Metaproteogenomics resolution of a high-CO_2 aquifer community suggests an active symbiotic lifestyle of groundwater Gracilibacteria

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

[1]Environmental Metagenomics, Research Center One Health of the University Alliance Ruhr, Faculty of Chemistry, 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]Instrumental Analytical Chemistry and Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen,45141 Essen, Germany
[6]Microbiology, The Forsyth Institute, Cambridge 02142, Massachusetts, USA
[7]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston 02115, Massachusetts, USA
[8]Department of Geology, University of Duisburg-Essen, 45141 Essen, Germany

*authors contributed equally
#corresponding author: alexander.probst@uni-due.de

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. However, besides genomic evidence, little is known about the lifestyle of Gracilibacteria, their temporal dynamics, and activity in natural ecosystems, particularly in groundwater, where they were initially been genomically resolved. Therefore, we here aimed to investigate Gracilibacteria activity *in situ* and to discern expressed genes involved in their lifestyle, 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 670 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 Galionellales 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 encoded for a high number of proteins related to a symbiotic or even predatory lifestyle, *e.g.*, type IV and type II secretion system subunits and features related to cell-cell interactions and cell motility, which were also detected on protein level.

Conclusions. Coupling metagenomics to metaproteomics enabled us to identify microbial keystone taxa in a high-CO_2 aquifer, and to reveal microbial dynamics of Gracilibacteria. We posit that Gracilibacteria might be successful microbial predators in this ecosystem, potentially aiding in population control of this highly perturbed microbial geyser community from the deep biosphere.

Keywords: CPR bacteria, BD1-5, groundwater, genome-resolved metagenomics, proteomics, symbiotic bacteria, predatory bacteria

Background

Bacteria belonging to the Candidate Phyla Radiation (CPR, also known as Patescibacteria [1]) comprise at least one quarter of the bacterial domain [2,3], and are characterized by having reduced genomes as well as small cell sizes (<0.2 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. Interestingly, 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 environmental surveys, CPR bacteria have been detected in a

73 plethora of environments, including hydrothermal vents, soil, lakes, rivers, permafrost, the
74 (deep) terrestrial subsurface and groundwater, as well as oral cavities of mammals [10–18].
75 Considering the limited metabolic potential and small genome size of CPR bacteria, which is
76 comparable to symbiotic non-CPR bacteria [8], CPR bacteria likely depend on hosts in order
77 to obtain essential molecular building blocks, such as amino acids, lipids, or co-factors [6,8].
78 This host-dependency was first shown to be the case for the phylum Saccharibacteria, and in
79 particular for *Nanosynbacter lyticus* (also known as TM7x), which was isolated in co-culture
80 with its host, *Schaalia odontolyticus* XH001, from the human oral cavity [15,19]. While *N.*
81 *lyticus* can, however, achieve stable growth conditions along with its host [20–22], recent
82 reports of predatory CPR bacteria, specifically of the Absconditabacteria lineage, show that
83 the host is killed for the CPR bacteria to survive, regardless of the environmental conditions
84 [23,24].
85 Gracilibacteria, originally named BD1-5 [7,25], is a relatively understudied phylum of CPR
86 bacteria and forms a clade with the candidate phyla Abscondibacteria and Peregrinibacteria
87 [23]. A trademark characteristic of the Gracilibacteria lineage, similar to Absconditabacteria,
88 is their use of genetic code 25, where the codon UGA encodes for a glycine, as opposed to
89 being a stop codon [26–28]. Like other members of CPR bacteria, Gracilibacteria have very
90 limited metabolic capabilities (*e.g.*, some members encode for the oxidative branch of the
91 pentose pathway but lack steps for glycolysis and TCA). At the same time, Gracilibacteria
92 encode for several surface proteins related to type IV pili and secretion systems [28],
93 indicating that these organisms have the potential for cell-cell interactions.
94 One such example of predatory CPR bacteria is *Candidatus* Absconditicoccus praedator,
95 isolated from a hypersaline alkaline lake along with its host *Halorhodospira halophila*, an
96 obligately anaerobic photosynthetic purple-sulphur bacterium [24]. The symbiont's metabolic
97 limitations make it completely dependent on its host, whom it kills by lysing and consuming
98 its cytoplasmic contents. A second case of a predatory CPR bacterium was reported for
99 *Candidatus* Vampirococcus lugosii, discovered in an athalassic hypersaline lake in Spain.
100 Cultures derived from the lake's microbial mat showed that *Ca.* V. lugosii is an epibiotic
101 bacterium that feeds on *Halochromatium*, an anoxygenic photosynthetic
102 gammaproteobacterium [23]. Both of these CPR bacteria are members of the
103 Absconditabacteria phylum. Similar to *Ca.* A. praedator, the attachment of *Ca*. V. lugosii leads
104 to compromising the cell membrane of its host, so that *Ca*. V. lugosii can consume the

105 *Halochromatium* bacterium's cytoplasmic content, thereby killing it [23,24]. Based on *Ca.* A.
106 praedator and *Ca*. V. lugosii's predatory lifestyle, it was proposed that they might play a role
107 in population control of their hosts. As there is as of yet no cultured Gracilibacteria
108 representative, no information about their host range nor predatory character is available. In
109 addition, while metabolic potential can be inferred based on genomes, the only proteomic
110 report about Gracilibacteria focuses on their alternative genetic code and not on their protein
111 expression, *i.e.*, actual phenotype, as for example related to metabolism or cell-cell
112 interactions [7,26]. Consequently, whether they actually live as symbionts as predicted from
113 their genomes, has remained elusive.
114 In this study, we characterized a model ecosystem (cold-water geyser Wallender Born (WB;
115 50°09'13.4"N 6°43'13.1"E)) enriched in Gracilibacteria, to investigate proteogenome of
116 Gracilibacteria in a complex microbial community. The geyser is located in the town of
117 Wallenborn, in the Volcanic Eifel region of Germany. The driving force for its eruptions is the
118 accumulation of carbon dioxide (CO_2) in groundwater. CO_2 saturates the groundwater, which
119 leads to the geyser's eruption where groundwater and CO_2 are released to the atmosphere.
120 These eruptions occur in approximate 40 minute intervals. Taking advantage of these
121 groundwater eruptions, we used geyser WB as a study site to characterize its microbial
122 community on the molecular level. For this, we sampled erupted groundwater over a twelve-
123 day time series and sequentially filtered it to enrich small- and large-sized microbes on 0.1-
124 µm filters and 0.2-µm filters, respectively. We used genome-resolved metagenomics to obtain
125 high-quality genomes, and coupled this to metaproteomics analyses of the same time points,
126 to generate in-depth insights into Gracilibacteria metabolism and ecology. We identified traits
127 related to a predatory lifestyle of Gracilibacteria in the geyser, both in their metabolic
128 potential as well as in expressed proteins with metabolic similarities with *Ca.* A. praedator
129 and *Ca*. V. lugosii. Ultimately, we discerned a predatory lifestyle of groundwater
130 Gracilibacteria, hinting at their role as key members of the geyser's microbial community.

131 Material and methods

132 ***Geological context of sampling site***

133 The cold-water geyser Wallender Born (WB) is located in the town Wallenborn, southwest of
134 the subrecently active Quaternary volcanic field of the West Eifel region, in the Rhenish Massif

135 (50°09’13.4”N 6°43’13.1”E). It is located between the Manderscheid anticline to the
136 southeast and the Salmtal syncline to the northwest which formed during the Variscan
137 orogeny event. It consists of weakly metamorphic early Devonian (Emsian) sedimentary units
138 [29]. From a geological point of view, it belongs to the eastern edge of the old depression
139 zone of the Eifel north-south zone (Eifeler Nord-Südzone). This zone is characterized both by
140 the Middle and Late Devonian Eifel limestone synclines, preserved in the depression, and by
141 numerous N-S to NNE-SSW trending fault zones [30]. The strong CO_2 outgassing is related to
142 the magmatic activity of the West Eifel region, which together with the East Eifel volcanic field
143 is attributed to a mantle plume underneath the Eifel (Eifel plume). Helium R/Ra values from
144 the Wallender Born geyser show a clear mantle influx of ^3He, implying magmatic intrusion
145 [31]. Please refer to Supplementary Figure S1 for a scheme of the geyser.
146
147 ***Groundwater sampling and geochemical analyses***
148 Time-series sampling of biomass was done for twelve days, from the 14th – 25th of October
149 2020. Groundwater was collected in sterile, DNA-free containers during each eruption, for
150 immediate filtering on-site. The groundwater was filtered sequentially on 0.2-µm and then
151 0.1-µm filters (see Supplementary Figure S2A for a scheme of the filtering setup). Before
152 starting the filtration of each time-point, the system was rinsed with groundwater to get rid
153 of stagnant water left over from the previous sampled time-point. Additionally, one bulk
154 sample, filtered directly onto a 0.1-µm filter, was obtained. Approximately 100 L were filtered
155 on each 0.2-µm filter, with the flow-throughs of two 0.2-µm filters being combined via a T-
156 piece and filtered onto one 0.1-µm filter (please refer to Supplementary Table S1 for
157 individual filtered volumes). Filters were then immediately transferred to a sterile falcon tube,
158 stored on dry ice on-site, and later on transferred to -80° C until further processing.
159 Temperature and pH were measured in triplicates with a thermometer and pH strips,
160 respectively.
161 *Total iron.* 100 µL of groundwater were mixed with 900 µL of 1 M HCl on-site. Samples were
162 taken in triplicates and kept at 4°C till measurement. Total iron was measured by mixing of
163 100 µl of aliquot with 900 µl hydroxylamine, followed by addition of 80 µl reduced aliquot to
164 120 µl Ferrozine, incubation for 10 min in the dark, and photometric determination by a Tecan
165 plate reader at 560 nm [32].

166 *Ions*. 0.1-µm pore-size filtered water was used for ion measurements. Samples were
167 measured in dilutions of 1:100 and 1:500 in distilled water to be able to accurately quantify
168 both highly abundant ions and lower abundant ions. Samples were analyzed using the Dionex
169 Aquion ion chromatography system (Thermo Scientific, USA). Anions were analyzed with a
170 Dionex IonPac AG23-4 µm guard column, a Dionex IonPac AG23-4 µm 2 x 250 mm analytical
171 column, as well as an AERS 500 Carbonate 2 mm suppressor, Ultimate 3000 heating element,
172 and a DS6 heated conductivity cell detector. Cations were measured on a CS12A RFIC 2 x 250
173 mm analytical column and a CG12A RFIC detector. Measurements were done in technical
174 triplicates.
175 *Total organic carbon (TOC)*. Unfiltered groundwater samples were stored at 4 °C until
176 measurement. Prior to measurement, the samples were filtered through a 0.45-µm filter to
177 remove large particles. 10 ml of sample were augmented with 0.5 ml 1 M HCl to remove
178 residual carbonate present in cold-water geysers, verified via pH strips. For each time point,
179 two biological samples were measured in triplicate on a TOC-5050 Analyzer with ASL-5000
180 (Shimadzu) and averaged after accounting for the HCl dilution.
181
182 ***Epifluorescence microscopy***
183 For each time-point, groundwater was filtered on-site through a 0.2-µm syringe-filter using a
184 sterile syringe. The flow-through was collected in a sterile, DNA-free tube and subsequently
185 filtered through a 0.1-µm syringe-filter using a second sterile syringe. Filters were incubated
186 with 10 µg ml^{-1} DAPI in 2% (v/v) formaldehyde for five minutes, washed with distilled water,
187 dried for ten minutes and stored in the dark at 4 °C until use. Cells were quantified using the
188 Axio Imager M2m epifluorescence microscope equipped with an AxioCam MRm and a Zen 2
189 Pro software (Carl Zeiss Microscopy GmbH, Jena, Germany; version 2.0.0.0). Enumeration of
190 the cells was done in ten horizontal and ten vertical fields, extrapolating counts to the entire
191 filter area and normalizing through the filtered volume to calculate cells per millilitre.
192
193 ***DNA extraction and sequencing***
194 Each pair of 0.2-µm filters was pooled to have a comparable sample to the respective 0.1-µm
195 filter (see Supplementary Figure S2B for a scheme of the pooling for metagenomics). This
196 pooling resulted in 22 samples for the 0.2-µm filtered fractions, each corresponding to the
197 different time points obtained during the sampling campaign, with their respective 0.1-µm

198 filtered fraction. DNA extractions of the samples were performed using the Dneasy PowerMax
199 Soil DNA Extraction kit (Qiagen, 12988-10) according to the manufacturer's instructions and
200 further concentrated using ethanol precipitation with glycogen as the carrier. Of the 44
201 samples, 30 yielded enough biomass for library preparation via the Westburg NGS DNA
202 Library Prep Kit (cat. No. WB 9096). Of these, 28 were successfully sequenced using Illumina
203 NextSeq500 (2x150 bp paired-end reads). Some libraries which yielded low sequencing depth
204 in initial sequencing runs were sequenced up to a total of three times and the reads were
205 concatenated to reach about 20 Gbps sequencing depth per sample.
206
207 ***Sample preparation for metaproteomics***
208 Due to the higher biomass requirements for metaproteomics as compared to metagenomics,
209 sample sets analyzed differed from those for metagenomics. A total of three different
210 metaproteomics sample sets were analyzed (see Supplementary Figure S2C and
211 Supplementary Table S3 for details): I) We compiled a time-series consisting of individual 0.2-
212 µm filters taken at different time points. II) To compare the proteome of the size-fractionated
213 microbiomes, we separately pooled 0.2-µm filters and 0.1-µm filters of the respective time
214 points to generate biomass consistent of only microbes from the 0.2-µm fraction and of the
215 0.1-µm fraction, respectively. This provided sufficient biomass for the 0.1-µm filter for most
216 protein extraction but also generated one set of 0.2-µm filter samples without the
217 corresponding 0.1-µm sample due to too-low biomass. III) For a whole-community, we used
218 a bulk sample filtered directly onto a 0.1-µm filter (without size fractionation).
219
220 ***Protein extraction***
221 Protein extraction from the filters was performed according to previously published protocols
222 [33] with some modifications. In brief, the filters were cut in pieces, transferred to low protein
223 binding reaction tubes and covered with one volume of resuspension solution 1 (50 mM Tris-
224 HCl pH 7.5, 0.1 mg ml^{-1} chloramphenicol, 1 mM PMSF (phenylmethanesulfonyl fluoride)).
225 After vortexing, 1.5 volumes resuspension solution 2 (20 mM Tris-HCl pH 7.5, 2% (w/v) SDS)
226 were added, and the samples were incubated at 60 °C for 10 min with vigorous shaking. After
227 cooling down of the samples to room temperature, 5x volume of DNAse solution (1 µg ml^{-1}
228 DNAse I) were added. Samples were lysed by ultrasonication on ice for 6 min (Sonopuls MS72
229 (Bandelin, Berlin, Germany), amplitude 51-60%; cycle 0.5), and subsequently incubated at 37

230 °C for 10 min with vigorous shaking. After centrifugation, proteins were precipitated from the
231 supernatants by adding pre-cooled trichloroacetic acid (TCA) to reach 20% (v/v) TCA and
232 incubating at 4 °C for 30 min in an overhead inverter. Precipitates were pelleted and washed
233 with pre-cooled acetone. The dried pellet was resuspended in 2x SDS sampling solution (0.125
234 M Tris-HCl pH 6.8, 4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) ß-mercaptoethanol), and
235 incubated in an ultrasonic bath for 15 min before heating at 95 °C for 5 min. Samples were
236 centrifuged, supernatants were kept and the remaining pellet was again treated with 2x SDS
237 sampling solution. Both sample fractions, supernatant and pellet, were loaded onto pre-cast
238 SDS Gels (Criterion TGX 4-20%, 12+2 wells, Bio-Rad) and separated by SDS-PAGE. After
239 staining with Coomassie, proteins were in-gel digested as described previously [34]. Briefly,
240 gel lanes were excised in ten equidistant pieces and each piece was destained, washed, and
241 subjected to trypsin digestion (as well as mass spectrometric measurement) individually.
242 Peptides were eluted in water using an ultrasonic bath and desalted by the use of C18 ZipTip
243 columns (Merck) according to the manufacturer's guidelines.
244
245 ***Metagenome processing***
246 Illumina paired-end reads were checked for adapters and sequencing artefacts with Bbduk
247 v37.09 (Bushnell, https://sourceforge.net/projects/bbtools), quality-trimmed with Sickle
248 v1.33 (https://github.com/najoshi/sickle) and deduplicated with dedup (Bushnell,
249 https://sourceforge.net/projects/bbtools). The diversity coverage in the samples based on
250 raw reads was estimated using Nonpareil3 [35] in kmer mode with k=20 (*i.e.*, the minimal
251 read length). For each individual metagenome, quality-controlled reads were first assembled
252 with Metaviralspades 3.15.2 [36] to reconstruct viruses and plasmids, and paired reads where
253 no read mate mapped to assembled viruses or plasmids (via Bowtie2 v2.3.5.1, –sensitive
254 mode, [37]) were then assembled using metaspades 3.15.2 [38]. Viral and normal assemblies
255 were then concatenated to yield the final assembly. Unless otherwise stated, all mappings
256 were performed with Bowtie2 [37] in sensitive mode for, *e.g.*, determining relative scaffold
257 abundances of the assembled scaffolds. Open reading frames were predicted on scaffolds
258 with a minimum length of 1000 bps using Prodigal 2.6.3 [39] in meta-mode and annotated
259 against FunTaxDB 1.3 [40] to retrieve function and taxonomic profiles.
260
261 ***Mass spectrometry and data processing of metaproteomics samples***

262 Peptides were subjected to LC-MS/MS analyses on a LTQ Orbitrap Elite instrument
263 (ThermoFisher Scientific) coupled to an EASY-nLC 1200 liquid chromatography system.
264 Therefore, peptides were loaded on a self-packed analytical column (OD 360 µm, ID 100 µm,
265 length 20 cm) filled with 3 µm diameter C18 particles (Maisch) and eluted by a binary
266 nonlinear gradient starting at 5% up to 99% acetonitrile in 0.1% (v/v) acetic acid over 82 min
267 with a flow rate of 300 nL min^{-1}. For MS analysis, a full scan in the Orbitrap with a resolution
268 of 60,000 was followed by collision-induced dissociation (CID) of the 20 most abundant
269 precursor ions. Database searches against a concatenated forward-reverse protein database
270 derived from the described metagenome analyses (see "Generation of protein database from
271 metagenomes") (15,269,672 unique entries) was performed using Mascot (Matrix Science;
272 version 2.7.0.1), assuming the digestion enzyme trypsin and applying the following
273 parameters: fragment ion mass tolerance 0.5 Da and parent ion tolerance of 10 ppm, up to
274 two missed cleavages as well as methionine oxidation as a variable modification. Scaffold
275 v5.1.2 (Proteome Software Inc.) was used to merge the search results and to validate MS/MS
276 based peptide and protein identifications. During creation of the Scaffold file, an additional
277 X! Tandem search with default settings was performed for validation. Protein and peptide-
278 level false discovery rate (FDR) were set at 5 % and protein groups required to contain at least
279 1 unique peptide. Protein groups consisted of proteins which were not distinguishable based
280 on identified peptides.
281 For proteomic comparison across samples, the spectral counts were normalized by the
282 protein length, followed by total sample spectral count, as described previously [41].
283
284 ***Community profiling based on rpS3 marker gene***
285 Genes encoding for the *rpS3* marker gene were identified via hmmsearch [42] against the
286 Phylosift HMM model [43]; PhyloSift Marker: DNGNGWU00028) at an E-value of 1e-28 as well
287 as by pulling out genes annotated as *rpS3* during data processing (see above). All identified
288 *rpS3* sequences throughout all assemblies were then pooled and clustered at 99% similarity
289 as described in Sharon et al. [44] using the usearch -uclust module [45]. Mapping on just the
290 *rpS3* gene would result in inadequate abundance estimates, as reads mapping to the edges
291 of sequences map poorly, causing coverage to drop off at the edges. Hence, for each *rpS3*
292 gene, the scaffold area around the gene was taken into account by extending the scaffold
293 sequences around identified rpS3 genes by up to 1000 bp to either side [46]. The workflow

294 and scripts to identify, cluster, and extract rpS3 sequences as well as calculate their coverage
295 and breadth are provided at: https://github.com/ProbstLab/publication-
296 associated_scripts/tree/main/Figueroa-Gonzalez_Bornemann_etal. This yielded scaffold
297 fragments of about 3500 bp in length, which are expected to result in more accurate
298 abundance estimates than just mapping on the gene by itself. Representative *rpS3* sequences
299 were selected from each *rpS3* cluster by order of the following priorities: 1) centroid and can
300 be extended by 1000 bp into both sides of the gene, 2) not centroid and can be extended into
301 both sides by 1000 bp, and 3) longest scaffold region. The resulting *rpS3* sequences, ideally
302 extended in both directions by 1000 bp, are henceforth called rpS3extended. The taxonomy
303 of the rpS3extended sequences was determined by using usearch (ublast) against the pooled
304 archaeal and bacterial *rpS3* sequences of the GTDB (E-value 0.00001). The best hits from the
305 search were taken, and those rpS3extended sequences without a hit at this threshold were
306 assigned as "Unclassified". The abundance of the sequences in each sample was determined
307 by mapping of reads back to the rpS3extended sequences and their average coverage per
308 sample was determined. Additionally, the breadth, *i.e.,* the number of nucleotide positions
309 having a coverage of at least 1, was determined. The coverage of an rpS3extended sequence
310 in a given sample was set to zero if the breadth of the sequence in that sample was less than
311 95%. This was done to ensure that the respective sequences were present in their entirety in
312 the respective sample, instead of just fragments of them. After breadth correction, coverages
313 per sample were normalized by sequencing depth.
314
315 ***Reconstruction of metagenome-assembled genomes***
316 Reads of all metagenomes were mapped against each assembly to yield differential coverage
317 data used in binning. Metagenome-assembled genomes (MAGs) were reconstructed using
318 ABAWACA v1.0.0 (https://github.com/CK7/abawaca), CONCOCT v1.1.0
319 (https://github.com/BinPro/CONCOCT), and MaxBin 2.0 v2.2.7 [47]. The bin sets were
320 consolidated with DAS tool v1.1.2 [48] with default parameters. The consolidated bin sets
321 were curated in uBin v0.9.20 [40] and all curated MAGs of all samples were then pooled and
322 dereplicated with dRep[49] at 99% gANI to yield all unique strains present in the dataset.
323 CheckM1 [50] was used to assess and filter genomes based on completeness and
324 contamination, using 70% and 10% as the cutoffs, respectively.
325

326 ***Taxonomic classification of MAGs***

327 GTDB-tk v.2.1 [51] with the classify_wf workflow against GTDBr207 was used to determine
328 the taxonomic origin of the recovered MAGs. Genes on these genomes were predicted with
329 prodigal in normal mode, either with default genetic codes (for non-Gracilibacteria) or codon
330 code 25 (for Gracilibacteria).

331

332 ***Generation of protein database from metagenomes***

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

343

344 ***Tracking abundances of rpS3 genes and MAGs across samples***

345 Each metagenome was mapped to the dereplicated MAGs or rpS3extended sequences to
346 estimate their abundances across time and filter fractions. Additionally, the breadth of each
347 MAG and rpS3extended sequence in each sample was determined. The breadth is herein
348 defined as the percentage of the sequence length having a coverage of at least 1. This was
349 used to filter out MAGs and rpS3extended sequences that only had coverage because a small
350 proportion of their sequence was shared with other organisms. For genomes, the minimum
351 required breadth to have a valid coverage was set to 30%, for rpS3extended sequences to
352 95%. If the breadth of a sequence in a sample was below this threshold, the coverage was set
353 to zero.

354

355 ***Normalization of abundances of MAGs and rpS3 genes***

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

358 $NormMAGi = MAG * (\max(SeqDepth)/SeqDepthi)$

359 $NormMAG_i$ = Normalized MAG (or rpS3extended) abundance in sample i

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

361 $SeqDepth_i$ = Sequencing depth in sample i

362

363 ***Shannon and Simpson diversity estimates of assemblies***

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

367

368 ***Rank abundance estimates of phyla***

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

372

373 ***Non-Metric Multidimensional Scaling (NMDS)***

374 NMDS, based on Bray-Curtis dissimilarities, were calculated using the *metaMDS* function of
375 the *vegan* package, using species tables of either MAGs or rps3extended sequences. All
376 samples were analyzed while being pooled together as one database, as well as split into the
377 two filter fractions. Only NMDS plots with a convergent solution (after a maximum of 400
378 tries) as well as sufficiently low stress are shown.

379

380 ***Proportionality networks***

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

387

388 ***Metabolic analyses***

389 Metabolic potential of the microbial community was determined using METABOLIC [54].
390 Genome annotation of Gracilibacteria was done using the Genoscope platform MAGE [55].
391 In-depth metabolic potential was determined by combining the annotation information from
392 MAGE with information from MetaCyc [56], KEGG [57], and UniProt [58] (see Table S7 for
393 automatic annotation results). Furthermore, unclassified proteins were manually searched
394 against protein domain databases of NCBI protein domains
395 (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and AlphaFold-EMBL database
396 [(https://www.ebi.ac.uk/Tools/sss/fasta/), see Table S8 for manual annotation results). The
397 information obtained in the analyses of automatic annotation was compared with the
398 metaproteomics data to evidence expressed proteins.
399
400 ***Comparison of available microbial populations from cold-water geysers***
401 All dereplicated MAGs from Crystal Geyser (Utah, USA) [59], Geyser Andernach (Andernach,
402 Germany) [60] and the Geyser Wallender Born (Wallenborn, Germany) were pooled and
403 compared via FastANI[61] with a >= 50% coverage.
404
405 ***Visualization***
406 Figure panels, including the NMDS, were generated with *ggplot2* [62] in R [63], and
407 subsequently assembled and post-processed in AffinityDesigner V2
408 (https://affinity.serif.com/). The phylogenetic trees produced by GTDB-tk were visualized
409 using Dendroscope v3.5.10 [64] and processed with iTol v6 [65]. The proportionality networks
410 were generated using the *iGraph* package [66]. The number of shared genomes between the
411 cold-water geysers were analyzed using *ggupset (*https://github.com/const-ae/ggupset). The
412 metabolic reconstruction scheme was generated manually in Affinity Designer V2.
413

414 # Results and discussion

415 ***Microbial community of the geyser is dominated by Gallionella and Gracilibacteria***
416 Size fractionation of 22 groundwater samples discharged from Wallender Born resulted in 15
417 samples on 0.2-µm filters, 14 samples on 0.1-µm filters, and one bulk 0.1-µm filter sample.
418 The subsequent metagenomic sequencing resulted in 725 Tbps of Illumina NextSeq data.

419 These data covered 78% of the microbial diversity (on average) in the samples based on
420 Nonpareil3. After individually assembling the metagenomes, we determined that about
421 80.0% (+-3.2% STDEV) of the reads mapped back to the scaffolds, with only slight differences
422 between 0.1-µm and 0.2-µm filter fractions (78.2% +-2.9% STDEV and 82.6% +-1.9% STDEV,
423 respectively; see Supplementary Table S4 for reads, assembly and Nonpareil3 statistics). We
424 hence conclude that the assemblies are representative of the microbial community of the
425 geyser. The Shannon diversity index based on *rpS3* marker genes was on average 3.22 for 0.1-
426 µm size fractions, and 3.50 for 0.2-µm size fractions, while the Simpson diversity index was
427 on average 0.90 for both size fractions with similar intra-sample variation (0.42 to 0.11, and
428 0.30 to 0.01, respectively; Figure 1A). Based on *rpS3* abundances (Figure 1B), the communities
429 were dominated by Proteobacteria (58.0%) followed by Patescibacteria (26.2%) and
430 Bacteroidota (4.57%; please refer to Supplementary Figure S3 for the complete rank
431 abundances and Supplementary Figure S4 for phylum-level taxa abundances across time-
432 series samples). NMDS analyses revealed that the 0.1-µm filter fractions showed a great intra-
433 group dissimilarity, and that 0.2-µm samples and the bulk sample clustering in close proximity
434 (Figure 1C). Individual NMDS analyses of 0.1-µm and 0.2-µm filter fractions confirmed this
435 finding, with the NMDS of the 0.1-µm samples being much more dissimilar compared to their
436 0.2-µm counterparts (Figure 1C).

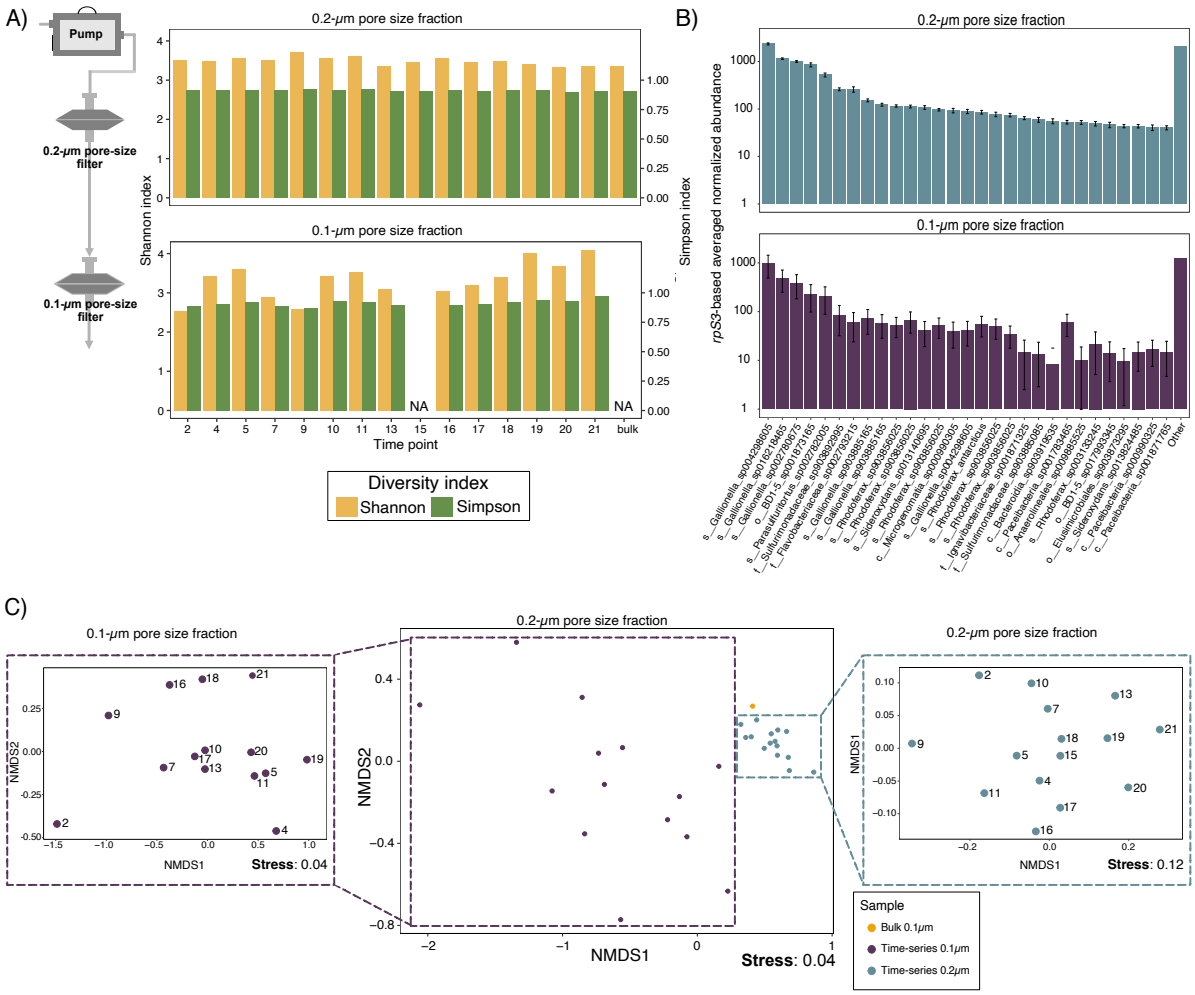


Figure 1. Microbial community composition of the geyser Wallender Born. A) The scheme to the left side illustrates the principle of sequential filtration we used for this study. The bargraph on the right displays the Shannon and Simpson diversity indices for the 0.2-µm and the 0.1-µm filtered fractions. 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. 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.

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 belonged to 26 different phyla, including Nitrosomonadales,

454 Proteobacteria, Nitrospirae and Bacteroidetes, and multiple bacteria of the Candidate Phyla
455 Radiation (CPR), in particular Nomurabacteria and Gracilibacteria (Figure 2). Microbial
456 communities, in particular those of the 0.2-µm filter fraction, were stable across the sampled
457 time series, with the filter fraction being the main discriminatory factor for microbial
458 communities (Figure 3A). Ordination analyses based on relative abundance of MAGs agreed
459 with *rpS3*-based ordinations, displaying that there was a significant difference between the
460 communities on the 0.1-µm filters compared to those of the 0.2-µm filters, which were
461 comparatively stable (Figure 1C). Comparison with known cold-water geyser microbial
462 communities [59,60] (Geyser Andernach (GA), Germany; and Crystal Geyser (CG), USA)
463 showed that most community members were specific to their respective ecosystems, with
464 only Altiarchaea and Hydrogenophilales being present in all three ecosystems
465 (Supplementary Figure S5). While Altiarchaea were the dominant organisms in both GA and
466 one aquifer in CG, they constituted a minority in WB. Out of the 123 WB genomes, 24 had
467 related populations based on FastANI (>= 50% genome coverage, >= 75% ANI, [61]) in the CG
468 ecosystem while only a single population genome was related between the WB and GA
469 populations, which might relate to the larger diversity of the CG ecosystem compared to the
470 other two ecosystems.
471 Although other high-CO_2 geysers have been metagenomically analyzed in the past [16,60,67],
472 little is known about the activity or the contribution of the organisms to the protein pool of
473 the community. To elucidate the expression profile (*i.e.*, expressed phenotype and metabolic
474 functions) of the detected organisms in the metagenomes of geyser WB, we performed
475 metaproteomics of the very same samples that were analyzed for metagenomics (see
476 Supplementary Figure S2 for sample distribution of metagenomics and metaproteomics).
477 Metaproteomic datasets showed a batch effect in the NMDS analysis, with the three
478 categories of samples (time-series, size, and bulk) clustering separately in the ordination
479 (Supplementary Figure S6). Subsequently, we focused our analyses on the 0.2-µm fraction
480 only and compared metaproteomic and metagenomic data of these samples. Across time, the
481 proteomics profile of the community in the 0.2-µm fraction was more dynamic than the
482 relatively stable genome abundances, with only few organisms showing consistent expression
483 throughout the time series (Figure 2B-D). Only the most abundant organisms in
484 metagenomics (including a Campylobacter, a Hydrogenophilales, and two Nitrosomonadales
485 closely related to Gallionella) were also fairly consistently represented in the proteomic

datasets (Figure 2). However, the most abundant members based on proteomics included organisms of the Anaerolineae, a Paceibacteria, a Gracilibacteria, as well as some Nitrosomonadales, which were all only minor community members in the metagenomic dataset (Figure 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.

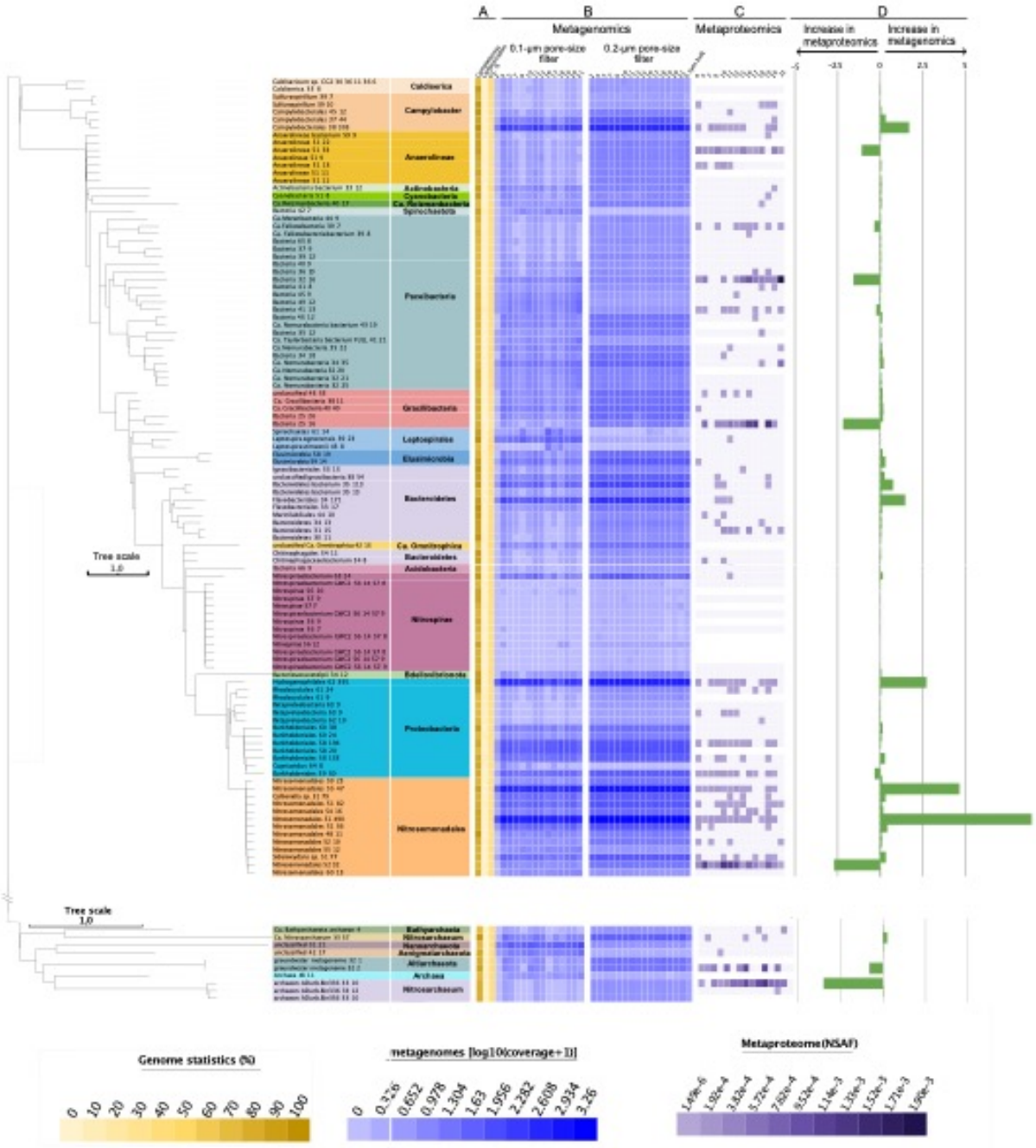


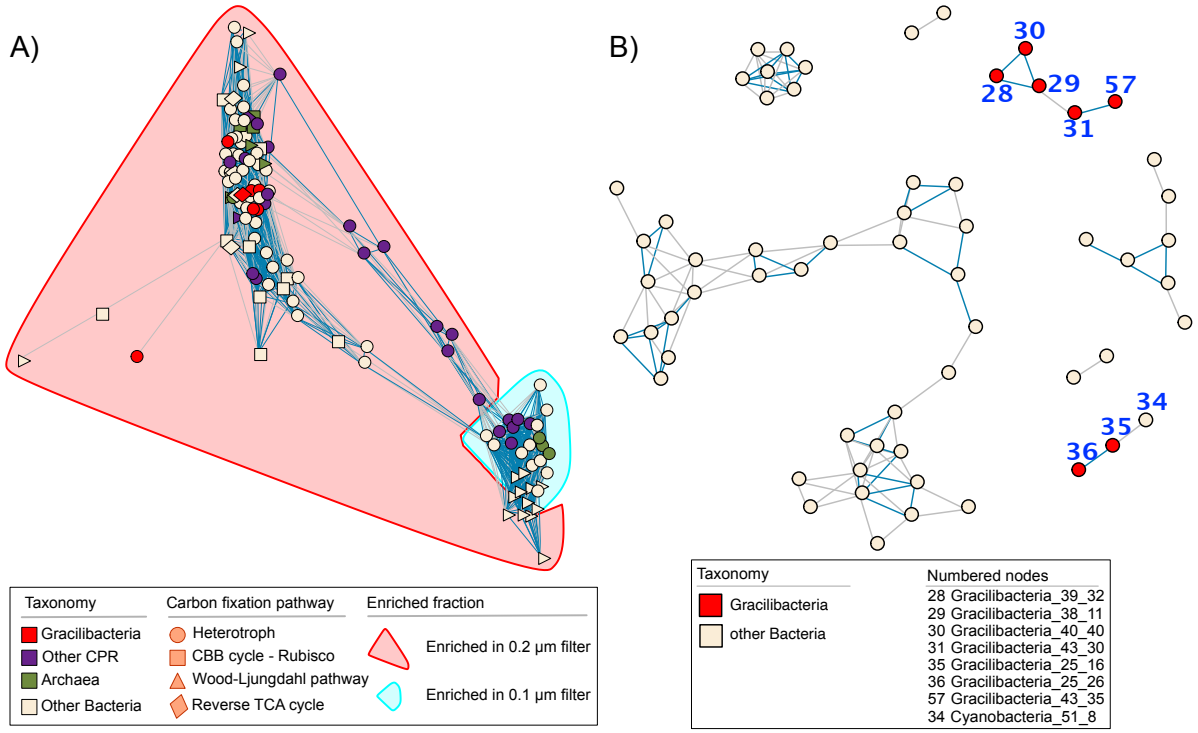
Figure 2. Phylogeny of high-quality genomes and their prevalence in metaproteogenomics. Colored sections to the right of the phylogenetic tree denote the phylum-level taxonomic classifications of high-quality genomes. **A)** Yellow-colored heatmap indicates the genome statistics (completeness, contamination, and GC % content).

496 **B)** The blue heatmap shows the abundance of the genomes throughout the time-series, separated into 0.1-µm
497 filter and the 0.2-µm filter fractions. **C)** The purple heatmap conveys the abundance of the genomes based on
498 the summed counts (NSAF) of proteins belonging to the respective genomes. **D)** Green bar charts denote the
499 difference in the median relative abundance of the metagenomic-based genome abundance relative to the
500 proteomic-based genome abundance on the 0.2-µm filters. The median relative abundance was calculated by
501 setting the total sample abundance to 100%, both in metagenomics as well as in metaproteomics, and
502 afterwards calculating the median percentage per genome. Negative numbers indicate higher relative
503 abundance in metagenomic data than in the metaproteomic dataset, and vice versa for the positive percentage
504 numbers.
505

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

508 CPR bacteria, in particular Nomurabacteria and Gracilibacteria, represented 31 of the 123
509 unique MAGs recovered from geyser WB. Due to their predicted symbiotic lifestyle [4,8], we
510 sought to reconstruct co-occurrence networks to establish potential symbiont-host
511 relationships. As correlating relative abundance data, such as genome abundances, is
512 problematic due to relative data by default fostering significant correlations [53], we utilized
513 the proposed alternative proportionality [53], implemented in the R package '*propr*' [52]. The
514 resulting network showed two clusters of organisms (Figure 3A), which has previously been
515 suggested to be related to different aquifers and thus groundwater sources at Crystal Geyser
516 [59,68]. Nonetheless, further analyses showed that the clusters observed for WB are due to
517 the fractionation, where microbes group based on their enrichment pattern in either the 0.1-
518 µm or 0.2-µm filter fraction. We observed that the WB Gracilibacteria genomes are enriched
519 in the 0.2-µm fraction, while other CPRs were either distributed between both network
520 clusters or situated between the clusters.
521 When focusing the analysis on the dataset from 0.2-µm filters, on which both hosts and host-
522 symbiont attachments should accumulate, the proportionality network showed only one
523 weak link for one out of seven Gracilibacteria with another organism, a Cyanobacterium
524 (genome Cyanobacteria_51_8; 5% FDR; Figure 3B). *Vampirococcus*, closely related to the WB
525 Gracilibacteria (Figure 4A), have been reported to predate on autotrophs such as *Chromatium*
526 and *Halochromatium* [23,69]. Although there is one report of viable Cyanobacteria surviving
527 on a hydrogen-based metabolism in a 613-m-deep borehole in Spain [70], these organisms
528 can be contaminants of open geysers [60]. While these Cyanobacteria may still fall prey to

529 Gracilibacteria, correlation and proportionality networks have previously been discussed to
530 be biased towards organisms with scavenging lifestyles (based on Lotka-Volterra models;
531 [71]). This reported bias, and given the absence of co-correlation with Gracilibacteria, we
532 suggest that the detected Gracilibacteria populations follow a parasitic lifestyle as previously
533 described for the closely related Abscondidabacteria [23,24].

534 

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

541 ***Metaproteomics suggests an active symbiotic/predatory lifestyle of WB Gracilibacteria***

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

560 The seven Gracilibacteria genomes also encoded for interconversion steps for the purine and
561 pyrimidine biosynthesis pathways, with one genome (WB Gracilibacteria 26_16) encoding for
562 the conversion of ribose-5P to phosphoribosyl diphosphate (PRPP) (Figure 4A). The pyrimidine
563 interconversion steps start with UMP/dUMP and ultimately yield CTP/dCTP that can be used
564 to synthesize RNA and DNA, respectively. As for purines, the conversion steps go from IMP to
565 ATP and dATP which, in a similar manner as pyrimidines, can serve as substrate for the
566 synthesis of RNA and DNA, respectively. While PRPP is an important metabolic intermediate,
567 in particular for the biosynthesis of nucleotides [72], most of the WB Gracilibacteria do not
568 encode for its synthesis (Figure 4) suggesting that this intermediate has to be retrieved from
569 external sources.

570

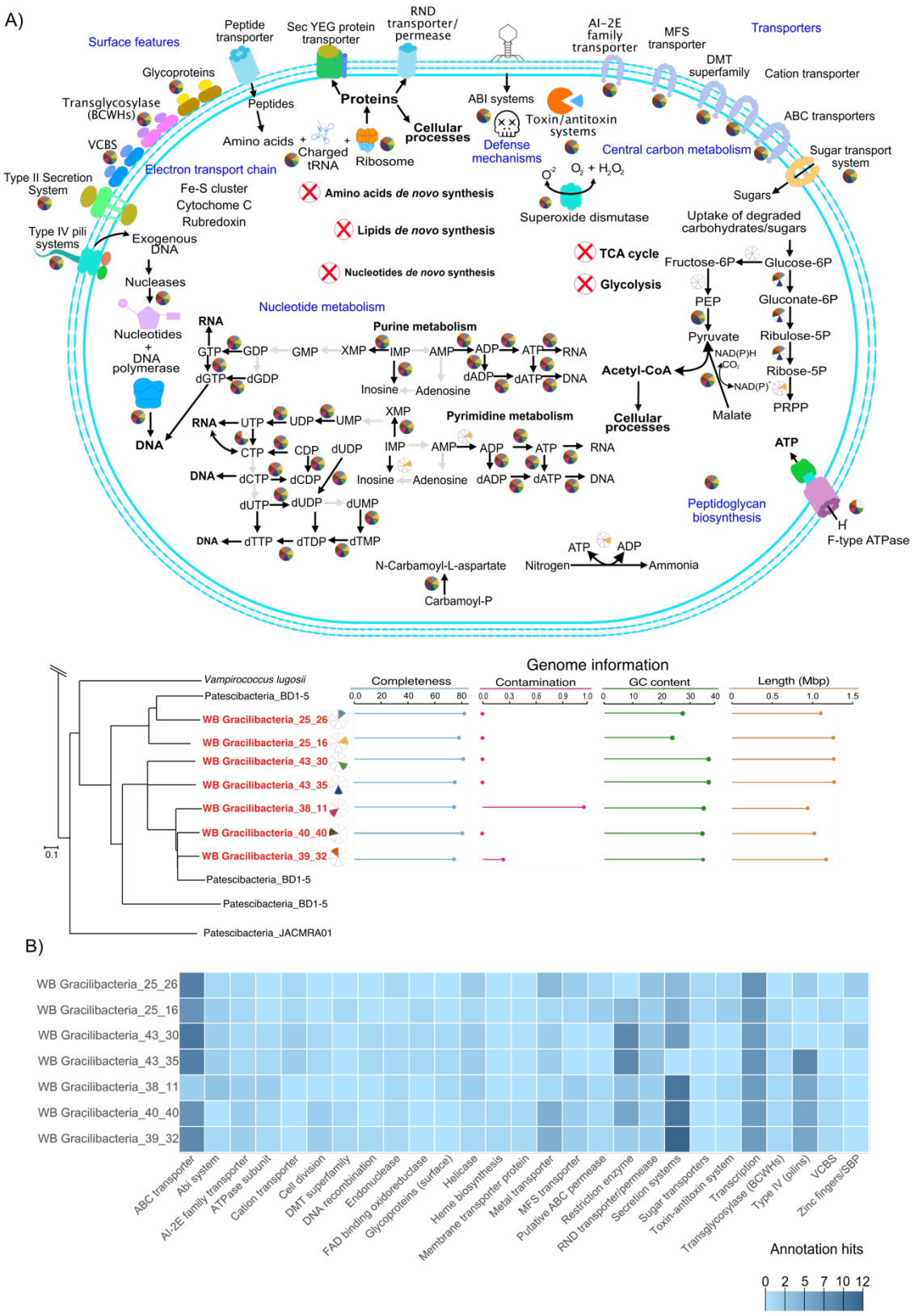


Figure 4. Metabolic overview of the seven WB Gracilibacteria genomes. A) Schematic metabolic overview of the major traits found in the WB Gracilibacteria and phylogenetic placement of the seven genomes along with genome statistics. Please note that the individual wedges of the pie-charts relate to the presence of individual

575 metabolic steps encoded in the different genomes, depicted in the metabolic scheme. **B)** Specific proteins of
576 interest that were identified in the genome annotation of Gracilibacteria.
577
578 The limited metabolism of the WB Gracilibacteria suggests that these bacteria require
579 molecular building blocks from other organisms (refer to Supplementary Figure S7 and Table
580 S7 for the full KEGG annotation of these genomes), but it remains unclear how they obtain
581 these necessary biomolecules. Given that no potential host(s) for the WB Gracilibacteria could
582 be identified using co-occurrence patterns, as well as their extremely limited central
583 metabolism, we set out to investigate their potential lifestyle based on metaproteogenomics,
584 focusing on (trans)membrane and cell surface features (see Figure 4B and Supplementary
585 Table S8 for proteome annotations of Gracilibacteria MAGs). We identified membrane-
586 associated proteins predicted to be functioning as transporters, in particular ABC
587 transporters, which aid in the uptake of a broad spectrum of molecules from the surrounding
588 of the cell [73–76]. Other transporters identified in the genome, such as the DMT superfamily
589 and major facilitator superfamily (MFS), enable the transport of drugs and a broad range of
590 molecules across the membrane, including antibiotics [77]. Consistent with other
591 Gracilibacteria, the WB Gracilibacteria expressed type IV pili systems,which can aid in the
592 uptake of DNA from the environment [78,79] and have been shown to be essential for host
593 adhesion in the case of TM7i and its *Leucobacter aridicollis* host [80]. The DNA uptake is
594 further supported by the observed strong expression of endonucleases. Consequently, WB
595 Gracilibacteria were predicted to attach to their host or make use of their appendages to
596 obtain molecules, which is likely necessary during the turbulent eruptions of the geyser.
597 Furthermore, most WB Gracilibacteria encoded for the resistance-nodulation-division (RND)
598 family of proteins, which also have transporter functions and efflux activity [81] but were not
599 detected in the proteome. Two genomes (WB Gracilibacteria 25_16 and 25_26) encoded for
600 a putative Vibrio, Colwellia, Bradyrhizobium, and Shewanella (VCBS) repeat-containing
601 protein, a group of proteins that are proposed to contain domains with functions related to
602 cell adhesion (InterPro entry: IPR010221). We were able to identify them in the proteomic
603 data by means of manual search of the unclassified proteins (Supplementary Figure S8,
604 “Adhesion” category). Translocases and peptidases identified in the proteome could aid in
605 obtaining amino acids probably to mainly carry out protein synthesis but also other reactions,
606 like deamination, and mobilization of molecules across the membrane. Strikingly, we

607 identified a vast number of type II and IV secretion systems in all seven genomes, of which
608 some subunits were also identified in the proteome (Supplementary Figure S8, "Secretion
609 system" category). Type IV secretion systems are related to pili and its assembly subunits [82–
610 84], while secretion systems II are used by Gram-negative bacteria to excrete proteins and
611 other molecules that aid in signaling or microbial pathogenicity [82,85–88]. Further evidence
612 for the active expression of type II secretion systems in Gracilibacteria is their differential
613 expression when comparing the most abundant proteins between the 0.2-µm filter and the
614 0.1-µm filter, showing that subunits for these secretion systems are highly expressed by WB
615 Gracilibacteria (Supplementary Figure S9). We also identified, in both genome and proteome
616 data, subunits of the translocase Sec system, which is important for the maturation and
617 secretion of proteins and their insertion into the inner membrane [82,89–91]. Two proteins
618 were annotated as transglycosylase, which are involved in cell wall interactions, such as
619 anchoring and even cell lysis [92–94]. Added to these cell surface features, the WB
620 Gracilibacteria also encoded for a complete peptidoglycan biosynthesis.
621 Publicly available Gracilibacteria genomes are missing steps in central metabolic pathways,
622 such as glycolysis and the tricarboxylic acid cycle (TCA), but encode for the complete oxidative
623 branch of the pentose pathway [23,24,28]. Likewise, the WB Gracilibacteria showed similar
624 traits and encode for several cell surface features associated with a predatory lifestyle (Figure
625 4). Annotations included gene functions related to transporters (*e.g*, ABC, protein exporters,
626 efflux systems), secretion systems type II and IV, pilus and flagella assembly, oxidoreductases,
627 peptidases, DNA repair, peptidoglycan assembly, and nucleotide interconversions, all of
628 which agree to the previously described metabolic patterns encoded in cultivated and
629 uncultivated Gracilibacteria [23,24].
630 Taken together, the presence of the abovementioned findings suggests an active predatory
631 lifestyle in the groundwater fluids discharged by the Geyser, where the WB Gracilibacteria
632 obtained their needed biomolecules by means of adhesion using pili and flagella,
633 subsequently lysed the target cell using their surface features such as the injection and efflux
634 systems, and ultimately made use of their transporters and appendages repertoire to take in
635 the cellular components released. Once taken in, the range of enzymes (*e.g.,* endonucleases,
636 proteases, metalloenzymes) in the Gracilibacteria would be in charge of breaking down big
637 molecules and incorporating smaller molecules into its metabolism.
638

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

640 While the functional significance of Gracilibacteria in the geyser's ecosystem is still unclear,
641 their symbiotic lifestyle coupled to their great abundance in metagenomics and
642 metaproteomics demonstrates that they are key players in the geyser's community. Taking
643 into account that cell counts for the 0.2-µm filter samples ranged between $2.07x10^4$ to
644 $5.94x10^4$ cells ml^{-1} (Table S1), cell numbers that were on the low end for a shallow subsurface
645 environment [95], and considering the regular flushing of the geyser's ecosystem (eruptions
646 occur in 30-40 minute intervals), it can be assumed that availability of molecules is low. Added
647 to this, it has been proposed that the small cell size of CPR, in this particular case of
648 Gracilibacteria, is an advantage in oligotrophic environments since the increased surface-to-
649 volume ratio of the cell is optimal for the uptake of nutrients [96,97]. Additionally, the small
650 cell size would allow for more Gracilibacteria to target the same host cells [21]. Considering
651 that most of the Gracilibacteria genomes did not co-correlate with other prokaryotes in the
652 community, we assume that the Gracilibacteria can target a broad range of microbes to feed
653 on. This assumption and the high abundance of Gracilibacteria compared to other members
654 of the overall geyser's microbial community (Figure 2) indicates that the Gracilibacteria thrive
655 in this environment, likely by means of attaching to other microbes to predate, as opposed to
656 scavenging in such a low-nutrient environment. If one considers a similar case as with *Ca.* A.
657 praedator, which is proposed to play a role in population control by a "Kill the Winner"
658 approach [24], Gracilibacteria in WB could potentially target other organisms in the geyser,
659 lysing them and effectively keeping any one species from outcompeting the rest of the
660 microbial community. In oligotrophic environments, such as geyser WB, it is expected that
661 predators, deemed "competition specialists", dominate the community since competition is
662 a selective pressure [98]. It should be noted that the "winner" in such scenarios is not
663 necessarily the most abundant organism (*e.g.,* Gallionella in WB) but rather the competition
664 specialist [98], in this case the predatory Gracilibacteria or key stone organisms that are low
665 abundant (infrequent targets) but highly active. We propose that Gracilibacteria likely serve
666 as a population control in WB, aiding in maintaining the dynamics of the geyser's microbiome
667 stable throughout time.

668 **Conclusions**

669 We present an in-depth metaproteogenomics study of a groundwater microbial community
670 accessed through a cold-water geyser. Metagenomic abundances show that the microbial
671 community is dominated by Nitrosomonadales, specifically by members of Gallionella spp.,
672 Campylobacter and Hydrogenophilales, which also showed expression in the metaproteomic
673 data. Nonetheless, metaproteomics showed that Anaerolineae, Paceibacteria,
674 Nitrosomonadales, and in particular Gracilibacteria were the most active in terms of protein
675 abundance. Analyzing metaproteogenomic data, we were able to draw conclusions regarding
676 the lifestyle of Gracilibacteria. They show metabolic overlaps with other members of the
677 Gracilibacteria lineage, with very limited central metabolism, while encoding for
678 interconversion steps of nucleotide biosynthesis and peptidoglycan synthesis. However, their
679 core metabolism patchiness is compensated by encoding and expression of several types of
680 transporters, membrane-related proteins, and enzymes. Added to this, we identified proteins
681 that point towards cell-cell interactions and type II and type IV secretion systems. Based on
682 these results, we propose that the WB Gracilibacteria follow a predatory lifestyle, likely being
683 capable of lysing other members of the geyser's microbial community. Considering the
684 membrane features we identified and the lack of identification of a potential host using co-
685 occurrence patterns, we hypothesize that the WB Gracilibacteria are predators that rely on
686 other members of the community to obtain molecular building blocks to survive. At the same
687 time, WB Gracilibacteria potentially assist in population control for the microbial community
688 of WB geyser. Additionally, they have the potential to become a liability to the community if
689 nutrients deplete too much or too fast, or if stress conditions become challenging, during
690 which their predatory lifestyle would be an advantage to survive. The results of this
691 metaproteogenomic study focusing on groundwater Gracilibacteria provides *in situ* evidence
692 regarding the activity of Gracilibacteria suggesting a predatory lifestyle in oligotrophic
693 groundwater ecosystems.

694

695 **Declarations**

696 **Ethics approval and consent to participate**

697 Not applicable

698 **Consent for publication**

699 Not applicable

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.

Funding

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.

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. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [99] partner repository with the dataset identifier reviewer_pxd042980@ebi.ac.uk and the password goTjmuGe.

Competing interests

The authors declare that they have no competing interests.

Author 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

732 as carried out the geochemical measurements. T.H., S.M., and D.B. conducted proteomic
733 extraction and measurements, as well as data interpretation of it. J.S., C.M., S.P.E., and J.P.
734 aided in the metagenomic analyses and binning. U.S. provided geological information. B.B.
735 aided in discussion and insight into data interpretation. A.J.P. conceptualized the study.
736 P.A.F.G., T.L.V.B. and A.J.P. wrote the manuscript with revisions from all co-authors.
737

738 **Bibliography**

739 [1] D.H. Parks, M. Chuvochina, D.W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P.
740 Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny
741 substantially revises the tree of life, Nat Biotechnol. 36 (2018) 996–1004.
742 https://doi.org/10.1038/nbt.4229.
743 [2] L.A. Hug, B.J. Baker, K. Anantharaman, C.T. Brown, A.J. Probst, C.J. Castelle, C.N.
744 Butterfield, A.W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D.A. Relman, K.M.
745 Finstad, R. Amundson, B.C. Thomas, J.F. Banfield, A new view of the tree of life, Nat
746 Microbiol. 1 (2016) 16048. https://doi.org/10.1038/nmicrobiol.2016.48.
747 [3] D.H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B.J. Woodcroft, P.N. Evans, P.
748 Hugenholtz, G.W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes
749 substantially expands the tree of life, Nat Microbiol. 2 (2017) 1533–1542.
750 https://doi.org/10.1038/s41564-017-0012-7.
751 [4] C.T. Brown, L.A. Hug, B.C. Thomas, I. Sharon, C.J. Castelle, A. Singh, M.J. Wilkins, K.C.
752 Wrighton, K.H. Williams, J.F. Banfield, Unusual biology across a group comprising more
753 than 15% of domain Bacteria, Nature. 523 (2015) 208–211.
754 https://doi.org/10.1038/nature14486.
755 [5] B. Luef, K.R. Frischkorn, K.C. Wrighton, H.-Y.N. Holman, G. Birarda, B.C. Thomas, A.
756 Singh, K.H. Williams, C.E. Siegerist, S.G. Tringe, K.H. Downing, L.R. Comolli, J.F. Banfield,
757 Diverse uncultivated ultra-small bacterial cells in groundwater, Nat Commun. 6 (2015)
758 6372. https://doi.org/10.1038/ncomms7372.
759 [6] C.J. Castelle, J.F. Banfield, Major New Microbial Groups Expand Diversity and Alter our
760 Understanding of the Tree of Life, Cell. 172 (2018) 1181–1197.
761 https://doi.org/10.1016/j.cell.2018.02.016.
762 [7] K.C. Wrighton, B.C. Thomas, I. Sharon, C.S. Miller, C.J. Castelle, N.C. VerBerkmoes, M.J.
763 Wilkins, R.L. Hettich, M.S. Lipton, K.H. Williams, P.E. Long, J.F. Banfield, Fermentation,
764 Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla, Science. 337
765 (2012) 1661–1665. https://doi.org/10.1126/science.1224041.
766 [8] C.J. Castelle, C.T. Brown, K. Anantharaman, A.J. Probst, R.H. Huang, J.F. Banfield,
767 Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN
768 radiations, Nat Rev Microbiol. 16 (2018) 629–645. https://doi.org/10.1038/s41579-
769 018-0076-2.
770 [9] R. Méheust, D. Burstein, C.J. Castelle, J.F. Banfield, The distinction of CPR bacteria from
771 other bacteria based on protein family content, Nat Commun. 10 (2019) 4173.
772 https://doi.org/10.1038/s41467-019-12171-z.
773 [10] J. Gong, Y. Qing, X. Guo, A. Warren, “Candidatus Sonnebornia yantaiensis”, a member
774 of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium

775 bursaria (Ciliophora, Oligohymenophorea), Systematic and Applied Microbiology. 37
776 (2014) 35–41. https://doi.org/10.1016/j.syapm.2013.08.007.
777 [11] R.E. Danczak, M.D. Johnston, C. Kenah, M. Slattery, K.C. Wrighton, M.J. Wilkins,
778 Members of the Candidate Phyla Radiation are functionally differentiated by carbon-
779 and nitrogen-cycling capabilities, Microbiome. 5 (2017) 112.
780 https://doi.org/10.1186/s40168-017-0331-1.
781 [12] C. Wurzbacher, R.H. Nilsson, M. Rautio, S. Peura, Poorly known microbial taxa
782 dominate the microbiome of permafrost thaw ponds, ISME J. 11 (2017) 1938–1941.
783 https://doi.org/10.1038/ismej.2017.54.
784 [13] E.P. Starr, S. Shi, S.J. Blazewicz, A.J. Probst, D.J. Herman, M.K. Firestone, J.F. Banfield,
785 Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria
786 utilize microbially-processed plant-derived carbon, Microbiome. 6 (2018) 122.
787 https://doi.org/10.1186/s40168-018-0499-z.
788 [14] P. Geesink, C. Wegner, A.J. Probst, M. Herrmann, H.T. Dam, A. Kaster, K. Küsel,
789 Genome-inferred spatio-temporal resolution of an uncultivated Roizmanbacterium
790 reveals its ecological preferences in groundwater, Environmental Microbiology. 22
791 (2020) 726–737. https://doi.org/10.1111/1462-2920.14865.
792 [15] J.S. McLean, B. Bor, K.A. Kerns, Q. Liu, T.T. To, L. Solden, E.L. Hendrickson, K. Wrighton,
793 W. Shi, X. He, Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome
794 Bacteria to Mammalian Hosts, Cell Reports. 32 (2020) 107939.
795 https://doi.org/10.1016/j.celrep.2020.107939.
796 [16] C. He, R. Keren, M.L. Whittaker, I.F. Farag, J.A. Doudna, J.H.D. Cate, J.F. Banfield,
797 Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR
798 bacteria and DPANN archaea in groundwater ecosystems, Nat Microbiol. 6 (2021) 354–
799 365. https://doi.org/10.1038/s41564-020-00840-5.
800 [17] A.M. Nicolas, A.L. Jaffe, E.E. Nuccio, M.E. Taga, M.K. Firestone, J.F. Banfield, Soil
801 Candidate Phyla Radiation Bacteria Encode Components of Aerobic Metabolism and
802 Co-occur with Nanoarchaea in the Rare Biosphere of Rhizosphere Grassland
803 Communities, mSystems. 6 (2021) e01205-20.
804 https://doi.org/10.1128/mSystems.01205-20.
805 [18] S. Batinovic, J.J. Rose, J. Ratcliffe, R.J. Seviour, S. Petrovski, Cocultivation of an
806 ultrasmall environmental parasitic bacterium with lytic ability against bacteria
807 associated with wastewater foams, Nat Microbiol. 6 (2021).
808 [19] X. He, J.S. McLean, A. Edlund, S. Yooseph, A.P. Hall, S.-Y. Liu, P.C. Dorrestein, E.
809 Esquenazi, R.C. Hunter, G. Cheng, K.E. Nelson, R. Lux, W. Shi, Cultivation of a human-
810 associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle,
811 Proc Natl Acad Sci USA. 112 (2015) 244–249.
812 https://doi.org/10.1073/pnas.1419038112.
813 [20] B. Bor, N. Poweleit, J.S. Bois, L. Cen, J.K. Bedree, Z.H. Zhou, R.P. Gunsalus, R. Lux, J.S.
814 McLean, X. He, W. Shi, Phenotypic and Physiological Characterization of the Epibiotic
815 Interaction Between TM7x and Its Basibiont Actinomyces, Microb Ecol. 71 (2016) 243–
816 255. https://doi.org/10.1007/s00248-015-0711-7.
817 [21] B. Bor, J.S. McLean, K.R. Foster, L. Cen, T.T. To, A. Serrato-Guillen, F.E. Dewhirst, W. Shi,
818 X. He, Rapid evolution of decreased host susceptibility drives a stable relationship
819 between ultrasmall parasite TM7x and its bacterial host, Proc. Natl. Acad. Sci. U.S.A.
820 115 (2018) 12277–12282. https://doi.org/10.1073/pnas.1810625115.

[22] B. Bor, A.J. Collins, P.P. Murugkar, S. Balasubramanian, T.T. To, E.L. Hendrickson, J.K. Bedree, F.B. Bidlack, C.D. Johnston, W. Shi, J.S. McLean, X. He, F.E. Dewhirst, Insights Obtained by Culturing Saccharibacteria With Their Bacterial Hosts, J Dent Res. 99 (2020) 685–694. https://doi.org/10.1177/0022034520905792.

[23] D. Moreira, Y. Zivanovic, A.I. López-Archilla, M. Iniesto, P. López-García, Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii, Nat Commun. 12 (2021) 2454. https://doi.org/10.1038/s41467-021-22762-4.

[24] M.M. Yakimov, A.Y. Merkel, V.A. Gaisin, M. Pilhofer, E. Messina, J.E. Hallsworth, A.A. Klyukina, E.N. Tikhonova, V.M. Gorlenko, Cultivation of a vampire: ‘ *Candidatus* Absconditicoccus praedator,’ Environmental Microbiology. 24 (2022) 30–49. https://doi.org/10.1111/1462-2920.15823.

[25] L. Li, C. Kato, K. Horikoshi, Bacterial diversity in deep-sea sediments from different depths, Biodiversity & Conservation. (1999) 659–667.

[26] A. Hanke, E. Hamann, R. Sharma, J.S. Geelhoed, T. Hargesheimer, B. Kraft, V. Meyer, S. Lenk, H. Osmers, R. Wu, K. Makinwa, R.L. Hettich, J.F. Banfield, H.E. Tegetmeyer, M. Strous, Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat, Front. Microbiol. 5 (2014). https://doi.org/10.3389/fmicb.2014.00231.

[27] N.N. Ivanova, P. Schwientek, H.J. Tripp, C. Rinke, A. Pati, M. Huntemann, A. Visel, T. Woyke, N.C. Kyrpides, E.M. Rubin, Stop codon reassignments in the wild, Science. 344 (2014) 909–913.

[28] C.M.K. Sieber, B.G. Paul, C.J. Castelle, P. Hu, S.G. Tringe, D.L. Valentine, G.L. Andersen, J.F. Banfield, Unusual Metabolism and Hypervariation in the Genome of a Gracilibacterium (BD1-5) from an Oil-Degrading Community, mBio. 10 (2019) e02128-19. https://doi.org/10.1128/mBio.02128-19.

[29] W. Meyer, Geologie der Eifel: mit 12 Tabellen, 4., völlig neu bearb. Aufl, Schweizerbart, Stuttgart, 2013.

[30] D. Dittrich, Der paläogeographische Nord-Anschluss des marin beeinflussten höheren Buntsandsteins der Trier-Luxemburger Bucht, Mainzer geowissenschaftliche Mitteilungen. (2021) 105–168.

[31] E. Griesshaber, R.K. O’Nions, E.R. Oxburgh, 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] L.L. Stookey, Ferrozine---a new spectrophotometric reagent for iron, Anal. Chem. 42 (1970) 779–781. https://doi.org/10.1021/ac60289a016.

[33] S. Deusch, J. Seifert, Catching the tip of the iceberg - Evaluation of sample preparation protocols for metaproteomic studies of the rumen microbiota, Proteomics. 15 (2015) 3590–3595. https://doi.org/10.1002/pmic.201400556.

[34] F. Bonn, J. Bartel, K. Büttner, M. Hecker, A. Otto, D. Becher, Picking Vanished Proteins from the Void: How to Collect and Ship/Share Extremely Dilute Proteins in a Reproducible and Highly Efficient Manner, Anal. Chem. 86 (2014) 7421–7427. https://doi.org/10.1021/ac501189j.

[35] L.M. Rodriguez-R, S. Gunturu, J.M. Tiedje, J.R. Cole, K.T. Konstantinidis, Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity, mSystems. 3 (2018) e00039-18. https://doi.org/10.1128/mSystems.00039-18.

867 [36] D. Antipov, M. Raiko, A. Lapidus, P.A. Pevzner, METAVIRALSPADES: assembly of viruses
868 from metagenomic data, Bioinformatics. 36 (2020) 4126–4129.
869 https://doi.org/10.1093/bioinformatics/btaa490.
870 [37] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat Methods. 9
871 (2012) 357–359. https://doi.org/10.1038/nmeth.1923.
872 [38] S. Nurk, D. Meleshko, A. Korobeynikov, P.A. Pevzner, metaSPAdes: a new versatile
873 metagenomic assembler, Genome Res. 27 (2017) 824–834.
874 https://doi.org/10.1101/gr.213959.116.
875 [39] D. Hyatt, G.-L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal:
876 prokaryotic gene recognition and translation initiation site identification, BMC
877 Bioinformatics. 11 (2010) 119. https://doi.org/10.1186/1471-2105-11-119.
878 [40] T.L.V. Bornemann, S.P. Esser, T.L. Stach, T. Burg, A.J. Probst, uBin:A manual refining
879 tool for genomes from metagenomes, Environmental Microbiology. (2023) 1462-
880 2920.16351. https://doi.org/10.1111/1462-2920.16351.
881 [41] B. Zybailov, A.L. Mosley, M.E. Sardiu, M.K. Coleman, L. Florens, M.P. Washburn,
882 Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces c*
883 *erevisiae*, J. Proteome Res. 5 (2006) 2339–2347. https://doi.org/10.1021/pr060161n.
884 [42] J. Mistry, R.D. Finn, S.R. Eddy, A. Bateman, M. Punta, Challenges in homology search:
885 HMMER3 and convergent evolution of coiled-coil regions, Nucleic Acids Research. 41
886 (2013) e121–e121. https://doi.org/10.1093/nar/gkt263.
887 [43] A.E. Darling, G. Jospin, E. Lowe, F.A. Matsen, H.M. Bik, J.A. Eisen, PhyloSift:
888 phylogenetic analysis of genomes and metagenomes, PeerJ. 2 (2014) e243.
889 https://doi.org/10.7717/peerj.243.
890 [44] I. Sharon, M.J. Morowitz, B.C. Thomas, E.K. Costello, D.A. Relman, J.F. Banfield, Time
891 series community genomics analysis reveals rapid shifts in bacterial species, strains,
892 and phage during infant gut colonization, Genome Res. 23 (2013) 111–120.
893 https://doi.org/10.1101/gr.142315.112.
894 [45] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST,
895 Bioinformatics. 26 (2010) 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
896 [46] A. Lavy, D.G. McGrath, P.B. Matheus Carnevali, J. Wan, W. Dong, T.K. Tokunaga, B.C.
897 Thomas, K.H. Williams, S.S. Hubbard, J.F. Banfield, Microbial communities across a
898 hillslope-riparian transect shaped by proximity to the stream, groundwater table, and
899 weathered bedrock, Ecology and Evolution. 9 (2019) 6869–6900.
900 https://doi.org/10.1002/ece3.5254.
901 [47] Y.-W. Wu, B.A. Simmons, S.W. Singer, MaxBin 2.0: an automated binning algorithm to
902 recover genomes from multiple metagenomic datasets, Bioinformatics. 32 (2016) 605–
903 607. https://doi.org/10.1093/bioinformatics/btv638.
904 [48] C.M.K. Sieber, A.J. Probst, A. Sharrar, B.C. Thomas, M. Hess, S.G. Tringe, J.F. Banfield,
905 Recovery of genomes from metagenomes via a dereplication, aggregation and scoring
906 strategy, Nat Microbiol. 3 (2018) 836–843. https://doi.org/10.1038/s41564-018-0171-
907 1.
908 [49] M.R. Olm, C.T. Brown, B. Brooks, J.F. Banfield, dRep: a tool for fast and accurate
909 genomic comparisons that enables improved genome recovery from metagenomes
910 through de-replication, ISME J. 11 (2017) 2864–2868.
911 https://doi.org/10.1038/ismej.2017.126.
912 [50] D.H. Parks, M. Imelfort, C.T. Skennerton, P. Hugenholtz, G.W. Tyson, CheckM: assessing
913 the quality of microbial genomes recovered from isolates, single cells, and

914 metagenomes, Genome Res. 25 (2015) 1043–1055.
915 https://doi.org/10.1101/gr.186072.114.
916 [51] P.-A. Chaumeil, A.J. Mussig, P. Hugenholtz, D.H. Parks, GTDB-Tk: a toolkit to classify
917 genomes with the Genome Taxonomy Database, Bioinformatics. (2019) btz848.
918 https://doi.org/10.1093/bioinformatics/btz848.
919 [52] T.P. Quinn, M.F. Richardson, D. Lovell, T.M. Crowley, propr: An R-package for
920 Identifying Proportionally Abundant Features Using Compositional Data Analysis, Sci
921 Rep. 7 (2017) 16252. https://doi.org/10.1038/s41598-017-16520-0.
922 [53] D. Lovell, V. Pawlowsky-Glahn, J.J. Egozcue, S. Marguerat, J. Bähler, Proportionality: A
923 Valid Alternative to Correlation for Relative Data, PLoS Comput Biol. 11 (2015)
924 e1004075. https://doi.org/10.1371/journal.pcbi.1004075.
925 [54] Z. Zhou, P.Q. Tran, A.M. Breister, Y. Liu, K. Kieft, E.S. Cowley, U. Karaoz, K.
926 Anantharaman, METABOLIC: high-throughput profiling of microbial genomes for
927 functional traits, metabolism, biogeochemistry, and community-scale functional
928 networks, Microbiome. 10 (2022) 33. https://doi.org/10.1186/s40168-021-01213-8.
929 [55] D. Vallenet, A. Calteau, M. Dubois, P. Amours, A. Bazin, M. Beuvin, L. Burlot, X. Bussell,
930 S. Fouteau, G. Gautreau, A. Lajus, J. Langlois, R. Planel, D. Roche, J. Rollin, Z. Rouy, V.
931 Sabatet, C. Médigue, MicroScope: an integrated platform for the annotation and
932 exploration of microbial gene functions through genomic, pangenomic and metabolic
933 comparative analysis, Nucleic Acids Research. (2019) gkz926.
934 https://doi.org/10.1093/nar/gkz926.
935 [56] R. Caspi, R. Billington, I.M. Keseler, A. Kothari, M. Krummenacker, P.E. Midford, W.K.
936 Ong, S. Paley, P. Subhraveti, P.D. Karp, The MetaCyc database of metabolic pathways
937 and enzymes - a 2019 update, Nucleic Acids Research. 48 (2020) D445–D453.
938 https://doi.org/10.1093/nar/gkz862.
939 [57] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids
940 Research. 28 (2000).
941 [58] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids
942 Research. 47 (2019) D506–D515. https://doi.org/10.1093/nar/gky1049.
943 [59] A.J. Probst, B. Ladd, J.K. Jarett, D.E. Geller-McGrath, C.M.K. Sieber, J.B. Emerson, K.
944 Anantharaman, B.C. Thomas, R.R. Malmstrom, M. Stieglmeier, A. Klingl, T. Woyke, M.C.
945 Ryan, J.F. Banfield, Differential depth distribution of microbial function and putative
946 symbionts through sediment-hosted aquifers in the deep terrestrial subsurface, Nat
947 Microbiol. 3 (2018) 328–336. https://doi.org/10.1038/s41564-017-0098-y.
948 [60] T.L.V. Bornemann, P.S. Adam, V. Turzynski, U. Schreiber, P.A. Figueroa-Gonzalez, J.
949 Rahlff, D. Köster, T.C. Schmidt, R. Schunk, B. Krauthausen, A.J. Probst, Genetic diversity
950 in terrestrial subsurface ecosystems impacted by geological degassing, Nat Commun.
951 13 (2022) 284. https://doi.org/10.1038/s41467-021-27783-7.
952 [61] C. Jain, L.M. Rodriguez-R, A.M. Phillippy, K.T. Konstantinidis, S. Aluru, High throughput
953 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries, Nat
954 Commun. 9 (2018) 5114. https://doi.org/10.1038/s41467-018-07641-9.
955 [62] H. Wickham, ggplot2: ggplot2, WIREs Comp Stat. 3 (2011) 180–185.
956 https://doi.org/10.1002/wics.147.
957 [63] A. Bunn, M. Korpela, Crossdating in dplR, (2013).
958 [64] D.H. Huson, D.C. Richter, C. Rausch, T. Dezulian, M. Franz, R. Rupp, Dendroscope: An
959 interactive viewer for large phylogenetic trees, BMC Bioinformatics. 8 (2007) 460.
960 https://doi.org/10.1186/1471-2105-8-460.

[65] I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments, Nucleic Acids Research. 47 (2019) W256–W259. https://doi.org/10.1093/nar/gkz239.
[66] G. Csardi, T. Nepusz, The igraph software package for complex network research, InterJournal, Complex Systems,. 1695 (2006).
[67] A.J. Probst, F.J. Elling, C.J. Castelle, Q. Zhu, M. Elvert, G. Birarda, H.-Y.N. Holman, K.R. Lane, B. Ladd, M.C. Ryan, T. Woyke, K.-U. Hinrichs, J.F. Banfield, Lipid analysis of CO2-rich subsurface aquifers suggests an autotrophy-based deep biosphere with lysolipids enriched in CPR bacteria, ISME J. 14 (2020) 1547–1560. https://doi.org/10.1038/s41396-020-0624-4.
[68] K. Schwank, T.L.V. Bornemann, N. Dombrowski, A. Spang, J.F. Banfield, A.J. Probst, An archaeal symbiont-host association from the deep terrestrial subsurface, Microbiology, 2018. https://doi.org/10.1101/486977.
[69] R. Guerrero, C. Pedrós-Alió, I. Esteve, J. Mas, D. Chase, L. Margulis, Predatory prokaryotes: Predation and primary consumption evolved in bacteria, Proc. Natl. Acad. Sci. U.S.A. 83 (1986) 2138–2142. https://doi.org/10.1073/pnas.83.7.2138.
[70] F. Puente-Sánchez, A. Arce-Rodríguez, M. Oggerin, M. García-Villadangos, M. Moreno-Paz, Y. Blanco, N. Rodríguez, L. Bird, S.A. Lincoln, F. Tornos, O. Prieto-Ballesteros, K.H. Freeman, D.H. Pieper, K.N. Timmis, R. Amils, V. Parro, Viable cyanobacteria in the deep continental subsurface, Proc. Natl. Acad. Sci. U.S.A. 115 (2018) 10702–10707. https://doi.org/10.1073/pnas.1808176115.
[71] M.-C. Anisiu, Lotka, Volterra and their models, Didáctica Mathematica. 32 (2014).
[72] B. Hove-Jensen, K.R. Andersen, M. Kilstrup, J. Martinussen, R.L. Switzer, M. Willemoës, Phosphoribosyl Diphosphate (PRPP): Biosynthesis, Enzymology, Utilization, and Metabolic Significance, Microbiol Mol Biol Rev. 81 (2017) e00040-16. https://doi.org/10.1128/MMBR.00040-16.
[73] A.H.F. Hosie, P.S. Poole, Bacterial ABC transporters of amino acids, Research in Microbiology. 152 (2001) 259–270. https://doi.org/10.1016/S0923-2508(01)01197-4.
[74] V.G. Lewis, M.P. Ween, C.A. McDevitt, The role of ATP-binding cassette transporters in bacterial pathogenicity, Protoplasma. 249 (2012) 919–942. https://doi.org/10.1007/s00709-011-0360-8.
[75] K.J. Tanaka, S. Song, K. Mason, H.W. Pinkett, Selective substrate uptake: The role of ATP-binding cassette (ABC) importers in pathogenesis, Biochimica et Biophysica Acta (BBA) - Biomembranes. 1860 (2018) 868–877. https://doi.org/10.1016/j.bbamem.2017.08.011.
[76] C. Orelle, K. Mathieu, J.-M. Jault, Multidrug ABC transporters in bacteria, Research in Microbiology. 170 (2019) 381–391. https://doi.org/10.1016/j.resmic.2019.06.001.
[77] E.M. Quistgaard, C. Löw, F. Guettou, P. Nordlund, Understanding transport by the major facilitator superfamily (MFS): structures pave the way, Nat Rev Mol Cell Biol. 17 (2016) 123–132. https://doi.org/10.1038/nrm.2015.25.
[78] C.K. Ellison, T.N. Dalia, A. Vidal Ceballos, J.C.-Y. Wang, N. Biais, Y.V. Brun, A.B. Dalia, Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae, Nat Microbiol. 3 (2018) 773–780. https://doi.org/10.1038/s41564-018-0174-y.
[79] K.H. Piepenbrink, DNA Uptake by Type IV Filaments, Front. Mol. Biosci. 6 (2019) 1. https://doi.org/10.3389/fmolb.2019.00001.

1007 [80] B. Xie, J. Wang, Y. Nie, J. Tian, Z. Wang, D. Chen, B. Hu, X.-L. Wu, W. Du, Type IV pili
1008 trigger episymbiotic association of Saccharibacteria with its bacterial host, Proceedings
1009 of the National Academy of Sciences. 119 (2022) e2215990119.
1010 https://doi.org/10.1073/pnas.2215990119.
1011 [81] D. Fernando, A. Kumar, Resistance-Nodulation-Division Multidrug Efflux Pumps in
1012 Gram-Negative Bacteria: Role in Virulence, Antibiotics. 2 (2013) 163–181.
1013 https://doi.org/10.3390/antibiotics2010163.
1014 [82] E.R. Green, J. Mecsas, Bacterial Secretion Systems: An Overview, Microbiol Spectr. 4
1015 (2016) 4.1.13. https://doi.org/10.1128/microbiolspec.VMBF-0012-2015.
1016 [83] L. Craig, K.T. Forest, B. Maier, Type IV pili: dynamics, biophysics and functional
1017 consequences, Nat Rev Microbiol. 17 (2019) 429–440. https://doi.org/10.1038/s41579-
1018 019-0195-4.
1019 [84] T. Jacobsen, B. Bardiaux, O. Francetic, N. Izadi-Pruneyre, M. Nilges, Structure and
1020 function of minor pilins of type IV pili, Med Microbiol Immunol. 209 (2020) 301–308.
1021 https://doi.org/10.1007/s00430-019-00642-5.
1022 [85] M. Sandkvist, Type II Secretion and Pathogenesis, Infect Immun. 69 (2001) 3523–3535.
1023 https://doi.org/10.1128/IAI.69.6.3523-3535.2001.
1024 [86] N.P. Cianciotto, Type II secretion: a protein secretion system for all seasons, Trends in
1025 Microbiology. 13 (2005) 581–588. https://doi.org/10.1016/j.tim.2005.09.005.
1026 [87] T.L. Johnson, J. Abendroth, W.G.J. Hol, M. Sandkvist, Type II secretion: from structure
1027 to function, FEMS Microbiology Letters. 255 (2006) 175–186.
1028 https://doi.org/10.1111/j.1574-6968.2006.00102.x.
1029 [88] K.V. Korotkov, M. Sandkvist, Architecture, Function, and Substrates of the Type II
1030 Secretion System, EcoSal Plus. 8 (2019) ecosalplus.ESP-0034-2018.
1031 https://doi.org/10.1128/ecosalplus.ESP-0034-2018.
1032 [89] J. De Geyter, D. Smets, S. Karamanou, A. Economou, Inner Membrane Translocases and
1033 Insertases, Bacterial Cell Walls and Membranes. 92 (2019) 337–366.
1034 https://doi.org/10.1007/978-3-030-18768-2_11.
1035 [90] B.M. Forster, H. Marquis, Protein transport across the cell wall of monoderm Gram-
1036 positive bacteria: Protein transport across the bacterial cell wall, Molecular
1037 Microbiology. 84 (2012) 405–413. https://doi.org/10.1111/j.1365-2958.2012.08040.x.
1038 [91] J.A. Lycklama a Nijeholt, A.J.M. Driessen, The bacterial Sec-translocase: structure and
1039 mechanism, Phil. Trans. R. Soc. B. 367 (2012) 1016–1028.
1040 https://doi.org/10.1098/rstb.2011.0201.
1041 [92] E. Scheurwater, C.W. Reid, A.J. Clarke, Lytic transglycosylases: Bacterial space-making
1042 autolysins, The International Journal of Biochemistry & Cell Biology. 40 (2008) 586–
1043 591. https://doi.org/10.1016/j.biocel.2007.03.018.
1044 [93] G. Koraimann, Lytic transglycosylases in macromolecular transport systems of Gram-
1045 negative bacteria, Cellular and Molecular Life Sciences (CMLS). 60 (2003) 2371–2388.
1046 https://doi.org/10.1007/s00018-003-3056-1.
1047 [94] D. Zahrl, M. Wagner, K. Bischof, M. Bayer, B. Zavecz, A. Beranek, C. Ruckenstuhl, G.E.
1048 Zarfel, G. Koraimann, Peptidoglycan degradation by specialized lytic transglycosylases
1049 associated with type III and type IV secretion systems, Microbiology. 151 (2005) 3455–
1050 3467. https://doi.org/10.1099/mic.0.28141-0.
1051 [95] C. Magnabosco, L.-H. Lin, H. Dong, M. Bomberg, W. Ghiorse, H. Stan-Lotter, K.
1052 Pedersen, T.L. Kieft, E. van Heerden, T.C. Onstott, The biomass and biodiversity of the

continental subsurface, Nature Geosci. 11 (2018) 707–717. https://doi.org/10.1038/s41561-018-0221-6.

[96] S.M. Sowell, L.J. Wilhelm, A.D. Norbeck, M.S. Lipton, C.D. Nicora, D.F. Barofsky, C.A. Carlson, R.D. Smith, S.J. Giovanonni, Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea, ISME J. 3 (2009) 93–105. https://doi.org/10.1038/ismej.2008.83.

[97] 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.

[98] C. Winter, T. Bouvier, M.G. Weinbauer, T.F. Thingstad, Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: the "Killing the Winner" Hypothesis Revisited, Microbiol Mol Biol Rev. 74 (2010) 42–57. https://doi.org/10.1128/MMBR.00034-09.

[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).