, contamination, and GC % content). **B** The blue heatmap shows the abundance of the genomes throughout the time-series, separated into 0.1-µm filter and the 0.2-µm filter fractions. **C** The purple heatmap conveys the abundance of the genomes based on the summed counts (NSAF) of proteins on individual 0.2-µm pore size filters belonging to the respective genomes. **D** Green bar charts denote the difference in the median relative abundance of the metagenomic-based genome abundance relative to the proteomic-based genome abundance on the 0.2-µm filters. The median relative abundance was calculated by setting the total sample abundance to 100%, both in metagenomics and in metaproteomics, and afterwards calculating the median percentage per genome. Negative numbers indicate higher relative abundance in metagenomic data than in the metaproteomic dataset, and vice versa for the positive percentage numbers

only a single population genome was related between the WB and GA populations, which might relate to the larger diversity of the CG ecosystem compared to the other two ecosystems.

Although other high-CO_2 geysers have been metagenomically analyzed in the past [16, 60, 67], little is known about the activity or the contribution of the organisms to the protein pool of the community. To elucidate the expression profile (i.e., expressed phenotype and metabolic functions) of the detected organisms in the metagenomes of geyser WB, we performed metaproteomics of the same samples that were analyzed for metagenomics (see Supplementary Figure S2 for sample distribution of metagenomics and metaproteomics). Metaproteomic datasets showed a batch effect in the NMDS analysis, with the three categories of samples (time-series, size, and bulk) clustering separately in the ordination (Supplementary Figure S6). Subsequently, we focused our analyses on the 0.2-µm fraction only and compared metaproteomic and metagenomic data of these samples. Across time, the proteomics profile of the community in the 0.2-µm fraction was more dynamic than the relatively stable genome abundances, with only few organisms showing consistent expression throughout the time series (Fig. 2B–D). Only the most abundant organisms in metagenomics (including a *Campylobacter*, a

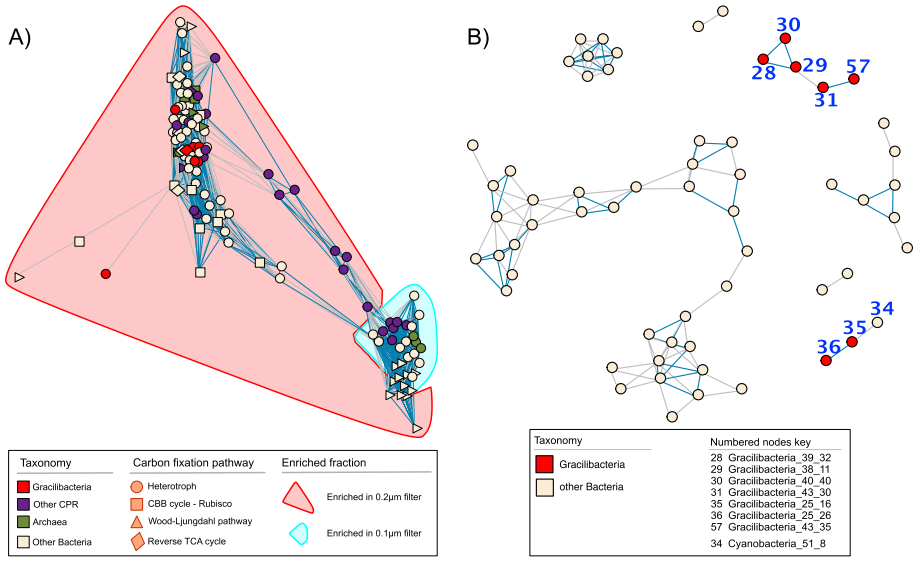

Fig. 3 Co-occurrence networks of unique genomes. **A** Co-occurrence of the overall prokaryotic community of WB; color of shape denotes taxonomy; shape relates to carbon fixation pathway detected in each genome; enrichment of genomes in either the 0.2-µm or 0.1-µm filter fraction is shown by the areas highlighted in red or blue. **B** Co-occurrence network of 0.2-µm filter samples focusing only on Gracilibacteria and non-CPR bacteria. Blue connecting lines indicate 1% FDR, while gray lines signify 5% FDR. Gracilibacteria were co-correlated and showed only one weak correlation with a cyanobacterial MAG

Hydrogenophilales, and two Nitrosomonadales closely related to *Gallionella*) were also fairly consistently represented in the proteomic datasets (Fig. 2). However, the most abundant members based on proteomics included organisms of the Anaerolineae, a Paceibacteria, a Gracilibacteria, and some Nitrosomonadales, which were all only minor community members in the metagenomic dataset (Fig. 2B–D). Based on their consistently high abundance in both metagenomes and metaproteomes, we conclude that these CPR bacteria likely represent key organisms in the microbial community of the geyser.

Gracilibacteria are enriched in the 0.2-µm fraction and show no co-correlation with other prokaryotes

CPR bacteria, in particular Nomurabacteria and Gracilibacteria, represented 31 of the 123 unique MAGs recovered from geyser WB. Due to their predicted symbiotic lifestyle [4, 8], we sought to reconstruct co-occurrence networks to establish potential symbiont-host relationships. As correlating relative abundance data, such as genome abundances, is problematic due to relative data by default fostering significant correlations [53], we utilized the proposed alternative proportionality [53], implemented in the R package "*propr*" [52]. The resulting network showed two clusters of organisms (Fig. 3A), which has previously been suggested to be related to different aquifers and thus groundwater sources at Crystal Geyser [59, 68]. Nonetheless, further analyses showed that the clusters observed for WB are due to the fractionation, where microbes group based on their enrichment pattern in either the 0.1-µm or 0.2-µm filter fraction. We observed that the WB Gracilibacteria genomes are enriched in the 0.2-µm fraction, indicating that the majority of them were attached at the time of sampling, while other CPR were either distributed between both network clusters or situated between the clusters.

When focusing the analysis on the dataset from 0.2-µm filters, on which both hosts and host-symbiont attachments should accumulate, the proportionality network showed only one weak link for one out of seven Gracilibacteria with another organism, a Cyanobacterium (genome Cyanobacteria_51_8; 5% FDR; Fig. 3B). *Vampirococcus*, closely related to the WB Gracilibacteria (Fig. 4A), have been reported to predate on autotrophs such as *Chromatium* and *Halochromatium* [23, 69]. Although there is one report of viable Cyanobacteria surviving on a hydrogen-based metabolism in a 613-m-deep borehole in Spain [70], these organisms can be contaminants of open geysers [60]. Assuming their inactivity in this geyser system, these Cyanobacteria may still fall prey to Gracilibacteria. However, missing correlations with other organisms could also be related to the study design (e.g., increasing/decreasing time interval for sampling [71]). Nevertheless, correlation and proportionality networks have previously been discussed to be biased towards organisms with

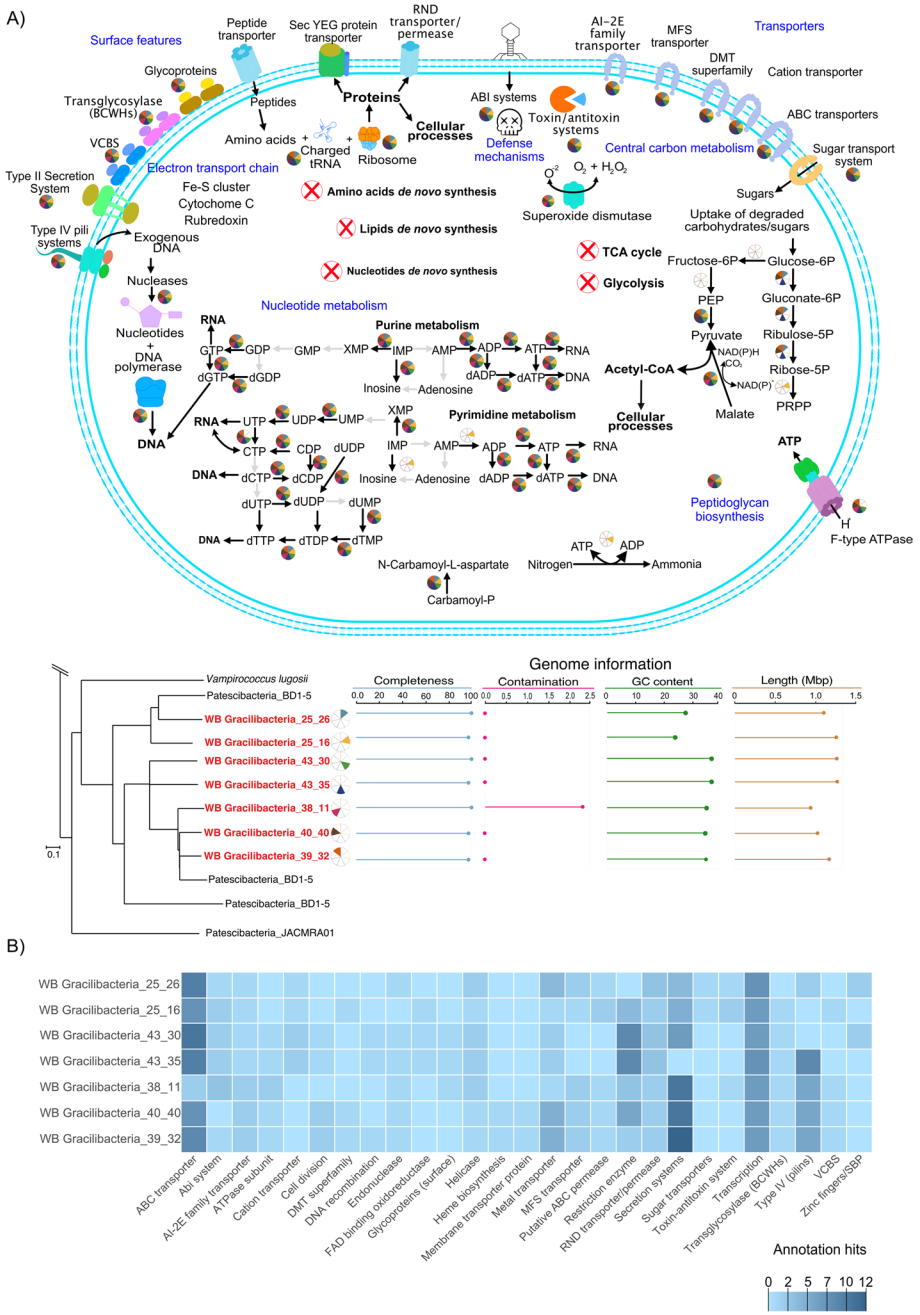

Fig. 4 Overview of WB Gracilibacteria metabolism as inferred metagenomic analysis. **A** Schematic overview of the major metabolic traits found in the WB Gracilibacteria and phylogenetic placement of the seven genomes along with genome statistics. Completeness and contamination were estimated using a custom CheckM1 model for CPR bacteria (see "Methods" for further details). Please note that the individual wedges of the pie-charts relate to the presence of individual metabolic steps encoded in the different genomes, depicted in the metabolic scheme. **B** Specific proteins of interest that were identified in the genome annotation of Gracilibacteria

scavenging lifestyles (based on Lotka-Volterra models; [72]). Based on their enrichment on the 0.2-μm filters throughout all samples signifying attachment to other microbes and based on the fact that no clear host links were revealed based on our network analysis, we conclude that the Gracilibacteria in WB geyser target multiple hosts of different phylogenetic lineages.

Metaproteomics suggests an active host-dependent lifestyle of WB Gracilibacteria

WB Gracilibacteria lacked all of the steps related to glycolysis, with the exception of genes encoding for the conversion of phosphoenolpyruvate (PEP) to pyruvate (Fig. 4A, Supplementary Figure S7). The latter could enable the generation of ATP via substrate level phosphorylation using pyruvate kinase (EC. 2.7.1.40), as previously hypothesized for *Vampirococcus lugosii*, a cultivated member of the related Abscondidabacteria [23]. Additionally, WB Gracilibacteria MAGs encoded for a variety of electron carrier proteins (e.g., Cytochrome C, Fe-S cluster proteins, rubredoxin, peroxidases) and F-type ATPase subunits. Given the presence of electron transport systems, the WB Gracilibacteria could establish a proton gradient to make use of the ATPase for energy generation and use the electron carrier proteins to maintain oxidative stress under control [23]. Energy generation via the use of ATPase is further evidenced by the F-type subunits identified in the proteomic analyses (see Supplementary Figure S8, "ATPase" category). By contrast, it has been argued that CPR bacteria in general could take electrons from their host and use ATPase in a reverse manner (consuming ATP) to drive their antiporters [23], aiding in the uptake of external molecules. Given the patchiness of their central carbon metabolism, the annotation of two enzymes in the TCA cycle, namely pyruvate formate lyase (activating enzyme, EC. 1.97.1.4) and ATP citrate (pro-S)-lyase (EC. 2.3.3.8), suggests that WB Gracilibacteria need to fill metabolic gaps with externally derived compounds, such as citrate.

The seven Gracilibacteria genomes also encoded for interconversion steps for the purine and pyrimidine biosynthesis pathways, with one genome (WB Gracilibacteria 26_16) encoding for the conversion of ribose-5P to phosphoribosyl diphosphate (PRPP) (Fig. 4A). The pyrimidine interconversion steps start with UMP/dUMP and ultimately yield CTP/dCTP that can be used to synthesize RNA and DNA, respectively. As for purines, the conversion steps go from IMP to ATP and dATP which, in a similar manner as pyrimidines, can serve as substrate for the synthesis of RNA and DNA, respectively. While PRPP is an important metabolic intermediate, in particular for the biosynthesis of nucleotides [73], most of the WB Gracilibacteria do not encode for its synthesis (Fig. 4) suggesting that this intermediate has to be retrieved from external sources.

The limited metabolism of the WB Gracilibacteria suggests that these bacteria require molecular building blocks from other organisms (refer to Supplementary Figure S7 and Table S7 for the full KEGG annotation of these genomes), but it remains unclear how they obtain these necessary biomolecules. Given their extremely limited central metabolism, and that we did not identify potential host(s) for the WB Gracilibacteria using co-occurrence patterns, we set out to investigate their potential lifestyle based on metaproteogenomics. For that, we focused on (trans)membrane and cell surface features (see Fig. 4B and Supplementary Table S8 for proteome annotations of Gracilibacteria MAGs). We identified membrane-associated proteins predicted to be functioning as transporters, in particular ABC transporters, which aid in the uptake of a broad spectrum of molecules from the surrounding of the cell [74–77]. Other transporters identified in the genome, such as the DMT superfamily and major facilitator superfamily (MFS), enable the transport of drugs and a broad range of molecules across the membrane, including antibiotics [78]. Consistent with other Gracilibacteria, the WB Gracilibacteria expressed type IV pili systems, which can aid in the uptake of DNA from the environment [79, 80] and have been shown to be essential for host adhesion in the case of TM7i and its *Leucobacter aridicollis* host [81]. The DNA uptake is further supported by the observed strong expression of endonucleases. Consequently, WB Gracilibacteria were predicted to attach to organisms or make use of their appendages to obtain molecules. Furthermore, most WB Gracilibacteria encoded for the resistance-nodulation-division (RND) family of proteins, which also have transporter functions and efflux activity [82] but were not detected in the proteome. Two genomes (WB Gracilibacteria 25_16 and 25_26) encoded for a putative Vibrio, Colwellia, Bradyrhizobium, and Shewanella (VCBS) repeat-containing protein, a group of proteins that are proposed to contain domains with functions related to cell adhesion (InterPro entry: IPR010221). We were able to identify them in the proteomic data by means of manual search of the unclassified proteins (Supplementary Figure S8, "Adhesion" category). Translocases and peptidases identified in the proteome could aid in obtaining amino acids probably to mainly carry out protein synthesis but also other reactions, like deamination, and mobilization of molecules across the membrane. Strikingly, we identified a vast number of type II and IV secretion systems in all seven genomes, of which some subunits were also identified in the proteome (Supplementary Figure S8, "Secretion system" category). Type IV secretion systems are related to pili and their assembly subunits [83–85], while secretion systems II are used by Gram-negative bacteria to excrete proteins and other molecules that aid in signaling or microbial pathogenicity [83, 86–89]. Further evidence for the active expression of type II secretion systems in Gracilibacteria is their differential expression when comparing the most abundant proteins between the 0.2-µm filter and the 0.1-µm filter, showing that subunits for these secretion systems

are highly expressed by WB Gracilibacteria (Supplementary Figure S9). We also identified, in both genome and proteome data, subunits of the translocase Sec system, which is important for the maturation and secretion of proteins and their insertion into the inner membrane [83, 90–92]. Two proteins were annotated as transglycosylase, which are involved in cell wall interactions, such as anchoring and even cell lysis [93–95]. Added to these cell surface features, the WB Gracilibacteria also encoded for a complete peptidoglycan biosynthesis.

Publicly available Gracilibacteria genomes are missing steps in central metabolic pathways, such as glycolysis and the tricarboxylic acid cycle (TCA), but encode for the complete oxidative branch of the pentose pathway [23, 24, 28]. Likewise, the WB Gracilibacteria showed similar traits and encode for several cell surface features associated with a host-dependent lifestyle (Fig. 4). Annotations included gene functions related to transporters (e.g., ABC, protein exporters, efflux systems), secretion systems type II and IV, pilus and flagella assembly, oxidoreductases, peptidases, DNA repair, peptidoglycan assembly, and nucleotide interconversions, all of which agree to the previously described metabolic patterns encoded in cultivated and uncultivated Gracilibacteria [23, 24].

Taken together, our findings suggests a host-dependent lifestyle of WB Gracilibacteria by adhesion to hosts using pili and/or flagella with subsequently transporting biomolecules into their cell via their plethora of expressed transporters and/or pili systems. It remains unclear whether this process requires the lysis of host cells (i.e., making the Gracilbacteria predatory) or whether the transfer can occur without cell lysis (i.e., making them a form of symbiont). Once taken in, the range of enzymes (e.g., endonucleases, proteases, metalloenzymes) in the Gracilibacteria would be in charge of breaking down big molecules and incorporating smaller molecules into its metabolism.

Gracilibacteria as key members of the microbial community of geyser Wallender Born

While the functional significance of Gracilibacteria in the geyser's ecosystem is still unclear, their great abundance in metagenomics and metaproteomics demonstrates that they are key players in the geyser's community. Taking into account that cell counts for the 0.2-µm filter samples ranged between 2.07×10^4 and 5.94×10^4 cells mL^{-1} (Table S1), cell numbers that were on the low end for a shallow subsurface environment [96], and considering the regular flushing of the geyser's ecosystem (eruptions occur in 30–40 min intervals), it can be assumed that availability of molecules is low. It has been proposed that the small cell size of CPR, in this particular case of Gracilibacteria, is an advantage in oligotrophic environments since the increased surface-to-volume ratio of the cell is optimal for the uptake of nutrients [97, 98]. Additionally, the small cell size would allow for more Gracilibacteria to target the same cells [21]. Considering that most of the Gracilibacteria genomes did not co-correlate with other prokaryotes in the community, we assume that the Gracilibacteria can target a broad range of microbes to retrieve biomolecules from. This assumption and the high abundance of Gracilibacteria compared to other members of the overall geyser's microbial community (Fig. 2) indicates that the Gracilibacteria thrive in this environment, likely by means of attaching to other microbes, as opposed to scavenging in such a low-nutrient environment.

Conclusions

We present an in-depth metaproteogenomics study of a groundwater microbial community accessed through a cold-water geyser. Metagenomic abundances show that the microbial community is dominated by Nitrosomonadales, specifically by members of Gallionella spp., Campylobacter, and Hydrogenophilales, which we also detected in the metaproteomic data. Nonetheless, metaproteomics elicited that Anaerolineae, Paceibacteria, Nitrosomonadales, and in particular Gracilibacteria were the most active in terms of protein abundance. The data also supported the hypothesis that Gracilibacteria directly rely on molecular building blocks from other community members to survive, based on the expression of type II and type IV secretion systems, transporters, and defense systems. Based on these results, we propose that the WB Gracilibacteria compensate for their patchy core metabolism by a host-dependent lifestyle. While the exact type of interaction (ranging from mutualistic to predatory) cannot be determined, the enrichment of WB Gracilibacteria on 0.2-µm pore-size filters together with missing hosts in co-occurrence analyses led to the conclusion that this interaction may be unspecific, i.e., involves multiple hosts. The combination of proteomics and metagenomics of this study provide evidence for complex cellular responses of Gracilibacteria, likely enabling them to interact with other organisms at multiple levels.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-024-01889-8.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
We would like to thank Ines Pothmann for laboratory maintenance, Ken Dreger for server administration and maintenance, as well as Rashi Halder of the Luxembourg Centre for Systems Biomedicine (LCSB) for sequencing. PAFG and TB would like to thank Diana and Frank for their hospitality during the sampling campaign, as well as the Wallender Born municipality for access to sampling the geyser. We would also like to thank the Aquatic Microbiology group for allowing us the use of their ion chromatography instruments. Sebastian Grund is acknowledged for technical assistance in proteomics sample preparation and Anke Trautwein-Schult for support in metaproteomics data processing.

Authors' contributions
P.A.F.G. and T.L.V.B. carried out sampling, DNA extractions, and cell counting. T.L.V.B. performed geochemical measurements, assisted by T.H. during TOC measurements. P.A.F.G. did the metabolic and proteomic analyses, data interpretation and prepared the corresponding figures, as well as preparation of samples for protein extractions. T.L.V.B. performed bioinformatics and statistical analyses, and developed the relevant figures, as well as carried out the geochemical measurements. Tj.H., S.M., and D.B. conducted proteomic extraction and measurements, as well as data interpretation of it. J.S., C.M., S.P.E., and J.P. aided in the metagenomic analyses and binning. U.S. provided geological information. B.B. aided in discussion and insight into data interpretation. A.J.P. conceptualized the study. P.A.F.G., T.L.V.B. and A.J.P. wrote the manuscript with revisions from all co-authors.

Funding
Open Access funding enabled and organized by Projekt DEAL. This study was funded by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (Nachwuchsgruppe Dr. Alexander Probst). J.P. was supported by Aker BP within the framework of the GeneOil Project given to A.J.P. We also acknowledge support by the German Research Foundation (DFG)—CRC 1439/1: project number 426547801 and by the ERC Synergy grant "Archean Park" under 101118631. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and material
Raw sequencing data and MAGs used in this study from geyser Wallender Born have been deposited at SRA and Genbank, respectively, and are available under the BioProject PRJNA1001268 (individual accession numbers for the raw reads and genomes can be found in the Supplementary Tables S4 and S6, respectively). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [99] partner repository under the dataset identifier PXD042980.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Environmental Metagenomics, Faculty of Chemistry, Research Center One Health of the University Alliance Ruhr, University of Duisburg-Essen, 45151 Essen, Germany. [2]Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, 45141 Essen, Germany. [3]Microbial Proteomics, Institute of Microbiology, University of Greifswald, 17489 Greifswald, Germany. [4]Department of Pathogen Evolution, Helmholtz Institute for One Health, 17489 Greifswald, Germany. [5]Microbial Physiology and Molecular Biology, Institute of Microbiology, University of Greifswald, Greifswald 17489, Germany. [6]Instrumental Analytical Chemistry and Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen 45141, Germany. [7]Microbiology, The Forsyth Institute, Cambridge, MA 02142, USA. [8]Department of Geology, University of Duisburg-Essen, 45141 Essen, Germany.

Received: 18 December 2023 Accepted: 29 July 2024
Published online: 05 October 2024

References

1. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004. https://doi.org/10.1038/nbt.4229.
2. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. A new view of the tree of life. Nat Microbiol. 2016;1:16048. https://doi.org/10.1038/nmicrobiol.2016.48.
3. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42. https://doi.org/10.1038/s41564-017-0012-7.
4. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523:208–11. https://doi.org/10.1038/nature14486.
5. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun. 2015;6:6372. https://doi.org/10.1038/ncomms7372.
6. Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell. 2018;172:1181–97. https://doi.org/10.1016/j.cell.2018.02.016.
7. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science. 2012;337:1661–5. https://doi.org/10.1126/science.1224041.
8. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 2018;16:629–45. https://doi.org/10.1038/s41579-018-0076-2.
9. 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. https://doi.org/10.1038/s41467-019-12171-z.
10. Gong J, Qing Y, Guo X, Warren A. "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). Syst Appl Microbiol. 2014;37:35–41. https://doi.org/10.1016/j.syapm.2013.08.007.
11. Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome. 2017;5:112. https://doi.org/10.1186/s40168-017-0331-1.
12. Wurzbacher C, Nilsson RH, Rautio M, Peura S. Poorly known microbial taxa dominate the microbiome of permafrost thaw ponds. ISME J. 2017;11:1938–41. https://doi.org/10.1038/ismej.2017.54.
13. Starr EP, Shi S, Blazewicz SJ, Probst AJ, Herman DJ, Firestone MK, Banfield JF. Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived carbon. Microbiome. 2018;6:122. https://doi.org/10.1186/s40168-018-0499-z.
14. Geesink P, Wegner C, Probst AJ, Herrmann M, Dam HT, Kaster A, Küsel K. Genome-inferred spatio-temporal resolution of an uncultivated Roizmanbacterium reveals its ecological preferences in groundwater. Environ Microbiol. 2020;22:726–37. https://doi.org/10.1111/1462-2920.14865.
15. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W, He X. Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to mammalian hosts. Cell Rep. 2020;32:107939. https://doi.org/10.1016/j.celrep.2020.107939.

16. He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, Banfield JF. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol. 2021;6:354–65. https://doi.org/10.1038/s41564-020-00840-5.
17. Nicolas AM, Jaffe AL, Nuccio EE, Taga ME, Firestone MK, Banfield JF. Soil candidate phyla radiation bacteria encode components of aerobic metabolism and co-occur with nanoarchaea in the rare biosphere of rhizosphere grassland communities. mSystems. 2021;6:e01205-20. https://doi.org/10.1128/mSystems.01205-20.
18. Batinovic S, Rose JJ, Ratcliffe J, Seviour RJ, Petrovski S. Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams, Nat Microbiol. 2021;6:703-11. https://doi.org/10.1038/s41564-021-00892-1.
19. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, Dorrestein PC, Esquenazi E, Hunter RC, Cheng G, Nelson KE, Lux R, Shi W. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci USA. 2015;112:244–9. https://doi.org/10.1073/pnas.1419038112.
20. Bor B, Poweleit N, Bois JS, Cen L, Bedree JK, Zhou ZH, Gunsalus RP, Lux R, McLean JS, He X, Shi W. Phenotypic and physiological characterization of the epibiotic interaction between TM7x and its basibiont actinomyces. Microb Ecol. 2016;71:243–55. https://doi.org/10.1007/s00248-015-0711-7.
21. Bor B, McLean JS, Foster KR, Cen L, To TT, Serrato-Guillen A, Dewhirst FE, Shi W, He X. Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. Proc Natl Acad Sci USA. 2018;115:12277–82. https://doi.org/10.1073/pnas.1810625115.
22. Bor B, Collins AJ, Murugkar PP, Balasubramanian S, To TT, Hendrickson EL, Bedree JK, Bidlack FB, Johnston CD, Shi W, McLean JS, He X, Dewhirst FE. Insights obtained by culturing Saccharibacteria with their bacterial hosts. J Dent Res. 2020;99:685–94. https://doi.org/10.1177/0022034520905792.
23. Moreira D, Zivanovic Y, López-Archilla AI, Iniesto M, López-García P. Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. Nat Commun. 2021;12:2454. https://doi.org/10.1038/s41467-021-22762-4.
24. Yakimov MM, Merkel AY, Gaisin VA, Pilhofer M, Messina E, Hallsworth JE, Klyukina AA, Tikhonova EN, Gorlenko VM. Cultivation of a vampire: '*Candidatus* Absconditicoccus praedator'. Environ Microbiol. 2022;24:30–49. https://doi.org/10.1111/1462-2920.15823.
25. Li L, Kato C, Horikoshi K. Bacterial diversity in deep-sea sediments from different depths, Biodiversity & Conservation. 1999:659–667.
26. Hanke A, Hamann E, Sharma R, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R, Makinwa K, Hettich RL, Banfield JF, Tegetmeyer HE, Strous M. Recoding of the stop codon UGA to glycine by a BD1–5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat, Front. Microbiol. 2014:5. https://doi.org/10.3389/fmicb.2014.00231.
27. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. Stop codon reassignments in the wild. Science. 2014;344:909–13.
28. Sieber CMK, Paul BG, Castelle CJ, Hu P, Tringe SG, Valentine DL, Andersen GL, Banfield JF. Unusual metabolism and hypervariation in the genome of a gracilibacterium (BD1–5) from an oil-degrading community. mBio. 2019;10:02128–19. https://doi.org/10.1128/mBio.02128-19.
29. Meyer W. Geologie der Eifel: mit 12 Tabellen, 4, völlig neu bearb. Schweizerbart, Stuttgart: Aufl; 2013.
30. Dittrich D. Der paläogeographische Nord-Anschluss des marin beeinflussten höheren Buntsandsteins der Trier-Luxemburger Bucht, Mainzer geowissenschaftliche Mitteilungen. 2021;105–168.
31. Griesshaber E, O'Nions RK, Oxburgh ER. Helium and carbon isotope systematics in crustal fluids from the Eifel, the Rhine Graben and Black Forest, F.R.G., Chemical Geology 99. 1992;213–235. https://doi.org/10.1016/0009-2541(92)90178-8.
32. Stookey LL. Ferrozine---a new spectrophotometric reagent for iron. Anal Chem. 1970;42:779–81. https://doi.org/10.1021/ac60289a016.
33. Deusch S, Seifert J. Catching the tip of the iceberg - evaluation of sample preparation protocols for metaproteomic studies of the rumen microbiota. Proteomics. 2015;15:3590–5. https://doi.org/10.1002/pmic.201400556.
34. Bonn F, Bartel J, Büttner K, Hecker M, Otto A, Becher D. Picking vanished proteins from the void: how to collect and ship/share extremely dilute proteins in a reproducible and highly efficient manner. Anal Chem. 2014;86:7421–7. https://doi.org/10.1021/ac501189j.
35. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems. 2018;3:e00039-18. https://doi.org/10.1128/mSystems.00039-18.
36. Antipov D, Raiko M, Lapidus A, Pevzner PA. METAVIRALSPADES: assembly of viruses from metagenomic data. Bioinformatics. 2020;36:4126–9. https://doi.org/10.1093/bioinformatics/btaa490.
37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
38. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34. https://doi.org/10.1101/gr.213959.116.
39. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. https://doi.org/10.1186/1471-2105-11-119.
40. Bornemann TLV, Esser SP, Stach TL, Burg T, Probst AJ, uBin:a manual refining tool for genomes from metagenomes, Environmental Microbiology. 2023;1462–2920.16351. https://doi.org/10.1111/1462-2920.16351.
41. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res. 2006;5:2339–47. https://doi.org/10.1021/pr060161n.
42. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41:e121–e121. https://doi.org/10.1093/nar/gkt263.
43. Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2: e243. https://doi.org/10.7717/peerj.243.
44. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23:111–20. https://doi.org/10.1101/gr.142315.112.
45. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1. https://doi.org/10.1093/bioinformatics/btq461.
46. Lavy A, McGrath DG, Matheus Carnevali PB, Wan J, Dong W, Tokunaga TK, Thomas BC, Williams KH, Hubbard SS, Banfield JF. Microbial communities across a hillslope-riparian transect shaped by proximity to the stream, groundwater table, and weathered bedrock. Ecol Evol. 2019;9:6869–900. https://doi.org/10.1002/ece3.5254.
47. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinform. 2016;32:605–7. https://doi.org/10.1093/bioinformatics/btv638.
48. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol. 2018;3:836–43. https://doi.org/10.1038/s41564-018-0171-1.
49. 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;11:2864–8. https://doi.org/10.1038/ismej.2017.126.
50. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55. https://doi.org/10.1101/gr.186072.114.
51. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, Bioinformatics. 2019;btz848. https://doi.org/10.1093/bioinformatics/btz848.
52. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: an R-package for identifying proportionally abundant features using compositional data analysis. Sci Rep. 2017;7:16252. https://doi.org/10.1038/s41598-017-16520-0.
53. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol. 2015;11:e1004075. https://doi.org/10.1371/journal.pcbi.1004075.

54. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, Karaoz U, Anantharaman K. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome. 2022;10:33. https://doi.org/10.1186/s40168-021-01213-8.
55. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, Burlot L, Bussell X, Fouteau S, Gautreau G, Lajus A, Langlois J, Planel R, Roche D, Rollin J, Rouy Z, Sabatet V, Médigue C. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Res. 2019;gkz926. https://doi.org/10.1093/nar/gkz926.
56. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes - a, update. Nucleic Acids Res. 2019;48(2020):D445–53. https://doi.org/10.1093/nar/gkz862.
57. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
58. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
59. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CMK, Emerson JB, Anantharaman K, Thomas BC, Malmstrom RR, Stieglmeier M, Klingl A, Woyke T, Ryan MC, Banfield JF. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nat Microbiol. 2018;3:328–36. https://doi.org/10.1038/s41564-017-0098-y.
60. Bornemann TLV, Adam PS, Turzynski V, Schreiber U, Figueroa-Gonzalez PA, Rahlff J, Köster D, Schmidt TC, Schunk R, Krauthausen B, Probst AJ. Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing. Nat Commun. 2022;13:284. https://doi.org/10.1038/s41467-021-27783-7.
61. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114. https://doi.org/10.1038/s41467-018-07641-9.
62. Wickham H. ggplot2: ggplot2, WIREs Comp. Stat. 2011;3:180–5. https://doi.org/10.1002/wics.147.
63. Bunn A, Korpela M. Crossdating in dplR. 2013.
64. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics. 2007;8:460. https://doi.org/10.1186/1471-2105-8-460.
65. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47:W256–9. https://doi.org/10.1093/nar/gkz239.
66. Csardi G, Nepusz T. The igraph software package for complex network research, InterJournal, Complex Systems. 2006;1695.
67. Probst AJ, Elling FJ, Castelle CJ, Zhu Q, Elvert M, Birarda G, Holman H-YN, Lane KR, Ladd B, Ryan MC, Woyke T, Hinrichs K-U, Banfield JF. Lipid analysis of CO2-rich subsurface aquifers suggests an autotrophy-based deep biosphere with lysolipids enriched in CPR bacteria. ISME J. 2020;14:1547–60. https://doi.org/10.1038/s41396-020-0624-4.
68. Schwank K, Bornemann TLV, Dombrowski N, Spang A, Banfield JF, Probst AJ. An archaeal symbiont-host association from the deep terrestrial subsurface. Microbiology. 2018. https://doi.org/10.1101/486977.
69. Guerrero R, Pedrós-Alió C, Esteve I, Mas J, Chase D, Margulis L. Predatory prokaryotes: predation and primary consumption evolved in bacteria. Proc Natl Acad Sci USA. 1986;83:2138–42. https://doi.org/10.1073/pnas.83.7.2138.
70. Puente-Sánchez F, Arce-Rodríguez A, Oggerin M, García-Villadangos M, Moreno-Paz M, Blanco Y, Rodríguez N, Bird L, Lincoln SA, Tornos F, Prieto-Ballesteros O, Freeman KH, Pieper DH, Timmis KN, Amils R, Parro V. Viable cyanobacteria in the deep continental subsurface. Proc Natl Acad Sci USA. 2018;115:10702–7. https://doi.org/10.1073/pnas.1808176115.
71. Moore CJ, Bornemann TLV, Figueroa-Gonzalez PA, Esser SP, Moraru C, Soares AR, Hinzke T, Trautwein-Schult A, Maaß S, Becher D, Starke J, Plewka J, Rothe L, Probst AJ. Time-series metaproteogenomics of a high-CO2 aquifer reveals active viruses with fluctuating abundances and broad host ranges, microLife. 2024;uqae011. https://doi.org/10.1093/femsml/uqae011.
72. Anisiu MC. Lotka, Volterra and their model. Didáctica mathematica. 2014;32:9–17. https://www.math.ubbcluj.ro/~didactica/pdfs/2014/didmath2014-02.pdf.
73. Hove-Jensen B, Andersen KR, Kilstrup M, Martinussen J, Switzer RL, Willemoës M. Phosphoribosyl diphosphate (PRPP): biosynthesis, enzymology, utilization, and metabolic significance. Microbiol Mol Biol Rev. 2017;81:e00040-e116. https://doi.org/10.1128/MMBR.00040-16.
74. Hosie AHF, Poole PS. Bacterial ABC transporters of amino acids. Res Microbiol. 2001;152:259–70. https://doi.org/10.1016/S0923-2508(01)01197-4.
75. Lewis VG, Ween MP, McDevitt CA. The role of ATP-binding cassette transporters in bacterial pathogenicity. Protoplasma. 2012;249:919–42. https://doi.org/10.1007/s00709-011-0360-8.
76. Tanaka KJ, Song S, Mason K, Pinkett HW. Selective substrate uptake: the role of ATP-binding cassette (ABC) importers in pathogenesis. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2018;1860:868–77. https://doi.org/10.1016/j.bbamem.2017.08.011.
77. Orelle C, Mathieu K, Jault J-M. Multidrug ABC transporters in bacteria. Res Microbiol. 2019;170:381–91. https://doi.org/10.1016/j.resmic.2019.06.001.
78. Quistgaard EM, Löw C, Guettou F, Nordlund P. Understanding transport by the major facilitator superfamily (MFS): structures pave the way. Nat Rev Mol Cell Biol. 2016;17:123–32. https://doi.org/10.1038/nrm.2015.25.
79. Ellison CK, Dalia TN, Vidal Ceballos A, Wang JCY, Biais N, Brun YV, Dalia AB. Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae. Nat Microbiol. 2018;3:773–80. https://doi.org/10.1038/s41564-018-0174-y.
80. Piepenbrink KH. DNA uptake by type IV filaments. Front Mol Biosci. 2019;6:1. https://doi.org/10.3389/fmolb.2019.00001.
81. Xie B, Wang J, Nie Y, Tian J, Wang Z, Chen D, Hu B, Wu X-L, Du W. Type IV pili trigger episymbiotic association of Saccharibacteria with its bacterial host. Proc Natl Acad Sci. 2022;119: e2215990119. https://doi.org/10.1073/pnas.2215990119.
82. Fernando D, Kumar A. Resistance-nodulation-division multidrug efflux pumps in gram-negative bacteria: role in virulence. Antibiotics. 2013;2:163–81. https://doi.org/10.3390/antibiotics2010163.
83. Green ER, Mecsas J. Bacterial secretion systems: an overview. Microbiol Spectr. 4 (2016) 4.1.13. https://doi.org/10.1128/microbiolspec.VMBF-0012-2015.
84. Craig L, Forest KT, Maier B. Type IV pili: dynamics, biophysics and functional consequences. Nat Rev Microbiol. 2019;17:429–40. https://doi.org/10.1038/s41579-019-0195-4.
85. Jacobsen T, Bardiaux B, Francetic O, Izadi-Pruneyre N, Nilges M. Structure and function of minor pilins of type IV pili. Med Microbiol Immunol. 2020;209:301–8. https://doi.org/10.1007/s00430-019-00642-5.
86. Sandkvist M. Type II secretion and pathogenesis. Infect Immun. 2001;69:3523–35. https://doi.org/10.1128/IAI.69.6.3523-3535.2001.
87. Cianciotto NP. Type II secretion: a protein secretion system for all seasons. Trends Microbiol. 2005;13:581–8. https://doi.org/10.1016/j.tim.2005.09.005.
88. Johnson TL, Abendroth J, Hol WGJ, Sandkvist M. Type II secretion: from structure to function. FEMS Microbiol Lett. 2006;255:175–86. https://doi.org/10.1111/j.1574-6968.2006.00102.x.
89. K.V. Korotkov, M. Sandkvist, Architecture, function, and substrates of the type II secretion system, EcoSal Plus 8 (2019) ecosalplus.ESP-0034–2018. https://doi.org/10.1128/ecosalplus.ESP-0034-2018.
90. De Geyter J, Smets D, Karamanou S, Economou A. Inner membrane translocases and insertases. Bacterial Cell Walls and Membranes. 2019;92:337–66. https://doi.org/10.1007/978-3-030-18768-2_11.
91. Forster BM, Marquis H. Protein transport across the cell wall of monoderm Gram-positive bacteria: protein transport across the bacterial cell wall. Mol Microbiol. 2012;84:405–13. https://doi.org/10.1111/j.1365-2958.2012.08040.x.
92. J.A. Lycklama a Nijeholt, A.J.M. Driessen, The bacterial Sec-translocase: structure and mechanism, Phil. Trans. R. Soc. B 367 (2012) 1016–1028. https://doi.org/10.1098/rstb.2011.0201.
93. Scheurwater E, Reid CW, Clarke AJ. Lytic transglycosylases: bacterial space-making autolysins. Int J Biochem Cell Biol. 2008;40:586–91. https://doi.org/10.1016/j.biocel.2007.03.018.
94. Koraimann G. Lytic transglycosylases in macromolecular transport systems of Gram-negative bacteria. Cellular and Molecular Life Sciences (CMLS). 2003;60:2371–88. https://doi.org/10.1007/s00018-003-3056-1.

95. Zahrl D, Wagner M, Bischof K, Bayer M, Zavecz B, Beranek A, Ruckenstuhl C, Zarfel GE, Koraimann G. Peptidoglycan degradation by specialized lytic transglycosylases associated with type III and type IV secretion systems. Microbiology. 2005;151:3455–67. https://doi.org/10.1099/mic.0.28141-0.
96. Magnabosco C, Lin L-H, Dong H, Bomberg M, Ghiorse W, Stan-Lotter H, Pedersen K, Kieft TL, van Heerden E, Onstott TC. The biomass and biodiversity of the continental subsurface. Nature Geosci. 2018;11:707–17. https://doi.org/10.1038/s41561-018-0221-6.
97. Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, Carlson CA, Smith RD, Giovanonni SJ. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J. 2009;3:93–105. https://doi.org/10.1038/ismej.2008.83.
98. M. Herrmann, C.-E. Wegner, M. Taubert, P. Geesink, K. Lehmann, L. Yan, R. Lehmann, K.U. Totsche, K. Küsel, Predominance of Cand. Patescibacteria in groundwater is caused by their preferential mobilization from soils and flourishing under oligotrophic conditions, Front. Microbiol. 10 (2019) 1407. https://doi.org/10.3389/fmicb.2019.01407.
99. PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences | Nucleic Acids Research | Oxford Academic, (n.d.). https://academic.oup.com/nar/article/50/D1/D543/6415112 (accessed September 20, 2023).

Publisher's Note