Mapping the soil microbiome functions shaping wetland methane emissions

Authors: Angela M. Oliverio[1,2+], Adrienne B. Narrowe[1+], Jorge A. Villa[3], Christian Rinke[4], David W. Hoyt[5], Pengfei Liu[1,6], Bridget B. McGivern[1], Emily K. Bechtold[1], Jared B. Ellenbogen[1], Rebecca A. Daly[1], Garrett J. Smith[7], Jordan C. Angle[7], Rory M. Flynn[1], Andrew P. Freiburger[8], Katherine B. Louie[9,10], Brooke Stemple[1], Trent R. Northen[9,10], Christopher Henry[8], Christopher S. Miller[11], Timothy H. Morin[12], Gil Bohrer[13], Mikayla A. Borton[1*], Kelly C. Wrighton[1*]

[1] Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, 80521
[2] Department of Biology, Syracuse University, Syracuse, NY, 13210
[3] School of Geosciences, University of Louisiana at Lafayette, Lafayette, LA 70503
[4] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD, Australia
[5] Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, 99352
[6] Center for Pan-third Pole Environment, Lanzhou University, Lanzhou 730000, China
[7] Department of Microbiology, The Ohio State University, Columbus, OH 43210
[8] Data Science and Learning, Argonne National Laboratory, Lemont, IL 60439
[9] Lawrence Berkeley National Lab, Berkeley, CA, 94720
[10] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720
[11] Department of Integrative Biology, University of Colorado Denver, Denver, CO, 80217
[12] Department of Environmental Resources Engineering, State University of New York College of Environmental Sciences and Forestry, Syracuse, New York, 13210
[13] Department of Civil, Environmental & Geodetic Engineering, The Ohio State University, Columbus, OH, 43210

[+]**Authors contributed equally to this work:** Angela M. Oliverio and Adrienne B. Narrowe
***Co-Corresponding Authors:** Kelly C. Wrighton and Mikayla A. Borton

Email: wrighton@colostate.edu and mborton@colostate.edu

Keywords: genome, MAG, metagenomics, metatranscriptomics, global change, greenhouse gases, methanogen

Abstract

Accounting for only 8% of Earth's land coverage, freshwater wetlands remain the foremost contributor to global methane emissions. Yet the microorganisms and processes underlying

methane emissions from wetland soils remain poorly understood. Over a five-year period, we surveyed the microbial membership and in situ methane measurements from over 700 samples in one of the most prolific methane-emitting wetlands in the United States. We constructed a catalog of 2,502 metagenome-assembled genomes (MAGs), with nearly half of the 72 bacterial and archaeal phyla sampled containing novel lineages. Integration of these data with 133 soil metatranscriptomes provided a genome-resolved view of the biogeochemical specialization and versatility expressed in wetland soils. Centimeter-scale depth differences best explained patterns of microbial community structure and transcribed functionalities, even more so than land coverage or temporal information. Moreover, while extended flooding restructured soil redox, this perturbation failed to reconfigure the transcriptional profiles of methane cycling microorganisms, contrasting with theoretical expected responses to hydrological perturbations. Co-expression analyses coupled to depth resolved methane measurements exposed the metabolisms and trophic structures most predictive of methane hotspots. This compendium of biogeochemically-classified genomes and their spatiotemporal transcriptional patterns begins to untangle the microbial carbon, energy and nutrient processing contributing to soil methane production.

Introduction

Wetlands are characterized as waterlogged soils, rich in organic matter, with low concentrations of oxygen in the porewater. Consequently, the anoxic microbial decomposition of this soil organic matter leads to the production of biogenic methane (CH_4). As such, wetland soils are the largest and most variable source of this potent greenhouse gas, accounting for nearly a third of annual CH_4 emissions (1–4). Despite this global climate relevance, the microbial communities and metabolic pathways driving the conversion of soil organic carbon to CH_4, along with their environmental controls, remain unresolved. This knowledge gap hinders the incorporation of microbial processes into land surface and biogeochemical models, impeding the accurate management and predictions of greenhouse gas (GHG) fluxes from wetlands (3, 5, 6).

Over the last decade, metagenomic sequencing approaches have described the microbial community membership and function associated with CH_4 production across a range of habitats. While other high CH_4 emitting systems, like wastewater (7), ruminants (8), or subsurface (9), habitats have received considerable genomic scrutiny, wetland soils remain relatively underexplored in terms of microbial genomic sampling. Of the handful of genome-resolved studies from freshwater wetlands, nearly all are focused on organic soils from northern peatlands (10–13). One notable study, from thawing permafrost in Sweden, yielded ~1,500 metagenome assembled genomes (MAGs) (11), serving as the stand alone microbial catalog for all wetland soils. This means the highest CH_4-contributing wetland types, temperate and tropical marshes, have only a few genome-resolved studies to date (14–16). These studies emphasized the roles of a limited number of microorganisms to specific biogeochemical processes (14–17), such as methanotrophy, yet failed to inventory the overall composition, function, and interconnected metabolic pathways driving CH_4 production across diverse wetland gradients.

To bridge this knowledge gap, we undertook a comprehensive sampling within a temperate freshwater marsh, aiming to elucidate the spatiotemporal related variations in the microbiome. This freshwater wetland was chosen due to its exceptionally high CH_4 fluxes, exceeding ten times the median flux observed in wetlands with similar annual temperature ranges (18, 19), and because of its susceptibility to climate-induced flooding events influenced by fluctuating Lake Erie water levels (20). Leveraging more than 5 Tera base pairs (Tb) of sequencing, we reconstructed thousands of microbial genomes and integrated this data with soil metatranscriptomes, geochemistry, and greenhouse gas porewater measurements and fluxes. Through this extensive genome recovery effort, we delineated the expressed membership and metabolisms of the microbial community, shedding light on the biogeochemical contributions of hundreds of understudied lineages, including novel methanogens. This multi-omic dataset provides insights into the encoded and expressed microbial functions, as well as their distribution across environmental gradients. These findings offer new perspectives on the stability and interactions of metabolic guilds driving CH_4 production in temperate, freshwater wetland ecosystems.

Results and Discussion

Sampling microorganisms from wetland soils

As a model wetland we selected the highest methane-emitting wetland in the AmeriFlux network (site ID US-OWC), an investigator network with tailored instrumentation, data processing, and standardized procedures for flux measurements (19, 21–23). The Old Woman Creek wetland (OWC) was instrumented to measure wetland CH_4 at various spatiotemporal scales (**Supplementary Fig. 1**), facilitating a detailed assessment of the microbial processes that

101 impact soil CH_4 dynamics. Based on land coverage types, we parsed the wetland into 10 patches.
102 These patches included 5 land coverage types including emergent vegetation-*Typha* (n=1),
103 emergent vegetation-*Nelumbo* (n=3), standing freshwater or open water (n=3), and temporal mud
104 flats (n=3) (**Fig. 1A**). From selected sites, once per month during sampling campaigns, dialysis
105 peepers measured soil CO_2 and CH_4 concentrations along ~3cm-depth increments up to 30 cm
106 deep. This content was paired to surface chamber flux measurements for CO_2 and CH_4 and site
107 wide fluxes using the eddy covariance and meteorological station. Surrounding the peepers, soil
108 cores were pulled and extruded at 5 cm depth increments down to 25 cm to result in 705 samples
109 for soil geochemical, metabolite, and microbial analysis. In summary, across the five-year
110 sampling campaign (2013-2018), each one of these 10 ecological patch sites was sampled at least
111 52 times for paired microbial, soil geochemistry, flux and porewater GHGs (**Supplementary**
112 **Table S1**). The inventory of measured geochemistry and flux data and their relationships is
113 shown in **Supplementary Fig. S2**.
114 We screened 671 of these collected soil samples with 16S rRNA amplicon sequencing for
115 methanogen membership (**Supplementary Table S2**), identifying 42 samples for deep
116 metagenomic sequencing (**Supplementary Table S1**). We generated up to 300 Gbp of
117 sequencing per sample, or ~3.1 Tbp total metagenomic sequencing (**Supplementary Table S3**).
118 Collectively, this effort doubled the amount of metagenomic sequencing from wetland soils
119 available in public repositories (**Supplementary Fig. 1D**). Given that one of our primary
120 objectives was to increase the recovery of methanogen genomes from soils, multiple assembly
121 and binning methods targeted the recovery of these strains (**Supplementary Fig. 1E**). This
122 approach resulted in over 17,333 draft genomes, which were dereplicated into a final collection

123 of 3,217 genomic representatives, with 2,502 of these constituting high and medium quality
124 metagenome assembled genomes (MAGs) (**Supplementary Fig. S1**).
125 To profile which microbial genomes recruited transcripts and which metabolisms were
126 expressed across sites, depths, seasons, and years, we obtained 2.7 Tbp of metatranscriptome
127 sequencing from 133 of the 67116S rRNA screened soil samples. For more sample details see
128 **Supplementary Table S1** and **Methods**. Nearly 80% of the metatranscriptomes had paired
129 NMR detected metabolites to profile methanogenic substrates across the sample set
130 (**Supplementary Table S4**). This comprehensive multi-omics sampling across wetland
131 gradients, including multiple land coverage types, depths, and times (months, seasons and years),
132 provides the first insights into the microbial structural and functional diversity residing in
133 temperate freshwater wetland soils. Due to the enhanced accessibility across multiple, open
134 platforms we hope this resource will advance wetland microbiome science more broadly (see
135 **Data Availability**).
136
137 **Uncovering thousands of novel and transcribed genomes from wetland soils**
138 We reconstructed MAGs from 60 bacterial (N=2,204 MAGs) and 12 archaeal (n=298
139 MAGs) phyla that included 85 methanogen-related genome representatives. A majority of the
140 MAGs recovered here were designated newly sampled lineages at the genus level or higher
141 including representatives from 1 new phylum (here referred to as New Phylum01_b), 8 classes,
142 40 orders, 93 families, and 639 genera (**Supplementary Table S3**). In addition, hundreds of
143 genomes were from poorly or previously uncharacterized phyla often denoted only by an
144 alphanumeric identifier or placeholder name, such as QMZS01, CSP1-3, and TA06_A (**Fig. 1B**).
145 Across this genome-based collection, the most represented phyla are the archaeal Halobacteriota

146 (n=73 MAGs) and the bacterial Proteobacteria (n=484), Acidobacteriota (n=316 MAGs),
147 Chloroflexota (n=237 MAGs), and Desulfobacterota_H (n=140). The reconstruction of these
148 thousands of genomes, many metabolically uncharacterized, highlights the vast, unsampled
149 microbial diversity residing within wetlands.

150 We next quantified the proportion of this genomic catalog that recruited transcripts,
151 finding that approximately three quarters of the reconstructed genomes (n=1948) recruited
152 transcripts from the 133 metatranscriptomes (**Supplementary Table S4**). We report the
153 maximum transcription for each MAG (**Fig. 1B**) and denote the top 15 most transcriptionally
154 dominant MAGs from each bacterial or archaeal domain. These MAGs with dominant
155 transcription also have high transcriptional occupancy, with a minimum detection in 35% across
156 the 133 metastranscriptome samples **(Fig. 1C).** A significant majority (93%) of these
157 transcriptionally dominant genomes belong to poorly characterized genera as noted by their
158 names, including those MAGs first sampled here (denoted by "new") or as genera defined only
159 by alphanumeric code. In fact, two methanogen MAGs, assigned to the genera
160 *Methanomethylicus* and *Methanoregula,* are the only transcriptionally dominant genomes from
161 lineages with characterized isolates. This broad transcriptional recruitment from hundreds of
162 diverse samples to thousands of genomes in our customized database hints at the large proportion
163 of the soil microbiome that may be metabolically active along wetland gradients.

164 We extended our analysis to assess the relevance of the microbial lineages identified here
165 to other wetland environments. We conducted a comparison of the 16S rRNA genes obtained
166 from the OWC wetland samples with 505 samples collected from 5 different North American
167 wetland sites (**Supplementary Fig. S3**). At a 97% threshold cutoff, OWC ASVs covered 51% of
168 the composition of the external wetlands database, suggesting species level conservation across

169 wetlands. Additionally, we performed a genome-resolved methanogen comparison of our
170 wetland genomes and their transcriptional recruitment to similar data obtained from a thawing
171 Arctic peatland in northern Sweden (11). Many methanogens with detectable transcription from
172 the OWC wetland had closely related genomes detectable in the Arctic bog and fen wetland
173 samples, yet each site had unique strain specific transcriptional profiles (**Supplementary Fig.**
174 **S4**). Taken together, the microbial lineages identified in our study are prevalent within vastly
175 different wetland types and geographic regions, extending the utility of this multi-omics
176 resource. However, these data also indicated site endemic transcriptional profiles, articulating the
177 need for continued expression-based sampling from saturated soils.
178
179 **Connecting transcribed biogeochemical functions to bacterial phylogeny**
180 Another key objective was to use the genome-based expression profiles to resolve the
181 microbial contributions to biogeochemical cycling in these high CH_4-emitting soils. We
182 developed a genome-based rule set for 14 microbial energy metabolisms that impact the
183 biogeochemical transformations of oxygen (aerobic respiration, microaerophilic respiration),
184 methane (methanogenesis, aerobic methanotrophy, anaerobic methanotrophy), sulfur (sulfur
185 oxidization, sulfur reduction), nitrogen (nitrification, nitrogen reducer, nitrogen fixation, and
186 Dissimilatory Nitrate Reduction to Ammonium (DNRA)), iron (iron oxidation, iron reduction),
187 and other metabolisms (phototrophy, obligate fermentation) (**Supplementary Fig. S5)**. Each of
188 the 1,948 MAGs that recruited transcripts was assessed for expressed gene content relating to
189 these 14 biogeochemical attributes. This resulted in 1,136 MAGs that contained sufficient gene
190 transcript recruitment to be assigned to at least one of these 14 biogeochemical guilds
191 (**Supplementary Table S5)**.

192 We first characterized the functional profiles for 30 most transcriptionally dominant and
193 prevalent MAGs shown in **Fig. 1C**. For the Archaea, 3 of the 12 transcriptionally dominant
194 genera expressed genes for methanogenesis, including genus UBA472 from the
195 Methanomassiliicoccales and the 2 named methanogens. The remaining 9 genera from phylum
196 Asgardarchaeota (JACAEL01), Thermoproteota (PALSA-986, UBA11855, UBA8941,
197 DTJL01), and Thermoplasmatota (newGenus37_a, newGenus38_a, newGenus40_a) encoded
198 and expressed genes for fermentative carbon processing. For the bacterial counterparts, the
199 transcribed metabolic roles were more diverse, even within genera. The genera in the phylum
200 Bacteroidota (UBA5072, LD21, SR-FBR-E99, newGenus265_b) had the greatest transcriptional
201 contributions to organic carbon decomposition via aerobic, microaerophilic, and denitrifying
202 pathways. We also found transcriptional evidence for iron reduction in Myxococcota
203 (newGenus410_b), Acidobacteriota (newGenus341_b) and Nitrospirota (GW-Nitrospira-1);
204 nitrate reduction and DNRA in Chloroflexota (9FT-COMBO-55-16) and Desulfobacterota
205 (DSXZ01); and fermentative organic carbon processing in Firmicutes (JAAYHS01). Our multi-
206 omics framework reveals how these previously enigmatic lineages transcribe genes that could
207 contribute to wetland carbon, iron, and nitrogen biogeochemical cycling.
208 We next scaled this classification of biogeochemical guilds to all the MAGs. To
209 understand how bacterial phyla more broadly contribute to biogeochemical cycling, we
210 overlayed these encoded and transcribed traits on a bacterial phylogenetic tree (**Fig. 2A**). Certain
211 phyla, such as Patesibacteria and Omnitrophota, are characterized by a solid block of pink color
212 in the encoded and transcribed rings, indicating their specialization as obligate fermenters. In
213 contrast, phyla like the Proteobacteria exhibit a wide array of colors in the encoded and
214 transcribed rings, signifying their trait multifunctionality at the phylum level. This trait-

215 phylogenetic overlay reveals disproportional contributions from phyla to nutrient and redox
216 transformations. For instance, MAGs transcribing genes for sulfur reduction are enriched in the
217 Desulforobacterota, iron reduction are prevalent in the Nitrospirota and Myxococcota, aerobic
218 respiration are prevalent across the Verrucomicrobiota, and nitrogen reduction and sulfur
219 oxidation are enriched in some lineages of the Proteobacteria. This representation also highlights
220 keystone metabolisms, like nitrogen fixation and nitrification, which were identified as
221 transcribed functionalities in a small fraction (<5%) of phylogenetically diverse bacterial MAGs.
222 This trait-based phylogenetic mapping underscores the diverse, expressed roles that bacteria play
223 in shaping wetland biogeochemistry.

224 We were surprised to also observe in the bacterial tree that while aerobic and
225 microaerophilic metabolisms were widely encoded across bacterial lineages these were not
226 widely transcribed. To evaluate this further in both the bacterial and archaeal MAGs, we
227 compared the dominant encoded metabolic traits (i.e., detected in the greatest number of
228 genomes) to the most dominant transcribed traits. The transcribed functionalities did not follow
229 the encoded (**Fig. 2BC**). For example, the capacity for microaerophilic and aerobic respiration
230 was detected in the largest percentage of genomes (53% and 51%), but only detected in 8% and
231 6% of genomes in our transcriptomes respectively. In contrast, transcription for genes enabling
232 anaerobic metabolisms like obligate fermentation, iron reduction, and nitrogen reduction were
233 detected more so than aerobic metabolisms. This enrichment for anaerobic metabolisms in the
234 transcript signal is not entirely surprising given that most soils in our data set have pore-water
235 dissolved oxygen (DO) concentrations below the 1uM detection limit (**Supplementary Table**
236 **S4**). These findings showcase the metabolic transcribed tuning that can be differentiated from
237 encoded potential using our approaches.

238 While our phylogenetic analysis hinted at a multifunctionality at broad phylum levels
239 (**Fig. 2A**), we wanted to examine this attribute across Bacterial and Archaea at a MAG resolved
240 scale. For every MAG with transcribed trait content, we graphed the number of present traits
241 versus the number of transcribed traits. From this we defined a spectrum of metabolic
242 specialization to versatility. Focusing on the extreme ends of this continuum, "active specialists"
243 encoded 2 or fewer metabolic functions and transcribed these traits (**Fig. 3AB**, group 1). Of the
244 1,136 trait classifiable MAGs, we were surprised to find that 40% were assigned to this group,
245 with obligate fermenters (e.g., the archaeal Bathyarchaeia and bacterial Omnitrophota and
246 QNDG01) and methanogens (e.g., *Methanomethylicus*, *Methanothri*x, and *Methanoregula* in the
247 Halobacteriota and Thermoplasmatota) constituting most functionalities. Consistent with
248 specialized nature, these lineages also had a high proportion of the genes in a genome transcribed
249 and low per genome transcriptional variability at the phylum level (**Fig. 3C**, **Supplementary**
250 **Fig. 6**).
251 On the other end of the spectrum, we defined metabolic versatility as having the genomic
252 capacity for 6 metabolic traits. We then considered what proportion of these functions were
253 transcribed denoting "versatile with multifunctional transcription" as members that transcribed at
254 least 4 of these encoded metabolic traits (**Figure 3AB**, group 2). As might be expected, this
255 group was small, containing 12 'versatile with multifunctional transcription' genomes that were
256 predominantly members of Proteobacteria (e.g., Burkholderiales and Chromatiales) and
257 Desulfobacterota_F. The third and final group was "versatile with specialized transcription",
258 which despite encoding 6 or more functional traits transcribed only 2 or less across the 133
259 samples. This category included 85 MAGs composed of members in the Proteobacteria,
260 Nitrospirota (class UBA9217), Myxococcota, and Acidobacteriota.

261 There is growing interest in determining microbial metabolic strategies dispersed across
262 microbiomes, as this knowledge could inform ecosystem resilience and stability (24).
263 Traditionally, metabolic versatility and specialization were measured using phenotypic assays,
264 and more recently, inferred from genome-based methods (24, 25). Our use of soil-based
265 metatranscript data across thousands of genomes and numerous spatial and temporal samples
266 adds a new perspective to metabolic strategy assignments. At the wetland level, we showed that
267 the most encoded metabolisms were not the most transcribed (**Fig. 2**). At the genome level, we
268 showed it is more common that versatile genomes express only a few metabolisms across
269 hundreds of spatial and temporal samples, indicating phenotypic environmental specialization
270 despite genotypic multifunctionality (**Fig. 3**). To ultimately improve CH_4 predictions, these
271 results highlight the need for soil microbiome technologies that capture dynamic activity or
272 related proxy measurements, rather than static microbiome response variables.
273
274 **Identifying the dominant drivers shaping wetland microbiome structure and function**
275 We hypothesized that geospatial and temporal factors would explain microbial
276 community structuring. To test how different wetland environments influenced the soil
277 microbiome, we evaluated the impact of 17 spatial, temporal, and geochemical factors across
278 three levels of biological organization: (i) phylotype membership, (ii) transcriptionally 'active'
279 genome membership, and (iii) expressed functional guilds (**Fig. 4**). Our analyses revealed that,
280 regardless of the microbiome attribute measured, depth and depth-related geochemistry (such as
281 soil Fe (II) and cation exchange capacity, CEC) were the best predictors of wetland soil
282 microbiome structure and function, more so than differences from vastly different land coverages
283 or time. These environmental attributes played a more direct role in shaping expressed functional

284 guilds, explaining 40 to 50% of this variation (**Supplementary Fig. S7**). Overall, this
285 information can inform wetland microbiome sampling priorities, showing the value of adding
286 more depth-resolved measurements and affirming collection of high-priority geochemical
287 variables, such as CEC, Fe (II), pH, and acetate.

288 We next identified the bacterial and archaeal family-level lineages whose transcription
289 was coordinated to these environmental factors (**Fig. 5**). Laterally across wetland land coverage
290 types, Actinobacteriota transcription corresponded to open water sites, while Nitrospirota family
291 UBA9217 transcription corresponded to temporally exposed mudflats. Investigating the
292 metabolic traits expressed within this mud-associated Nitrospirota clade revealed 38 transcribed
293 MAGs, with iron and nitrogen reduction being dominant expressed metabolisms (**Fig. 5B**).
294 Along the depth defined redox gradient, iron reducers (e.g., Zixibacteria, New_Phylum01) and
295 obligate fermenters (e.g., Lokiarchaeia, Bathyarchaea) had transcription patterns corresponding
296 to features in deeper soils, while expression of phototrophy (Cyanobacteria) and aerobic respires
297 (e.g., Verrucomicrobiae, Desulfobaccia) had transcription patterns corresponding to features in
298 surface soils. Importantly, these findings show that while certain family-level lineages are
299 associated with distinct wetland habitats, metabolisms are still widely dispersed with transcripts
300 for aerobic metabolism detected in deeper sediments and iron reduction, methanogenesis and
301 fermentation detected in surface soils (**Fig. 5B**). These data represent a first step towards using
302 gene and genome transcript recruitment to understand microbial taxonomic distributions and
303 functional redundancy across both small-scale depth changes and larger lateral distances in
304 wetland gradients.
305
306

307 **Uncovering newly defined archaeal roles in plant litter decomposition**

308 A valuable component of our sampling was the broad recovery of archaeal genomes
309 (n=298 MAGs). To our knowledge this is one of the largest environmental genomic catalogs for
310 this domain, exceeding the archaeal sampling in MAG compendiums from groundwater (24
311 genomes) (26), permafrost wetland soils (95 genomes) (11), human gut (28 genomes) (27), or
312 rumen (126 genomes) (28). The OWC wetland archaeal genomes spanned 12 phyla with 45
313 families including 98 genera. Paralleling the bacterial novelty, this included 10 new archaeal
314 family level designations (**Fig. 6**). Nearly a third of these archaeal genomes were methanogens
315 (n=85), assigned to 4 methanogen phyla (Crenarchaeota, Euryarchaeota, Halobacteriota,
316 Thermoplasmatota).

317 When considering the functional roles of Archaea in wetlands, these microorganisms are
318 regarded for their functional roles as methanogens or metabolizing other single carbon (C1)
319 compounds (16, 29–31). Here we derived a gene-based rule set to assign these genomes to
320 carbon decomposition guilds of (i) polymer hydrolysis, (ii) sugar oxidation, (iii) organic acids
321 and nitrogen conversions, and (iv) C1 metabolisms (**Supplementary Table S7**). To unlock the
322 soil carbon cycle, microbial enzymes hydrolyze polymeric compounds like polyphenols (lignin),
323 hemicellulose, and cellulose abundant in plant litter (11, 32). We show phylogenetically diverse
324 archaeal lineages expressed genes for hydrolyzing these plant polymeric constituents **(Fig. 6)**.
325 Depolymerization results in the second trophic group, yielding sugars and other oligomers or
326 monomers. Lineages like Thorarchaeota and UBA233 transcribed genes to use 6 or more
327 substrates that span both polymeric and sugar trophic groups. We also identified sugar utilizing
328 specialist archaea (members of the TCS64 and a new OWC7 family within the Thermoplasmata)
329 that expressed genes for utilizing at least 3 sugar substrates. In support of these expressed sugar

330 metabolisms, raffinose, sucrose, and maltose like compounds were prevalent across the
331 metabolome data (**Supplementary Table S8-S9**). The third and final trophic guild before
332 methane production, is downstream sugar processing, which yields small molecular weight
333 compounds like organic acids (e.g., acetate, butyrate, lactate) and methylated (e.g., methylamine,
334 methanol) compounds. Some of these are methanogenic substrates in their own right, while
335 others can be further fermented, typically syntrophically, to yield methanogenic substrates. This
336 latter guild had the greatest transcriptional prevalence across the archaeal genomes, indicating
337 the critical role these archaea, most of them diagnosed as obligate fermenters, could play in cross
338 feeding methanogens.

339 Our transcription-based findings revealed Archaea were not latent members of a soil
340 seedbank (33), but instead genomes from 34 of 45 archaeal families were expressed in 75% of
341 the metatranscriptomes, many across all depth gradients. Most of these were assigned roles as
342 obligate fermenters or methanogens, demonstrating their importance to the carbon cycle in
343 saturated, predominantly anoxic soils. As more genome-resolved multi-omics analyses become
344 available from soils and sediments, it will be interesting to note if this broadened scope of
345 archaeal influence across the major carbon decomposition guilds is unique to soil organic matter
346 processing in saturated wetland soils, where aerobic respiration by bacterial (**Fig. 2**) and fungal
347 counterparts may not be as competitive (34, 35).

348

349 **Methane cycling guilds reside in depth resolved, temporally stable 'neighborhoods'**

350 The 85 methanogen MAGs belonged to 4 phyla and spanned 8 families with 25 genera,
351 15 of the latter were new genera first identified here (**Fig. 7A**). One of these families was a new
352 methanogen family (NewFamily01_m) in the Methanomicrobiales. Within this family, we

353 reconstructed 25 MAGs, which resolved to 4 new genera. Lastly beyond canonical methanogens,
354 we also recovered genomes from presumed non-methanogenic, *mcrA*-containing taxa such as the
355 anaerobic methanotrophic *Methanoperedens* genus and the alkane-oxidizing Helarchaeales. Prior
356 to this effort, the largest wetland MAG database contained 21 methanogen representatives with
357 4-fold less genus level diversity (11). Despite their direct climate feedback role, it is astonishing
358 that we continue to discover multiple new, highly active methanogens from a single wetland.

359 Beyond taxonomic novelty, metabolic reconstruction with paired transcriptional profiles
360 offered new insights into the CH_4 cycle (**Fig. 7**). First, acetoclastic methanogenesis was confined
361 exclusively to *Methanothrix* and closely related members (19 MAGs). Based on (36), we posited
362 that these obligate acetoclastic methanogens outcompeted the facultative acetoclast
363 *Methanosarcina* because substrate acquisition strategies were better adapted to the low (<1mM)
364 acetate concentrations found in these wetland soils (mean 45 μM) (**Supplementary Fig. S8**).
365 Across the methanogens, hydrogenotrophy was the most prevalently transcribed pathway, with
366 all 25 members of the new family expressing this capacity, while one of these new genera (new
367 genus03) also expressed genes for using methylsulfonated compounds. We showed
368 methylotrophy was expressed in historically considered hydrogenotrophic lineages
369 (Methanoregulaceae) and new genera in obligate methylotrophic families like the
370 Methanomassiliicoccales (new genus 8, 9, 10) and Methanomethylicales (new genus 12) (**Fig.
371 7A**). While it is often cited that wetland CH_4 is derived exclusively from acetoclastic and
372 hydrogenotrophic methanogens (3), the discovery that nearly a quarter of these wetland
373 methanogens expressed the capacity for methylotrophic methanogenesis indicate that organic
374 nitrogen, sulfur, and oxygenated methylated compounds may be an underappreciated source of
375 CH_4 in wetlands (16, 31).

376 To ascertain changes in methanogen transcriptional profiles across spatial-temporal
377 gradients, from a flooded mud-flat we collected triplicate metatranscript data from 3 depths,
378 monthly over a 3-month summer field season. Based on genome-based transcription patterns the
379 heatmap exposed methanogen and methanotroph spatiotemporal niches and possible depth-
380 defined ecotypes (**Supplementary Fig. S9**). We uncovered six core methanogens, from diverse
381 functional and phylogenetic backgrounds including a MAG from the new OWC family, with
382 high levels of mean transcription across all soil depths and timepoints. The remaining CH_4
383 cycling members were localized to specific soil depths, forming co-expressing clades that
384 persisted across this summer season. We report a greater proportional enrichment of acetoclastic
385 and aerobic methanotrophs in surface soils, while the middle and deepest soil layers included the
386 anaerobic methane-oxidizing *Methanoperedens* genome and proportionally more
387 hydrogenotrophic and methyl-utilizing methanogens. Reflecting these latter methanogens
388 requirements for hydrogen, we observed increased formate concentrations from deeper (10-30
389 cm) depths across our mud, plant, and open water land coverage patch types (**Supplementary**
390 **Fig. S8B**). Here we provide a highly resolved spatial and temporal view of CH_4 cycling
391 transcription patterns in soils, opening up new research avenues regarding the assembly and
392 persistence of depth-stratified methane neighborhoods, composed of many closely related strains
393 with a high-degree of functional redundancy.
394 Given the seasonal stability across summer 2018, we next compared the multi-omics data
395 collected from this site in August 2018 to parallel samples from the same mud flat in August
396 2015. During this three-year period, the wetland flooded and the exposed mudflat from 2015 was
397 in 2018 beneath 5.3 ft of overlying freshwater. Based on existing theory about the redox
398 sensitivity of methanogens and their responses to hydrological perturbations (37), with flooding

399 we hypothesized to observe decreased dissolved oxygen (DO) concentrations and increased
400 methanogen expression broadly across the community, but especially in the surface most soils.
401 Matching flooding expectations, DO decreased to non-detectable levels in surface soils
402 subsequent to flooding. In contrast to expectations, 90% of the methanogen and aerobic
403 methanotroph (Methylococcaceae, Methylomonadaceae) MAGs experienced no significant
404 change in mean genome transcription in the flooding restructured surface soils (**Fig. 7B)**. Worth
405 noting, a few taxa were flooding-responsive, like the methylotrophic *Methanomethylicus*, which
406 was one of the 30 most transcribed archaeal MAGs (**Fig. 1C**). Supporting this lack of drastic
407 restructuring in the CH_4 cycling microbiome, the concentrations of methanogenic substrates like
408 formate, methanol, and acetate did not change with flooding (**Supplementary Fig. S8**). This
409 flooding resistance and or resilience of the surface CH_4 cycling members may be due to the fact
410 that methanogens in these routinely submerged and drained soils could be already sheltered
411 within anoxic microsites, such as biofilms or soil aggregates (15, 38).
412 Taken together our CH_4 cycling genome-based transcriptional profiles imply a seasonal
413 and multi-year stability across the soil profile, despite hydrological and redox perturbations.
414 While we previously reported enhanced CH_4 production in oxygenated soils from this site in
415 2015 (15), context from our new flooding data further reinforced that methanogen genome-wide
416 transcription is not responsive to soil oxygen concentrations. These findings contribute to a
417 growing body of research that challenges the widely held paradigm that oxygen is a master
418 controller on soil CH_4 production (39–42). Our research has important ramifications for global
419 biogeochemical models (6), as current simulations regulate CH_4 production in wetland soil layers
420 based on oxygen concentrations.
421

422 **Co-expression networks reveal key lineages modulating soil greenhouse gasses**

423 We implemented a workflow to integrate the data collected across this manuscript to
424 render the carbon and energy trophic network predictive of soil CH_4 concentrations
425 (**Supplementary Fig. S10**). First, weighted gene co-expression identified 7 modules composed
426 of genomes with shared transcriptional patterns across our 133 metatranscriptome samples (**Fig.**
427 **8A**). Two of these modules predicted soil porewater CH_4 concentrations, with the surface-soil
428 associated turquoise module negatively correlated and the deeper-soil associated brown module
429 positively correlated to soil CH_4 concentrations (**Fig. 8A**). These surface and deep modules
430 contained 551 and 372 genomes, respectively. The top microbial genomes with the greatest
431 importance for CH_4 prediction and their transcribed biogeochemical traits are also noted (**Fig.**
432 **8CD**). Distilling the transcribed carbon and energy functional gene content of these predictive
433 genera, we mapped the trophic configurations that contribute to CH_4 concentrations in these
434 depth-defined compartments (**Fig. 8E**).

435 Consistent with our taxonomic-based analyses (e.g., **Fig. 5**), we found key
436 biogeochemical processes (such as iron reduction, microaerophilic respiration, and methane
437 production) were transcribed across depths in these predictive modules (**Fig. 8CD**). Yet a
438 different set of microorganisms were often responsible for mediating these processes, with the
439 exception of Zikibacteria and the Syntrophales, which were important predictors in both
440 compartments. For example, in both compartments microaerophilic respiratory organisms were
441 responsible for accessing polymeric carbon, catalyzed by Bacteroidales in the surface soils, but
442 Anaerolineales and Sedimentisphaerales in the deep soils.

443 Alternatively, the turquoise surface module also included genomes with functionalities
444 not represented in the deep module. This includes genomes transcribing genes for nitrate

445 reduction and low affinity cytochrome oxidases, which are adapted to higher oxygen
446 concentrations than the microaerophilic respiration genes in the deeper module (**Fig. 8E**).
447 Likewise, the turquoise surface module, but not the deeper module, contained transcriptional
448 evidence for aerobic methanotrophy (Methylococcaceae UBA6136 and Methylomonadaceae
449 KS41). Taken together, we infer CH_4 consumption by these lineages in the surface, more
450 oxygenated soils likely explained the negative correlation between the turquoise module and soil
451 CH_4 concentrations (**Fig. 8B**).
452 Our transcriptome-derived, process-based observations represent an important first step
453 in uncovering the carbon and energy trophic structure in wetland soils. Soil CH_4 concentrations
454 are an emergent property of the wetland microbiome, meaning CH_4 may not be accurately
455 predicted by the result of individual microorganisms or their activities (e.g. methanogens), but
456 arise from complex interactions and collective co-expression behaviors in stable, depth-resolved
457 transcriptional networks. Along these lines, the greatest predictors of CH_4 were not always
458 methanogens, especially in the deeper soils where iron reducers (*Zixibacteria*, *Eisenbacteria*,
459 *Geobacter*) and obligate fermenters (Amincenantales) likely cross-fed hydrogenotrophic
460 methanogens (*Methanoregula*) (**Fig. 8D**). These findings articulate the need for higher
461 throughput tools to study microbial interactions in light of ecosystem-level processes.
462

463 **Conclusion**

464 In this study, we employed the most comprehensive multi-omics sampling of wetland
465 soils to date, allowing us an unprecedented view of the microbiome structural and functional
466 attributes relevant to biogeochemical cycling. The elevated sequence depth and targeted
467 assemblies allowed us to capture more genomic diversity, unearthing a new phylum, 8 new

468 classes, 40 new orders, 93 new families, and 639 new genera (**Fig. 1B**), many of which included
469 novel archaea (**Fig. 6**), including methanogens (**Fig. 7A**). More importantly, we show that these
470 uncharacterized lineages were transcriptionally dominant and prevalent across 133 spatial and
471 temporally heterogenous wetland samples (**Fig. 1C**).

472 Based on the microbial seed bank theory, we assumed that a majority of the recovered
473 microbial genomes would be dormant and lack detectable transcript recruitment. Instead, we
474 found that 78% of the 2,502 genomes recruited transcripts, with 60% of these MAGs expressing
475 sufficient genome content to permit metabolic assignments to 14 traits associated with wetland
476 biogeochemistry. Moreover, the average MAG mean transcriptional occupancy was 59%,
477 signifying active members were maintained across samples that span diverse wetland gradients.
478 Instead of a seed bank, our data indicate a vast majority of this phylogenetically and functionally
479 diverse microbiome was not-dormant, at least based on transcriptional inferences.

480 We next considered functional redundancy as one of the main contributors to the
481 temporal stability observed across the wetland metatranscriptome samples. While traditionally
482 measured by inferences from phylotype data or estimated by the functional potential encoded in
483 (meta)genomes (43), here we instead evaluated the transcription of genomes and functional
484 genes to map these profiles across relevant wetland gradients. We observed that depth related
485 features best explained patterns of microbial community structure and transcribed functionalities
486 (**Figs. 4,5**). Further microbiome transcriptional patterns were stable across these depth layers
487 across multiple years and seasonally (**Figs. 7, 8**).

488 We also report that within a depth layer similar biogeochemical functionalities were often
489 transcribed by many closely related genomes, e.g. within the same genus (e.g., methanogens,
490 **Figs. 7, 8C**), as well as between more distantly related genomes (e.g., iron reduction by different

phyla, **Fig. 8D**). For the functionalities transcribed in both depth layers, these were often mediated by more phylogenetically distinct lineages (**Figs. 5,7,8E**), but this too was not always the case (e.g., *Zikibacteria*, **Fig. 8CD**). Taken together, our data indicate that functional redundancy is deeply entrenched in wetland microbiomes and likely critical for the stability of biogeochemical processes in these highly dynamic environments. This study underscores the value of integrating both genomic and transcriptomic data to uncover the functional dynamics of belowground microbial systems.

Here we used multi-omics measurements to untangle the metabolic processes in wetland soils. This approach can move the field of soil microbiomes beyond the historical black box of inputs and outputs, towards discovery of the compositional and functional microbiome-based responses to environmental gradients. To begin to distill this metatranscriptomic content, we devised gene-based rule sets for microbial energy metabolisms and for carbon trophic guilds. Using this metabolic trait framework, we illuminated new roles for archaeal gene expression in carbon decomposition, the diverse substrate utilization patterns and resistance to flooding exhibited by methanogens, and the stable, coordinated gene transcription-based behavior of surface and deep guilds predictive of soil CH_4 concentrations. These insights, while complex, offer a more nuanced view of the microorganisms, functional attributes, and dynamics that warrant representation and ongoing investigation in wetland soils.

References

1. S. C. Neubauer, J. P. Megonigal, Moving Beyond Global Warming Potentials to Quantify the Climatic Role of Ecosystems. Ecosystems 18, 1000–1013 (2015).
2. R. B. Jackson, et al., Increasing anthropogenic methane emissions arise equally from agricultural and fossil fuel sources. Environ. Res. Lett. 15, 071002 (2020).
3. S. D. Bridgham, H. Cadillo-Quiroz, J. K. Keller, Q. Zhuang, Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. Glob Chang Biol 19, 1325–1346 (2013).
4. K. Calvin, et al., "IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.," First (Intergovernmental Panel on Climate Change (IPCC), 2023).
5. R. Conrad, Importance of hydrogenotrophic, aceticlastic and methylotrophic methanogenesis for methane production in terrestrial, aquatic and other anoxic environments: A mini review. Pedosphere 30, 25–39 (2020).
6. W. J. Riley, et al., Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM4Me, a methane biogeochemistry model integrated in CESM. Biogeosciences 8, 1925–1953 (2011).
7. B. Kieft, et al., Genome-resolved correlation mapping links microbial community structure to metabolic interactions driving methane production from wastewater. Nat Commun 14, 5380 (2023).
8. L. M. Solden, et al., Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. Nat Microbiol 3, 1274–1284 (2018).

9. M. A. Borton, et al., Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. Proceedings of the National Academy of Sciences 115, E6585–E6594 (2018).

10. N. Taş, et al., Landscape topography structures the soil microbiome in arctic polygonal tundra. Nat Commun 9, 777 (2018).

11. B. J. Woodcroft, et al., Genome-centric view of carbon processing in thawing permafrost. Nature 560, 49–54 (2018).

12. J. Hultman, et al., Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature 521, 208–212 (2015).

13. R. M. Wilson, et al., Soil metabolome response to whole-ecosystem warming at the Spruce and Peatland Responses under Changing Environments experiment. Proceedings of the National Academy of Sciences 118, e2004192118 (2021).

14. G. J. Smith, et al., Members of the Genus Methylobacter Are Inferred To Account for the Majority of Aerobic Methane Oxidation in Oxic Soils from a Freshwater Wetland. mBio 9, e00815-18.

15. J. C. Angle, et al., Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions. Nat Commun 8, 1567 (2017).

16. A. B. Narrowe, et al., Uncovering the Diversity and Activity of Methylotrophic Methanogens in Freshwater Wetland Soils. mSystems 4, e00320-19 (2019).

17. M. Bahram, et al., Structure and function of the soil microbiome underlying N2O emissions from global wetlands. Nat Commun 13, 1430 (2022).

18. G. Bohrer, et al., Methane and CO2 chamber fluxes and porewater concentrations US-OWC Ameriflux wetland site, 2015-2018. Environmental System Science Data Infrastructure for

a Virtual Ecosystem (ESS-DIVE) (United States); AmeriFlux Management Project. https://doi.org/10.15485/1568865. Deposited 1 October 2019.

19. G. Bohrer, J. Kerns, AmeriFlux BASE US-OWC Old Woman Creek. AmeriFlux AMP. https://doi.org/10.17190/AMF/1418679. Deposited 2023.

20. T. H. Morin, et al., Water level changes in Lake Erie drive 21st century CO2 and CH4 fluxes from a coastal temperate wetland. Science of The Total Environment 821, 153087 (2022).

21. H. Chu, et al., AmeriFlux BASE data pipeline to support network growth and data sharing. Sci Data 10, 614 (2023).

22. G. Pastorello, et al., The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci Data 7, 225 (2020).

23. K. B. Delwiche, et al., FLUXNET-CH4: a global, multi-ecosystem dataset and analysis of methane seasonality from freshwater wetlands. Earth System Science Data 13, 3607–3689 (2021).

24. Y.-J. Chen, et al., Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. ISME J 15, 2986–3004 (2021).

25. B. R. Bochner, Global phenotypic characterization of bacteria. FEMS Microbiol Rev 33, 191–205 (2009).

26. K. Anantharaman, et al., Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7, 13219 (2016).

27. A. Almeida, et al., A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39, 105–114 (2021).

28. R. D. Stewart, et al., Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol 37, 953–961 (2019).

29. B. J. Baker, et al., Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. Nat Microbiol 1, 1–9 (2016).

30. V. De Anda, et al., Brockarchaeota, a novel archaeal phylum with unique and versatile carbon cycling pathways. Nat Commun 12, 2404 (2021).

31. J. B. Ellenbogen, et al., Methylotrophy in the Mire: direct and indirect routes for methane production in thawing permafrost. mSystems 0, e00698-23 (2023).

32. B. B. McGivern, et al., Decrypting bacterial polyphenol metabolism in an anoxic wetland soil. Nat Commun 12, 2466 (2021).

33. S. M. Gibbons, et al., Evidence for a persistent microbial seed bank throughout the global ocean. Proceedings of the National Academy of Sciences 110, 4651–4655 (2013).

34. F. Barbi, et al., Fungal ecological strategies reflected in gene transcription - a case study of two litter decomposers. Environ Microbiol 22, 1089–1103 (2020).

35. L. Žifčáková, P. Baldrian, Fungal polysaccharide monooxygenases: New players in the decomposition of cellulose. Fungal Ecology 5, 481–489 (2012).

36. M. S. M. Jetten, A. J. M. Stams, A. J. B. Zehnder, Methanogenesis from acetate: a comparison of the acetate metabolism in Methanothrix soehngenii and Methanosarcina spp. FEMS Microbiology Letters 88, 181–197 (1992).

37. R. Conrad, Methane Production in Soil Environments—Anaerobic Biogeochemistry and Microbial Life between Flooding and Desiccation. Microorganisms 8, 881 (2020).

38. M. J. Bogard, et al., Oxic water column methanogenesis as a major component of aquatic CH4 fluxes. Nat Commun 5, 5350 (2014).

39. Y. A. Teh, W. L. Silver, M. E. Conrad, Oxygen effects on methane production and oxidation in humid tropical forest soils. Global Change Biology 11, 1283–1297 (2005).

40. J. L. Wilmoth, et al., The role of oxygen in stimulating methane production in wetlands. Glob Chang Biol 27, 5831–5847 (2021).

41. J. K. Shoemaker, D. P. Schrag, Subsurface characterization of methane production and oxidation from a New Hampshire wetland. Geobiology 8, 234–243 (2010).

42. M. R. Turetsky, et al., A synthesis of methane emissions from 71 northern, temperate, and subtropical wetlands. Global Change Biology 20, 2183–2197 (2014).

43. H. Chen, et al., Functional Redundancy in Soil Microbial Community Based on Metagenomics Across the Globe. Front Microbiol 13, 878978 (2022).

44. E. Hassett, et al., Carbon flux measurements from chambers collected between July to October 2022 at Old Woman Creek, Huron, Ohio. urn:node:ESS_DIVE. https://doi.org/10.15485/2229438. Deposited 2023.

45. J. Villa, G. Bohrer, L. Kinsman-Costello, Y. Ju, Nutrient and carbon concentrations in dated soil cores at US-OWC Ameriflux wetland site (OWC NERR). urn:node:ESS_DIVE. https://doi.org/10.15485/1839288. Deposited 2022.

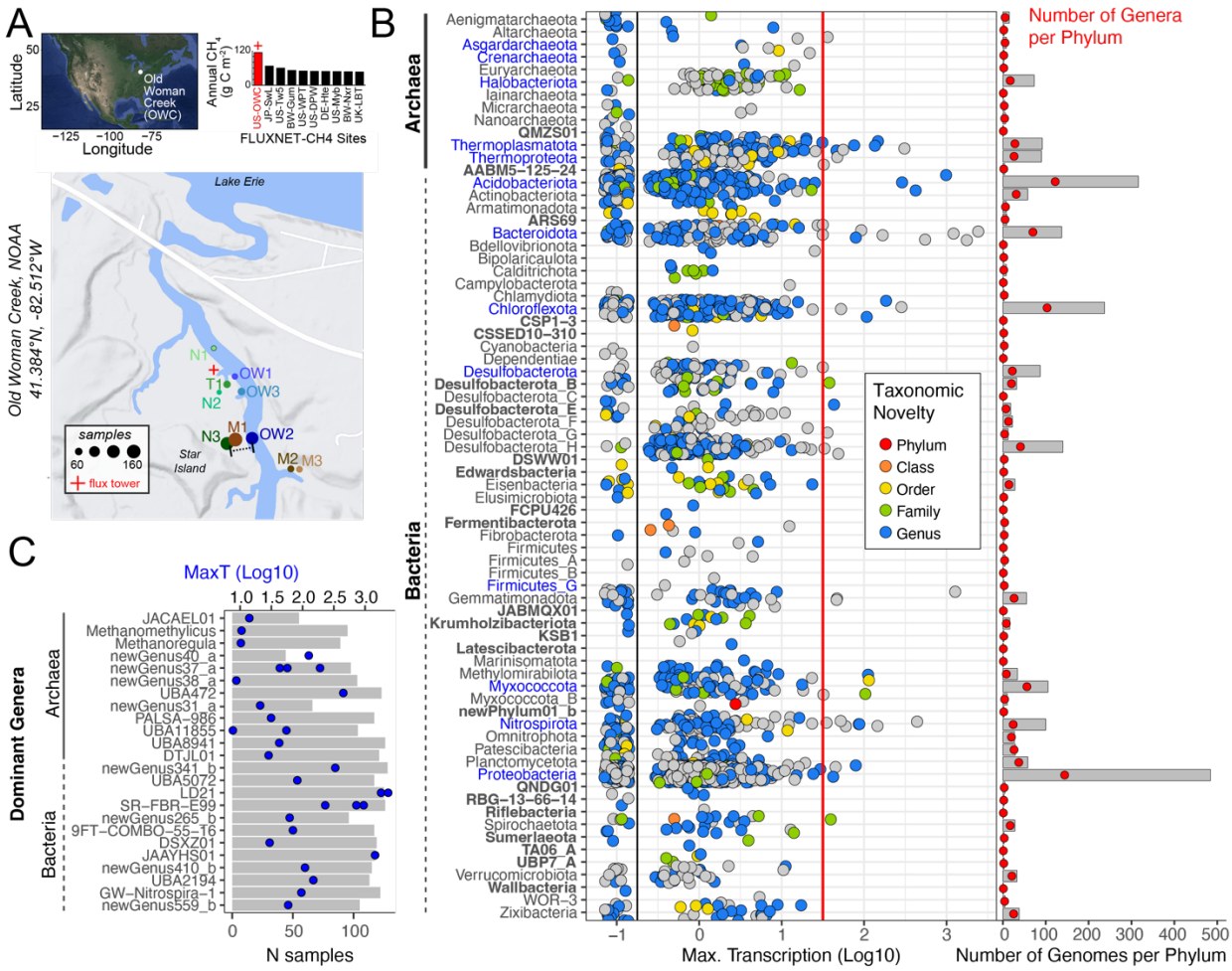


Figure 1. Freshwater wetland soils host thousands of taxonomically diverse MAGs that recruit metatranscripts. (A) Old Woman Creek (OWC) is located on the southern part of Lake Erie and is the highest US methane emitting wetland in the AmeriFlux Network. A site map of the OWC freshwater wetland shows where the 700 samples were collected from 10 major sampling sites that were historically defined as emergent vegetation, either *Nelumbo* (N1-3) or *Typha* (T1), mud cover (M1-3), or location in open water channel (OW1-3). Sites were sampled over multiple years and depths at 5 cm intervals from 0-30 cm, with sampling density proportional to circle size, 52 is the minimum sampling per location (**Supplementary Fig. S1**). **(B)** The taxonomic distribution of the 2,502 MAGs summarized at the phylum level. Bold phylum label indicates a candidate phylum, and blue text indicates the phylum contains a top 10 dominant genus (shown in **1C**). Each point represents a MAG and is colored by taxonomic novelty. Points are distributed by their log10 maximum transcription across the 133 metatranscriptome samples. MAGs to the right of the black line were considered positive for sufficient transcript recruitment (N = 1,948). MAGs to the right of the red line had the greatest transcription (log10 max. transcription > 1.5). The gray bar chart to the right summarizes the number of genomes per phylum recovered, and red points overlaid represent the number of genera recovered within each phylum, using the same x-axis. **(C)** The transcriptionally most dominant archaeal and bacterial MAGs (top 15 for each domain) from 1B, grouped by genus. Gray bar indicates metatranscriptome prevalence (gray bar), or the number (N) of the 133 samples where metatranscripts were recruited to the MAG for that genus. The axis is shown on bottom (N samples). The blue dot indicates the maximum transcription from **1B**, with the axis (MaxT, Log10) shown at the top of the graph.

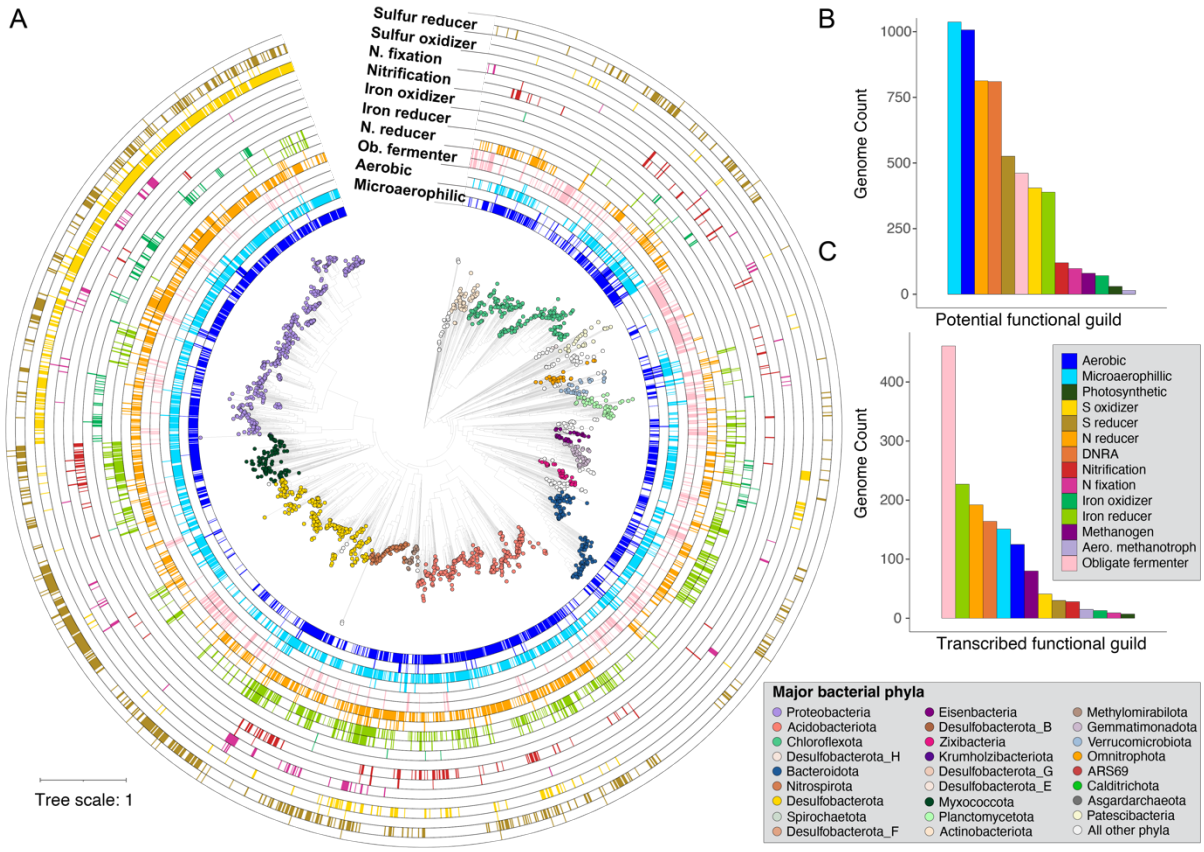


Figure 2. Microbial genomes from freshwater wetlands are phylogenetically and metabolically diverse. (A) The phylogenetic distribution of 2,204 bacterial MAGs and their corresponding potential (metagenome) and expressed (metatranscriptome) metabolic diversity. The phylogenetic tree is constructed from the 2,204 bacterial MAGs based on a subset of 30 genes from the PhyEco database. Nodes are colored by phylum level taxonomic assignment. Rings represent metabolic diversity at the trait level for each MAG. For each color, the inner ring represents potential (metagenome) metabolism, and the outer ring represents the expressed (metatranscriptome) metabolism. Bar plots summarize the genome counts across functional traits for genomic potential **(B)** and transcription **(C)**, respectively (n = 1,948 and includes bacterial and archaeal genomes).

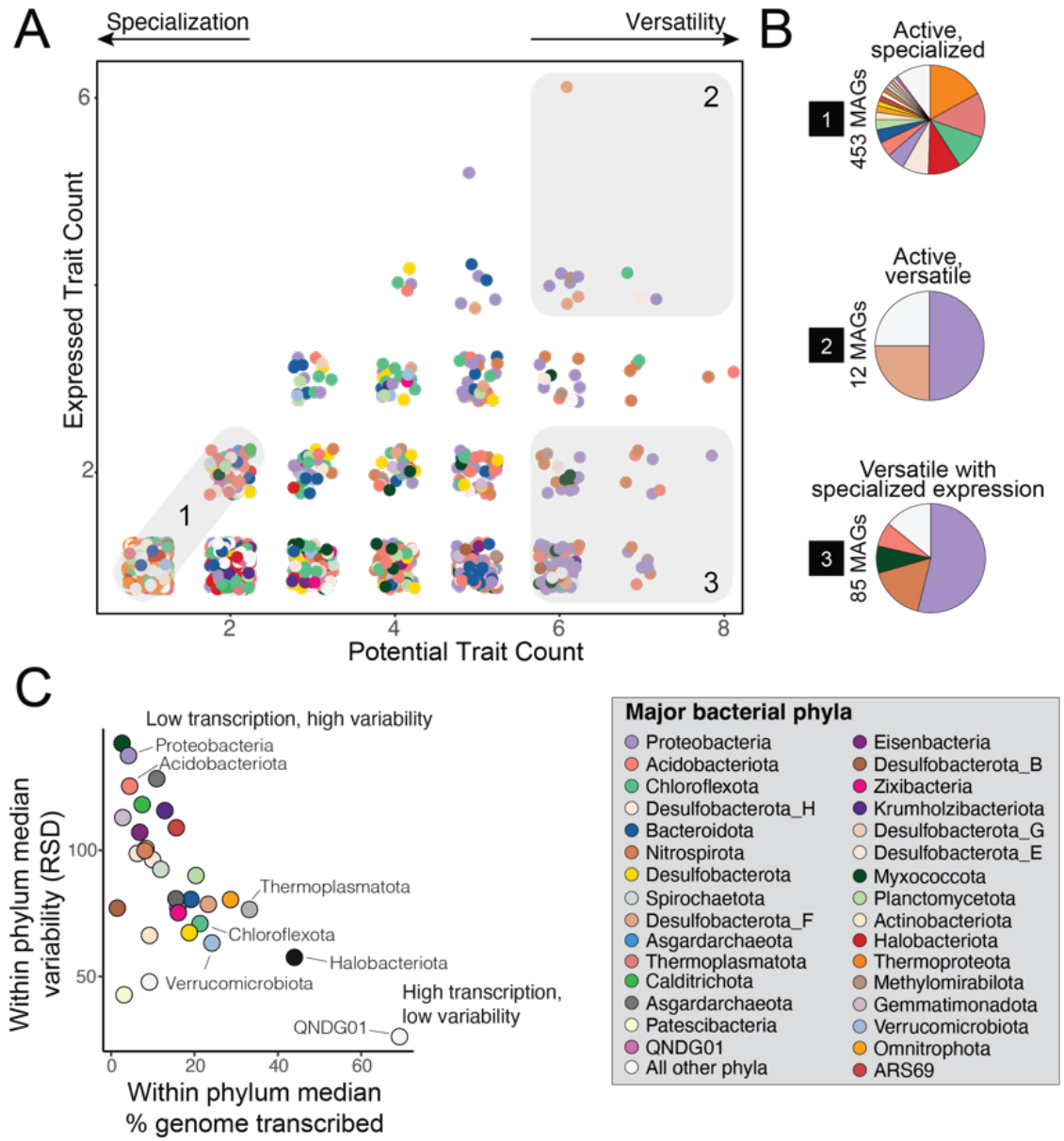


Figure 3. Microbial specialization and versatility. (A) The distribution of potential and expressed traits are shown by a scatterplot where each point is a MAG with transcriptional trait information, colored by phylum. Position on plot defines metabolic specialization (potential < 2 traits and these are expressed) or versatility (potential > 6 traits) as shown by arrows at the top. Gray boxes highlight extreme ends of each spectrum where Group 1) 'active specialized' encoded 2 or less traits and had an expressed to potential ratio of 1. Group 2) 'versatile with multi-functional transcription' encoded 6 or more traits and expressed more 4 or more traits. Group 3) 'versatile with specialized transcription' encoded 6 or more traits and expressed 1 to 2 traits. **(B)** Pie charts summarize the MAGs contained in each 1-3 category in A and are colored by MAG phylum. **(C)** Scatter plot summarizes the median percentage of genome transcribed within each phylum versus the median relative standard deviation (RSD). Points in the upper left-hand corner represent phyla comprised of MAGs with a low median % of overall genes transcribed and high variability across MAGs, and conversely, points in the lower right-hand corner represent phyla are comprised of MAGs with a consistently high % of the overall genome transcribed.

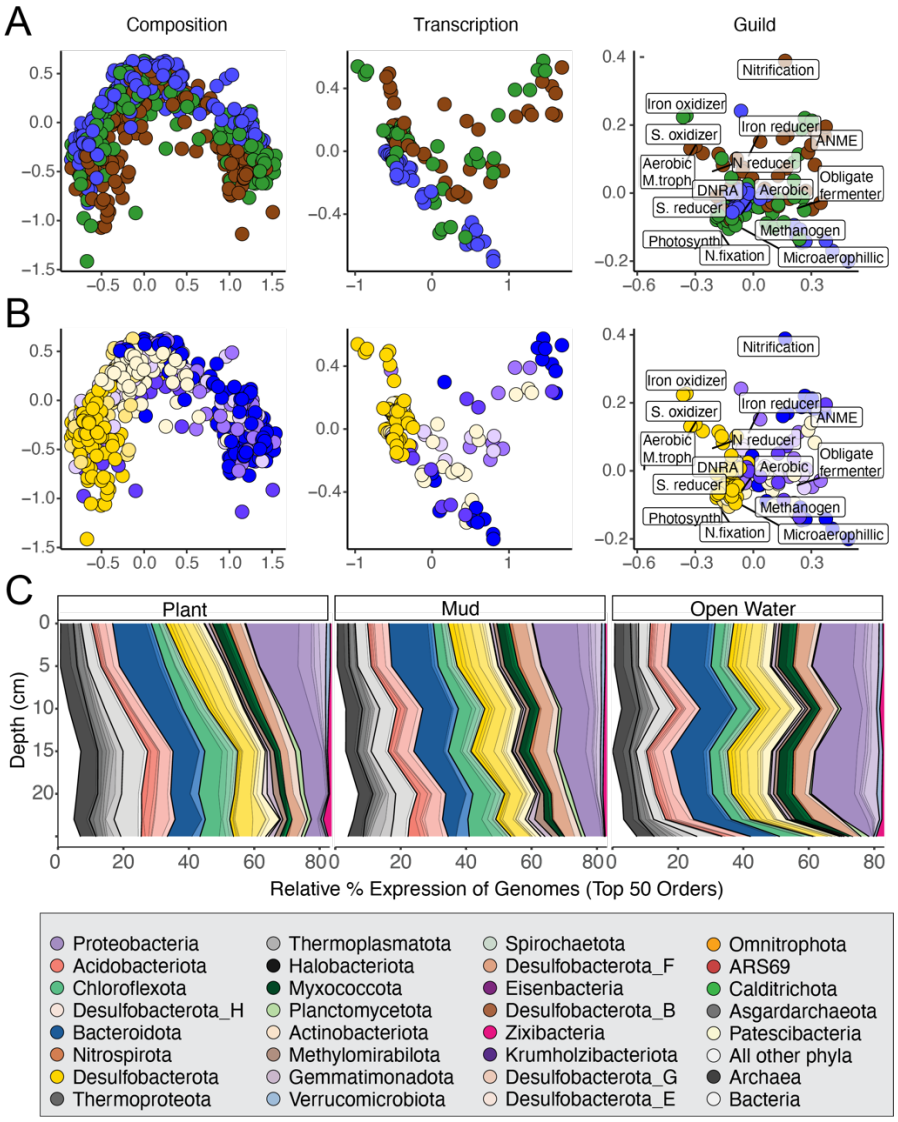


Figure 4. Nonmetric multidimensional scaling (NMDS) ordinations of wetland microbiome structural metrics are organized from (left) overall composition (N=671), (middle) genome transcription (N=133), and (right) guild transcription (N=133). **(A)** The samples are colored by land coverage (brown=mud; green=vegetative; blue=open water), while **(B)** is colored by depth (yellow=surface to blue=deep). A and B highligh the responses of the two major environmental predictors with depth being the strongest predictor across all soil microbiome attributes (**Supplementary Figure S7**). **(C)** Smoothed density plots for the top 50 orders, shaded by phylum level colors provided in the legend. Changes in the relative expression across the sampled depth gradient and across land coverage types (vegetative, mud, and open water).

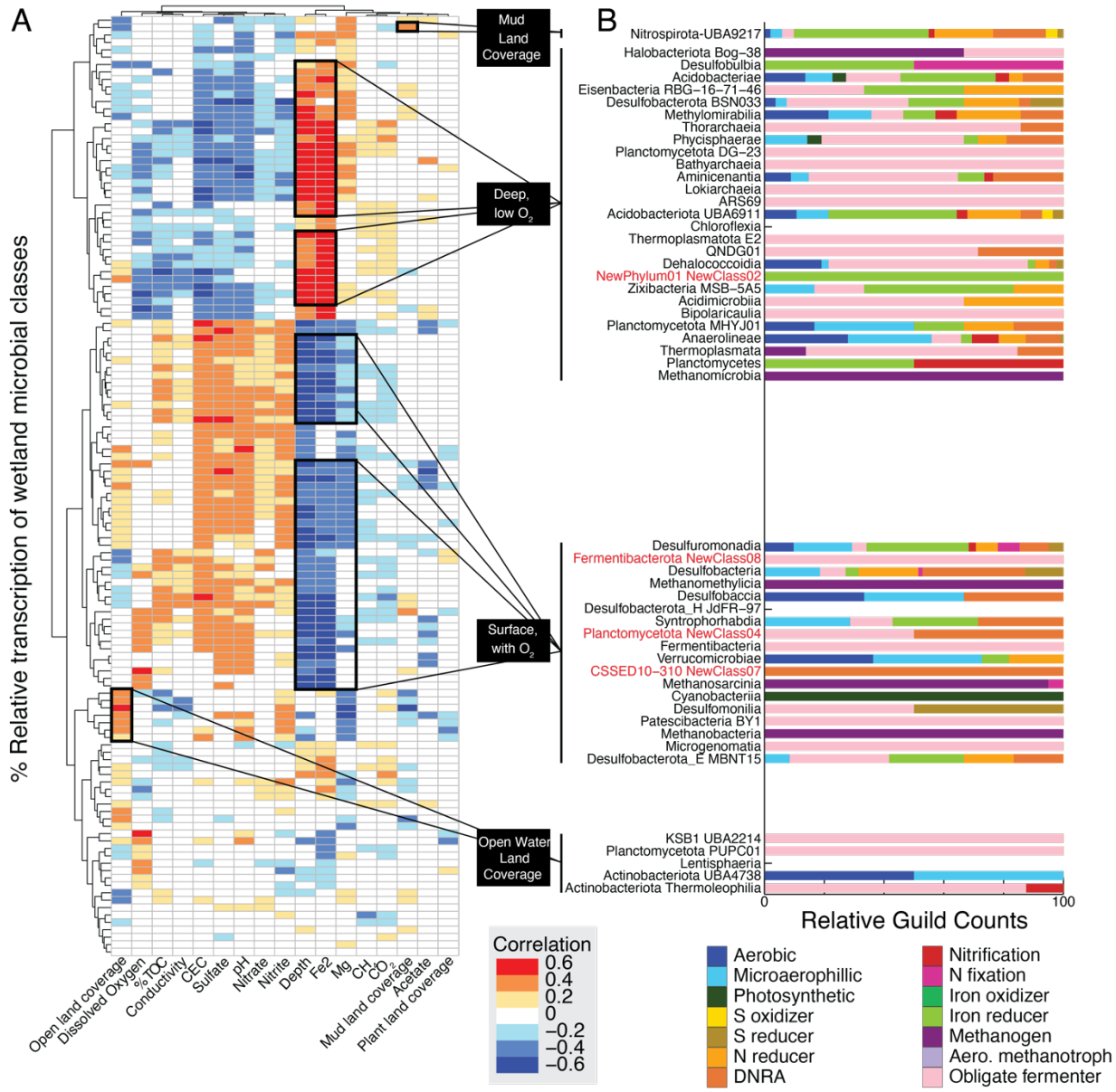


Figure 5. Genome expression of microbial family-level lineages associated with major wetland factors. (A) Heatmap shows correlations between the relative expression of microbial classes and environmental parameters with hierarchal clustering of microbial lineages shown on the left. Text in red are new lineages recovered in this study. Significant relationships between microbial lineages and depth, land coverage, or geochemistry are highlighted by black boxes with white text. **(B)** The stacked bar chart illustrates the distribution of guilds transcribed within each class. For each class, the guilds are tallied across all MAGs in the cohort, with the bar chart reporting the total proportion of all guilds within each class.

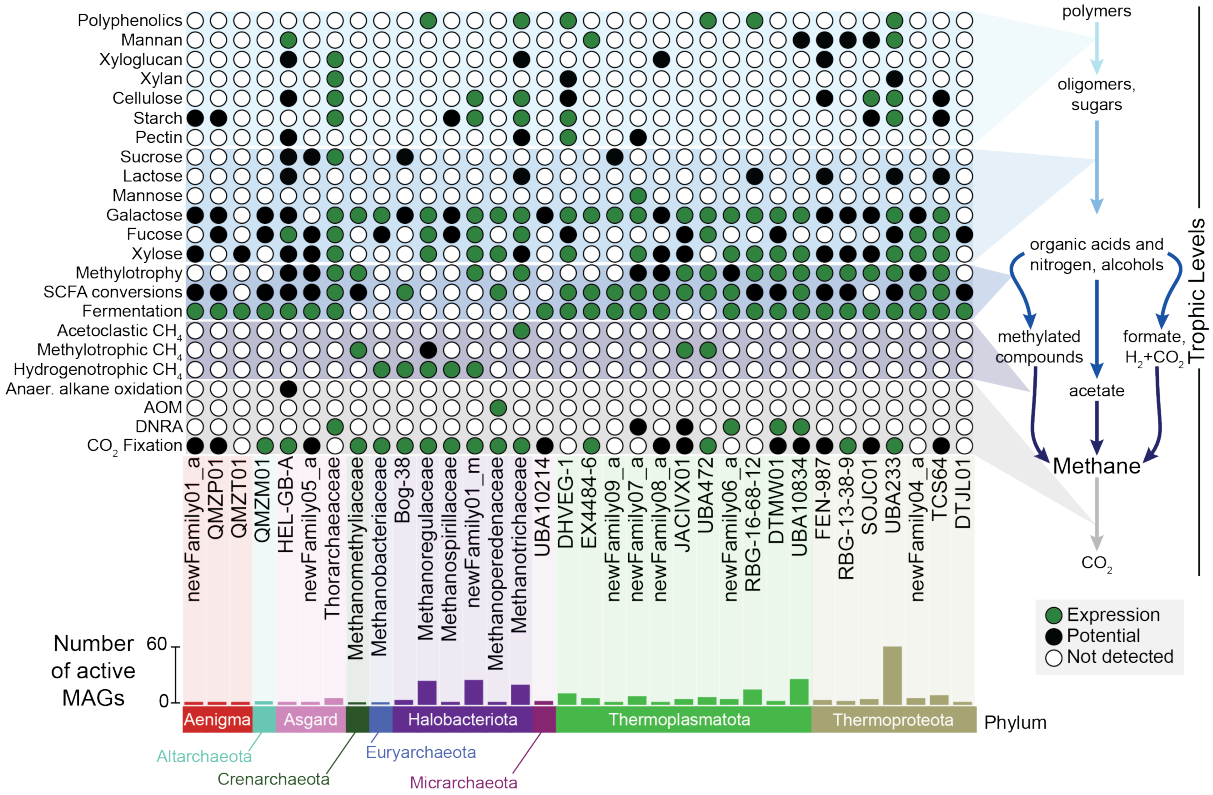


Figure 6. Assigning Archaeal roles in the wetland carbon cycle. (**Right**) The schematic on the right visualizes the major trophic levels contributing to soil carbon decomposition including: light blue polymers (polyphenolics, mannan, xyloglucan, xylan, cellulose, starch, pectin); medium blue oligomers and sugars (sucrose, lactose, mannose, galactose, fucose, xylose); dark blue organic acids, nitrogen, and alcohols (fermentation, methylotrophy, anaerobic alkane oxidation); purple methanogenesis (acetoclastic, methylotrophic, and hydrogenotrophic); and grey other 1 carbon metabolisms (anaerobic methane oxidation, CO_2 fixation). (**Left**) For each archaeal family the bubble plot denotes genes transcribed (green), genes present (black), or genes absent (white) for the major carbon utilization pathways assigned to the trophic levels. (**Bottom**) The bar chart summarizes the number of transcriptionally active MAGs recovered within each archaeal phylum and family.

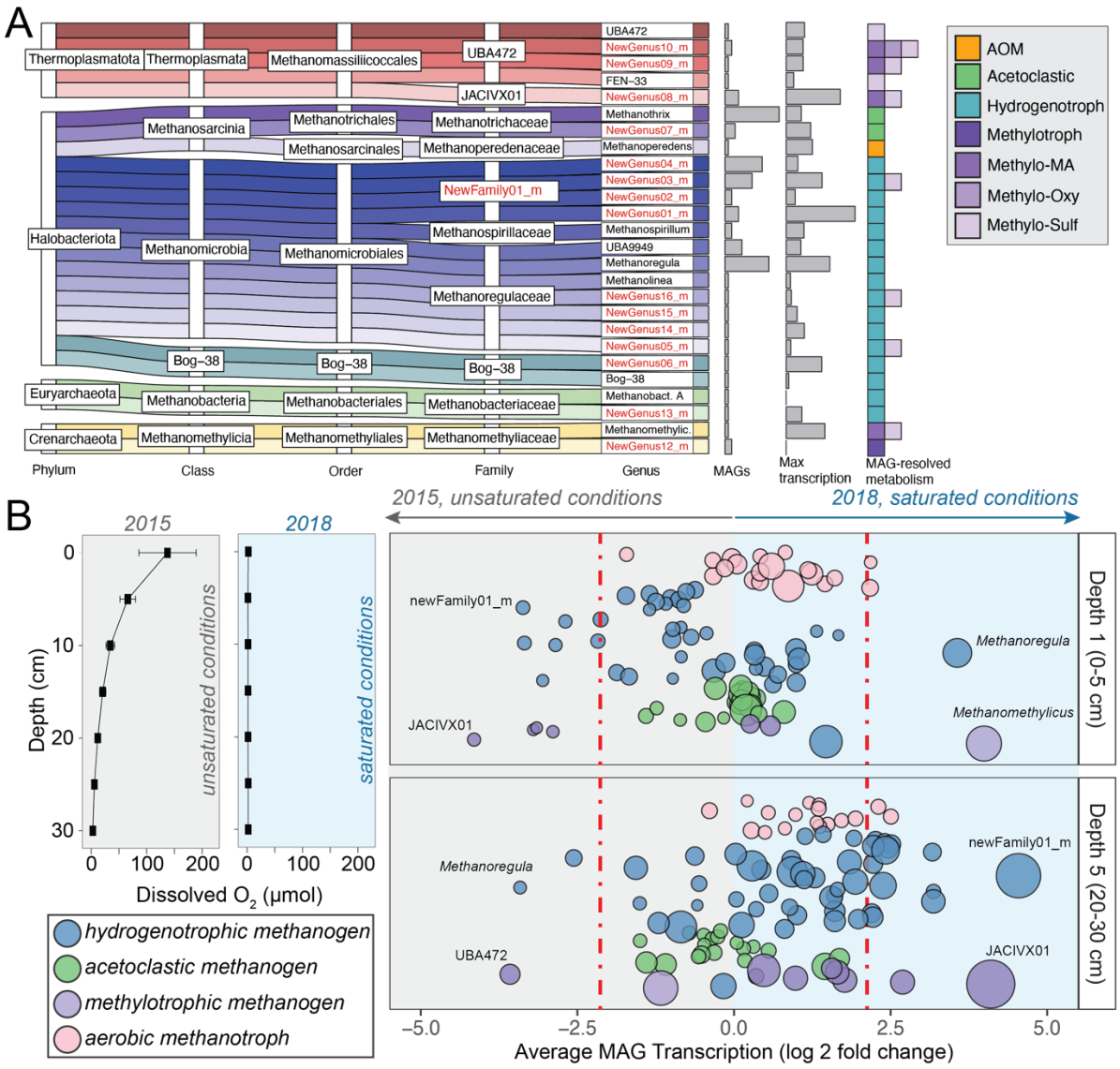


Figure 7. Methanogen diversity and stability in response to flooding. (A) Alluvial plot highlighting the taxonomic diversity of methanogen MAGs recovered here. The first gray bar chart summarizes the number of genomes per genus recovered and the second gray bar chart summarizes the max transcription detected of a MAG within the genus. The MAG-resolved metabolisms show transcription-based substrate utilization. Taxonomy in red are new lineages defined here. **(B)** Dissolved porewater oxygen depth profiles from August 2015 (grey) and 2018 (blue) highlight the redox impact of flooding. Scatterplot shows the change of mean MAG transcription from 2015 to 2018 from the methane cycling community within surface (0-5 centimeters) and deep (20-30 centimeters) soil compartments. Bubbles represent a single methane cycling MAG and are colored by metabolism and sized by mean transcription. MAGs within the red dashed lines do not have log2 fold change in mean transcription and are considered temporally stable. The naming follows the legend in A.

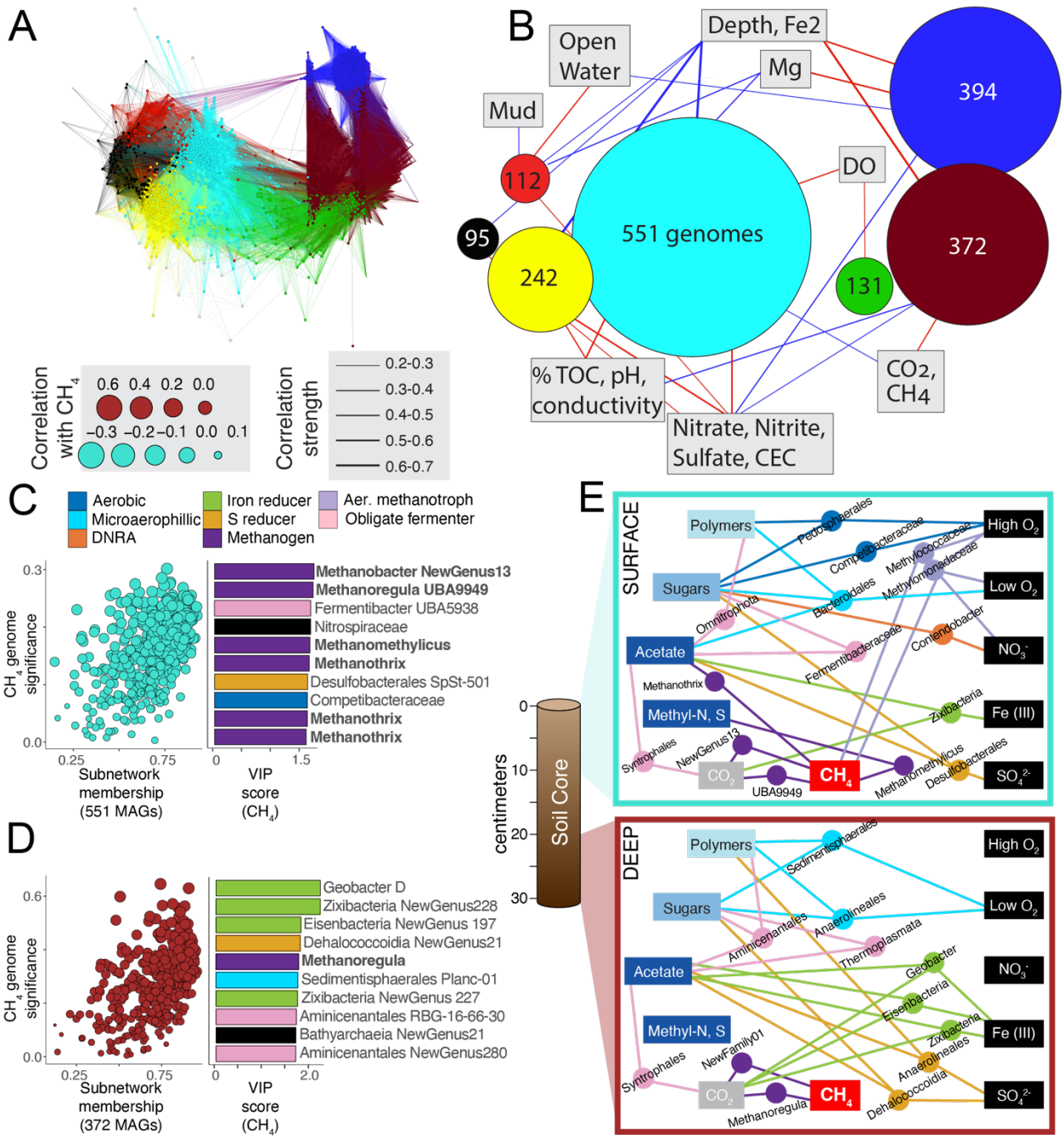


Figure 8. Coordinated gene expression networks predict wetland methane concentrations. **(A)** Overall co-expression network from 1,948 MAGs across 132 metatranscriptomics samples. Node outline colors highlight modules (where N > 100) identified with weighted gene co-expression analyses. **(B)** A simplified network diagram highlighting the associations between modules (sized scaled by number of MAGs) and environmental features. Red and blue edges correspond to positive and negative associations, respectively. **(C)** The genome significance plot and highest variable important genomes related to CH_4 concentrations in the turquoise, surface derived subnetwork (n = 551 MAGs, Rho = 0.53, P < 0.001). **(D)** The genome significance plot and highest variable important genomes related to CH_4 concentrations from

743 the brown, deep subnetwork (n = 372 MAGs, Rho = 0.58, P < 0.001). For panels C and D, the bar charts
744 (middle) highlight the relative contribution of the top 10 variably important genomes (colored by
745 metabolic classification). Methanogens are highlighted with bold text. **(E)** A process-based schematic
746 based on genome-resolved transcript data of the highest variably important taxa and their contribution to
747 methane production in the surface subnetwork from A (top) and the deep subnetwork from B (bottom).
748
749 **Methods**
750 **Field campaign and sample collection.** We extensively sampled the main wetland ecological
751 sites (ecosites), or ecohydrological patch types, across two major field campaigns to capture the
752 patch-level spatiotemporal heterogeneity. The patches included wetland plant cover (plant),
753 predominantly *Typha latifolia*, exposed mud -flat (mud), and standing freshwater (open water).
754 Our sampling was also depth-resolved every 5 cm from 0-25 cm, resulting in a total of 705 soils
755 included in this study (**Supplementary Fig. S1A**).
756
757 Sampling was conducted at the Old Woman Creek National Estuarine Research Reserve (OWC).
758 OWC is a freshwater wetland located on the southern shore of Lake Erie, OH, USA (41˚22'N,
759 82˚30'W). Samples were collected during field campaigns in 2013, 2014, 2015 and 2018.
760 Sample collection details for the 2013-2014 and 2014-2015 campaigns were previously
761 described (1–3). Consistent collection methods were employed for the 2018 campaign, which is
762 first reported in this publication. Briefly, in 2018 we selected 10 sampling sites that include
763 locations sampled during prior years and represent a range of current and historic patch types of
764 Typha (n=1), Nelumbo (n=3), Open Water Channel (n=3), and Temporal Mud flat (n=3) (**Fig.
765 1A**). At each location, cores were collected within a meter of each of the 10 installed pore-water
766 dialysis peepers (n=1 core for May, June, and n=3 cores for July, August, and September, during
767 peak methane production at the wetland). Prior to core collection, we measured soil dissolved
768 oxygen (DO) and temperature within 5 cm increments from the soil surface using the Fibox 4
769 meter and DP-Pst3 probe (PreSens Precision Sensing, Germany). Cores were hydraulically
770 extruded and immediately divided into 5 cm depth sections from 0-25 cm, as measured from the
771 core surface. Each 5 cm section was further subsampled for molecular and geochemical analyses:
772 subsamples were flash frozen in the field and then shipped on dry ice for storage at -80˚C prior
773 to RNA extraction.
774
775 **Geochemical analysis and collection of gasses.** Concentrations of soil nitrate, nitrite, acetate,
776 and sulfate were measured using ion chromatography as previously described (1). Briefly, 1:1
777 soil slurries were made using 5g of soil with 5 ml MilliQ water. Following pH measurement,
778 slurries were filtered using a 0.2µM filter, with the filtrate quantified on a Dionex ICS-2100 Ion
779 Chromatography System with an AS18 column. Using soil slurry filtrate, concentrations were
780 measured for: formate, methanol, acetate, at the Pacific Northwest National Laboratory using
781 proton magnetic resonance (^{1}H NMR) as previously described (4, 5). Fe^{2+} was measured in the
782 field using Hach FerroVer packets per the manufacturer's directions (Hach #927-9). Surface
783 emissions of CH_4 and CO_2 were measured by non-steady-state chambers in duplicate in each
784 ecosite and season and in sync with peeper measurement times (1). We also measured *in situ*
785 concentrations of dissolved CH_4 and CO_2 belowground monthly via porewater dialysis samplers
786 (1). In brief, peepers with a vertical resolution of 2.8 cm measured concentrations via a 0.1 µm
787 membrane, allowing water inside the peeper windows to equilibrate with dissolved gas
788 concentrations, with minimal soil disturbance. Gas concentrations from peepers were quantified

789 with a Shimadzu GC-2014 gas analyzer as previously described (1). Data from these analyses are
790 included in **Supplementary Table S4**.
791
792 **Soil metabolomics methods.** Water-soluble metabolites were extracted from soil by adding 7
793 mL of autoclaved Milli-Q water to 1 g of wet soil in a sterile 15-mL centrifuge tube. To identify
794 field-present metabolites in Old Woman Creek, including methanogenic substrates, we
795 performed ^{1}H NMR on aliquots water extractions. Samples (180 µL) were combined with 2,2-
796 dimethyl-2-silapentane-5-sulfonate-d_6 (DSS-d_6) in D_2O (20 µL, 5 mM) and thoroughly mixed
797 prior to transfer to 3 mm NMR tubes. NMR spectra were acquired on a Bruker Avance III
798 spectrometer operating at a field strength of 17.6 T (^{1}H ν_0 of 750.24 MHz) and equipped with a
799 5mm Bruker TCI/CP HCN (inverse) cryoprobe with Z-gradient and at a regulated temperature of
800 298 K. The 90° ^{1}H pulse was calibrated prior to the measurement of each sample. The one-
801 dimensional ^{1}H spectra were acquired using a nuclear Overhauser effect spectroscopy
802 (noesypr1d) pulse sequence with a spectral width of 12 ppm and 1024 transients. The NOESY
803 mixing time was 100ms and the acquisition time was 4s followed by a relaxation delay of 1.5 s
804 during which presaturation of the water signal was applied. Time domain free induction decays
805 (72114 total points) were zero filled to 131072 total points prior to Fourier transform. Chemical
806 shifts were referenced to the ^{1}H methyl signal in DSS-d_6 at 0 ppm. The 1D ^{1}H spectra were
807 manually processed, assigned metabolite identifications and quantified using Chenomx NMR
808 Suite 8.6. Metabolite identification was based on matching the chemical shift, J-coupling and
809 intensity of experimental signals to compound signals in the Chenomx and custom in-house
810 databases (**Supplementary Table S8).** Quantification was based on fitted metabolite signals
811 relative to the internal standard (DSS-d_6). Signal to noise ratios (S/N) were measured using
812 MestReNova 14 with the limit of quantification equal to a S/N of 10 and the limit of detection
813 equal to a S/N of 3. Data from these analyses are included in **Supplementary Tables S4 and S8**.
814
815 In preparation for LC-MS, water extracts were frozen and lyophilized dry (FreeZone 2.5 Plus,
816 Labconco). Each extract was resuspended in 170 µL of 100% methanol containing isotopically
817 labeled internal standards (5-50 µM of 13C,15N Cell Free Amino Acid Mixture, #767964,
818 Sigma; 1 ug/mL 2-amino-3-bromo-5-methylbenzoic acid, ABMBA, #R435902, Sigma) and
819 centrifuge-filtered (0.22 µm hydrophilic PVDF membrane, #UFC30GV00, Millipore) before
820 transferring to glass LC-MS autosampler vials. To detect polar metabolites using LC-MS/MS,
821 normal phase chromatography was performed using a QExactive HF Orbitrap MS (Thermo
822 Scientific, San Jose, CA) in line with an Agilent 1290 LC stack. Chromatography was
823 performed using a HILIC column (Agilent InfinityLab Poroshell 120 HILIC-Z, 2.1 x 150 mm,
824 2.7 µm, #673775-924) held at 40 ºC and mobile phase running at a flow rate of 0.45 mL/minute
825 with 2 µL injection volumes for each sample. The column was first equilibrated with 100%
826 buffer B (95:5 ACN:H2O with 5 mM ammonium acetate) for 1 minute, then diluting this down
827 to 89% with buffer A (100% H2O with 5 mM ammonium acetate and 5 µM
828 methylenediphosphonic acid) over 10 minutes, then down to 70% B over 4.75 minutes, then
829 down to 20% B over 0.5 minutes, followed by isocratic elution in 80% buffer A for 2.25 minutes
830 followed by re-equilibrating the column at 100% B prior to the next injection. Mass
831 spectrometry data was collected in centroid mode in both positive and negative polarity and
832 included full MS spectra from m/z 70-1050 at 60,000 resolution as well as MS2 fragmentation
833 spectra acquired using stepped then averaged 10, 20 and 40 eV collision energies at 17,500
834 resolution. Source settings included spray voltage of 3kV for both positive and negative

835 ionization, sheath gas flow of 55 (au), auxiliary gas flow of 20 (au), sweep gas flow of 2 (au),
836 and capillary temperature of 400 degrees C. Sample injection order was randomized, with an
837 injection blank of 100% methanol run between each sample and replaced by an internal standard
838 or quality control mixture only every 3 to 15 samples. Metabolomics data is available through
839 MassIVE under accession MSV000093935.
840
841 Metabolites were identified by comparing exact mass, retention time (RT) and MS/MS
842 fragmentation spectra to that of compound standards run using the same chromatography and
843 mass spectrometry methods. LC-MS spectra were analyzed using custom Python code (Yao et
844 al., 2015), with a score of 0 to 3 assigned to each feature (unique m/z and RT pair) that
845 represented the level of confidence in identifying the metabolite. A metabolite positively
846 identified had a detected m/z of ≤ 5 ppm or 0.001 Da from theoretical and RT ≤ 0.5 minutes
847 compared to that of a pure standard run using the same LC-MS conditions. Highest level of
848 positive identification (score of 3) was given for features having matching fragmentation spectra
849 to an internal database of standards run and collected on a Q Exactive Orbitrap MS, with
850 mismatching spectra invalidating a positive identification. Methodological processing these
851 analyses are included in **Supplementary Tables S8-S9**.
852
853 **DNA and RNA extraction.** DNA was extracted from soil samples using the PowerSoil DNA kit
854 (Qiagen). Total RNA was extracted using the Takara NucleoBond RNA Soil kit (Takara Bio
855 USA, CA) with 5g of soil as input. Extracted RNA was purified and concentrated using the
856 Zymo RNA Clean & Concentrator Kit (Zymo Research) and quantified with the Qubit RNA HS
857 Assay (Thermo Fisher Scientific). For low biomass samples, up to 5 RNA extractions of 5g were
858 pooled prior to concentration. RNA purity and DNA removal were verified by PCR using
859 universal V4 region 16S rRNA gene primers.
860
861 **16S rRNA gene amplicon sequencing.** 16S rRNA genes in extracted DNA were amplified with
862 the primers 515F/806R (6) and sequenced at Argonne National Laboratory on the Illumina
863 MiSeq with 251-bp paired end read. Read sets (n=701) were processed with QIIME2 (release
864 2018.11, (7)) and ASVs were recovered with DADA2 (8). Taxonomies were assigned with the
865 q2-feature-classifier with the SILVA database (release 132) (9). ASVs were only included in
866 downstream analyses if they were detected in at least 10% of samples. We set a minimum read
867 threshold of at least 5,000 sequences per sample, retaining 672 samples, and then normalized our
868 ASV table with cumulative sum scaling prior to downstream analyses (**Supplementary Table**
869 **S2**).
870
871 We validated the ubiquity of our identified wetlands microbial taxa at OWC with an independent
872 16S rRNA gene database assembled from 505 soils collected from 5 unique wetlands systems
873 that collectively represent a broad range of geographic locations (**Fig. S3**), span a range of
874 geochemical and climatic conditions, and represent different wetland types: marsh (P7, P8,
875 LA2), swamp (JLA), and bog (SPRUCE). All selected samples were previously amplified with
876 the EMP primers 515F/806R and sequenced on the Illumina Miseq. Using DADA2 (8) through
877 QIIME2 (7) 16S rRNA gene sequences were processed with a consistent methodology and used
878 the SILVA v138 and GDTB (10) (r202) databases to assign taxonomy across all 505 samples.
879 We then compared the representative sequences of our wetlands 16S rRNA to those recovered in
880 our assembled reference database using BLAST to compare sequence similarity at 100, 99, 97,

881 and 80% similarity cutoffs. The percentages of ASVs from each site and within the entire
882 wetland database that were detected in our OWC dataset were calculated for each cutoff. The
883 relative abundances of the ASVs within these cutoff groups was calculated for each site and the
884 entire database to understand how much community composition from other wetland sites was
885 shared with OWC.
886
887 **Metagenomic and metatranscriptomic sequencing.** Of the 42 Metagenomes, 25 are newly
888 reported data here (**Supplementary Table S1**). This includes 16 samples from the 2018 field
889 sampling representing a range of patch types, soil depths and time points to be used in building a
890 MAG database. For the 2018 field samples sequencing was performed by the Department of
891 Energy Joint Genome Institute following their established protocols. Briefly, plate-based DNA
892 library preparation for Illumina sequencing was performed on the PerkinElmer Sciclone NGS
893 robotic liquid handling system using Kapa Biosystems library preparation kit. 200.20 ng of
894 sample DNA was sheared to 472 bp using a Covaris LE220 focused-ultrasonicator. The sheared
895 DNA fragments were size selected by double-SPRI and then the selected fragments were end-
896 repaired, A-tailed, and ligated with Illumina compatible sequencing adaptors from IDT
897 containing a unique molecular index barcode for each sample library. The prepared libraries
898 were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run
899 on a Roche LightCycler 480 real-time PCR instrument. Sequencing of the flowcell was
900 performed on the Illumina NovaSeq sequencer using NovaSeq XP V1 reagent kits, S4 flowcell,
901 following a 2x151 indexed run recipe.
902
903 We also sequenced 9 samples from laboratory enrichments that are released in this study. Two of
904 these enrichment samples were sequenced at The Ohio State University Genomic Shared
905 Resource Facility with library preparation Nextera XT library system and sequencing done on
906 HiSeq 2500 using methods described in more detail (11). An additional 7 laboratory enrichments
907 were sequenced in 2018 from Genomics and Microarray Shared Resource at University of
908 Colorado Denver Cancer Center, Denver, CO, USA. For these samples, library preparation was
909 done (via Tecan/Nugen Ovation Ultralow DNA Input Kit, No. 0344) and sequenced on the
910 Illumina NovaSeq 6000, with an S4 flow cell. These enrichments were used to try to enhance
911 methanogen recovery.
912
913 We obtained 133 samples from our metatranscriptomics sequencing across the multi-year study,
914 with their distribution across land coverages, sites, cores, years, months, and depths reported
915 (**Supplementary Table S1, S4**). From the field in 2018 we obtained 109 metatransciptome
916 samples from triplicate cores for each of 3 depths (0-5 cm, 10-15 cm, and 20-25 cm) for July and
917 September, and 5 depths (0-5 cm, 5-10 cm, 10-15 cm, 15-20, and 20-25 cm) for August). The
918 remaining samples were collected in prior years and are described in main text in supplemental
919 tables.
920
921 Library preparation and sequencing were performed at the Department of Energy Joint Genome
922 Institute using standard methods. Plate-based RNA sample prep was performed on the
923 PerkinElmer Sciclone NGS robotic liquid handling system using Illumina Ribo-Zero rRNA
924 Removal Kit (Bacteria) and the TruSeq Stranded Total RNA HT sample prep kit following the
925 protocol outlined by Illumina with the following conditions: total RNA starting material of 100
926 ng per sample and 10 cycles of PCR for library amplification. The prepared libraries were

927 quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a
928 Roche LightCycler 480 real-time PCR instrument. Sequencing of the flow cell was performed on
929 the Illumina NovaSeq sequencer using NovaSeq XP v1 reagent kits, S4 flow cell, and following
930 a 2x150 indexed run recipe. For the remaining 24 samples, these were previously reported in (1,
931 3), with 12 collected in 2014 and 12 from 2015.
932
933 Two samples did not yield sufficient biomass to be sequenced at JGI. For these two samples and
934 nine additional samples from 2015, RNA-seq library construction and sequencing was performed
935 at The Genomics and Microarray Shared Resource at University of Colorado Denver Cancer
936 Center, Denver, CO, USA. Library preparation and rRNA depletion were performed using the
937 Zymo-Seq Ribo Free Total RNA Library Kit Cat No. R3000 and libraries were sequenced on the
938 Illumina NovaSeq 6000 using 2x150 paired end reads on an S4 flow cell.
939
940 **Genome assembly and binning.**
941 Metagenomic samples were assembled and binned using multiple approaches to maximize
942 assembly and metagenome assembled genome (MAG) recovery **(Supplementary Table S3)**.
943 Collectively, this resulted in 42 samples for the genome-resolved database made up of 2,502
944 unique MAGs, with assembly and binning methods described below. Additionally, the 2,502
945 dereplicated MAG database is available on Zenodo (see files MAGs.zip). Reported more
946 extensively in **Supplementary Table 1**, but briefly here MAGs were recovered from
947 1) 2013 soil field samples (n=9), sequencing and selected MAGs published previously (2, 12)
948 2) 2014 soil field samples (n=3), sequencing and selected MAGs published previously (1, 3)
949 3) 2015 soil field samples (n=3), sequencing and selected MAGs published previously (1, 3)
950 4) 2018 soil field samples, *released in this study* (n=16)
951 5) lab enrichment samples, sequencing and selected MAGs published previously (n=2) (4)
952 6) lab enrichment samples, *released in this study* (n=9)
953
954 Raw metagenomic reads from the 2018 metagenomes were trimmed using Sickle (pe) (13) and
955 then full assemblies were generated using Megahit (v1.2.2) (14) with parameters --k-min 41 --k-
956 max 121 --k-step 10. Subsampled assemblies using 30% of sequencing reads were generated
957 using IDBA-UD v 1.1.3 (15) with default parameters. To generate more finely resolved contig
958 depth signatures as an aid to binning, reads from multiple samples were mapped to each
959 assembly using BBMap with minid=0.95 and ambig=random (v 38.70) (16) and these profiles
960 were used with assembled contigs for binning using Metabat2 (17). Bins from these assemblies
961 were combined with bins from metagenomic assemblies derived from earlier sampling of this
962 wetland and from metagenomic sequencing of wetland soil enrichments (1, 3, 4).
963
964 This bin pool was dereplicated using dRep (v 1.4.3) (18) at 99% identity with the 'winning' bins
965 retained as a preliminary wetland genome database. Finally, reads from the 2018 assemblies
966 (n=16) that did not map to contigs in the preliminary MAG database were grouped by depth (e.g.
967 D1-D3 and D4-D5) and co-assembled using Megahit (14). These two co-assemblies were binned
968 as described above and dereplicated with the preliminary database yielding a final database of
969 3217 wetland MAGs. MAG completeness and contamination was estimated using CheckM (19)
970 and taxonomy assigned using GTDB-tk (v 1.5.0) with GDTB database release 202 (10, 20). Of
971 these 3217 MAGs, 2502 were of medium or high quality based on adapted MIMARKS standards
972 (completeness >=50% and contamination <10%) (21).

973
974 **MAG based phylogenomic assignments.** All MAGs were used as input for GTDB-tk (20)
975 taxonomy assignment using the *de novo* workflow and Nanoarchaeota as the outgroup using
976 release 202 of GTDB. The resulting multiple sequence alignment was used to produce a guide
977 tree using IQ-TREE (22) and the LG+F+G model. A PMSF tree was generated using IQ-TREE
978 with model LG+C10+G+F and 100 bootstraps. The resulting tree was annotated with iTOL (23).
979 Phylogenomic analysis of the 99% dereplicated methanogen MAGs from OWC (n=85) and
980 Stordalen Mire (n=74) was performed using the GTDB-Tk v2.3.0 r214 (20). The alignment was
981 based on 53 concatenated archaeal marker genes. A GTDB-derived MAG (GCA_002495465.1)
982 from the phylum Undinarchaeota was used as in outgroup. The tree was visually modified in R
983 using the ggtree package (24). Metatranscriptome data from Stordalen Mire included on the tree
984 was previously published (5). The Newick tree is available at on Zenodo at DOI
985 10.5281/zenodo.8194033.
986
987 **Assessment of genome transcription.** Raw metatranscriptomic reads were filtered and initial
988 QC performed using Joint Genome Institute standard protocols. Filtered reads were further
989 quality trimmed using sickle (pe) (13) and mapped to the full MAG database using BBMap (v
990 38.70) (16). Following mapping, bamfiles were filtered to retain mappings of at least 99% ID
991 and per-gene read counts were estimated using HT-Seq (v0.11.2) (25). The read count table was
992 first filtered to retain genes which have a minimum total count of 10 across all samples. The
993 filtered table was normalized for gene length and TMM normalized using the EdgeR 'cpm'
994 function with log2 normalization (26). MAG mean transcription values were calculated for those
995 MAGs having at least 20 genes transcribed. Due to file size constraints, the per MAG transcript
996 profiles across 133 metatranscriptome samples are accessible in the zenodo library that
997 accompanies this manuscript (10.5281/zenodo.8194033). This includes the MAG based analysis
998 **(owc_metat_table_mags.csv)** as well as annotated expressed DRAM genes
999 **(owc_metat_table_mags_genes.csv).**
1000
1001 To calculate the percentage of a genome transcribed for the 1,948 genomes recovered where at
1002 least 20 genes were transcribed, we first counted the number of genes transcribed per genome,
1003 and then divided that count by the number of predicted genes per genome as estimated by
1004 CoverM (27) and multiplied by 100. We performed a Mann-Whitney U test to assess if there was
1005 a significant difference in the percentage of a genome expressed in bacterial versus archaeal
1006 genomes. We also classified phyla by their median transcription and variability by calculating
1007 the median percentage of genomes transcribed (summarized by phylum) and by the variability of
1008 that % transcription of MAGs within a phylum, calculated as the median relative standard
1009 deviation (RSD, aka coefficient of variation). These data are accessible via the zenodo
1010
1011 **MAG and Gene Based Functional profiling.** Bacterial and archaeal MAGs were annotated
1012 using DRAM v1(28). Because of file size constraints, these data files are included on the zenodo
1013 (see data accessibility). This includes (1) the MAG amino acid gene sequences derived from
1014 DRAM gene calls (fasta file) as **OWC_HQMQ_DB_genes.faa.gz** and MAG DRAM
1015 annotations as **OWC_HQMQ_DB_ANNOTATIONS_20220208.txt.gz.** We also annotated
1016 bacterial and archaeal metagenome-resolved transcriptomes *via* DRAM (28), with these also on
1017 zenodo also, as (1) metatranscriptomic expression per gene across 133 metatranscriptomes (csv
1018 table)

1019 in file **owc_metat_table_mags_genes.csv** and (2) the corresponding DRAM annotations for
1020 these transcribed genes in file **owc_metat_table_mags_genes_annotations.csv**.
1021
1022 To assign traits to MAGs, we curated the MAG annotations made by DRAM (28) using a ruleset
1023 visualized in **Supplementary Fig. S5**. MAGs were assigned potential traits derived from
1024 metagenomics and expressed traits derived from metatranscriptomics. From metagenomics,
1025 MAGs were first classified by respiratory capacity based on the presence of >50% of the
1026 subunits required for Complex 1 of the electron transport chain and the presence of at least one
1027 gene for an electron acceptor. Based on which acceptor was predicted, a MAG would be
1028 assigned to a particular functional guild (e.g., N reducer, aerobic, iron reducer). As an example,
1029 for a MAG to be classified as microaerophillic, we required the MAG to have more than 50% of
1030 Complex 1 subunits and at least one subunit of a high affinity cytochrome oxidase. Likewise, if a
1031 MAG did not have more than 50% of the subunits required for Complex 1 of the electron
1032 transport chain or the potential for any electron acceptor, it was classified as an obligate
1033 fermenter.
1034
1035 Beyond respiratory metabolisms, other traits were assigned based on functional gene presence.
1036 For example, N fixation was assigned based on the presence of at least one subunit of any
1037 nitrogenase (**Table S5**). For specific metabolisms, such as methanogenesis and methanotrophy,
1038 assignment was based on the presence of a key functional gene (*mcrA* and *pmoA*, respectively) or
1039 by manual curation including taxonomic information given the conserved nature of these
1040 metabolisms at the genus level and the tendency for these key functional genes to be on short
1041 contigs, precluding them from binning (1, 3). From metatranscriptomics, trait calls were based
1042 on potential calls from MAGs, but also required various levels of transcription to be considered
1043 expressed. For example, to be considered an expressed microaerophile, a MAG had to 1) be
1044 classified as a microaerophile based on metabolic potential (more than 50% of complex 1 subunit
1045 and at least one subunit of a low affinity cytochrome oxidase) and 2) had to be expressing more
1046 than 30% of complex 1 subunits and at least one subunit of a low affinity cytochrome oxidase.
1047 For trait classified based on an absence of data (obligate fermenter) or taxonomy (methanogen,
1048 methanotroph), MAGs needed to be considered active (more than 20 genes transcribed) to be
1049 classified as expressed for a particular trait. As trait assignment was based on MAG annotation,
1050 we also assigned MAG-level transcription. Likewise, MAGs could be assigned to more than one
1051 functional guild, and totals were summed across samples. The metabolic guild assignments for
1052 the 2,502 MAGs in both potential and expression is contained in **Supplementary Table S5**.
1053
1054 We note that functional gene based phylogenetic analyses were performed to refine the
1055 annotation of nitrogen related metabolism including genes annotated as respiratory nitrate
1056 reductase (*nar*), nitrite oxidoreductase (*nxr*), ammonia monooxygenase (*amo*), or methane
1057 monooxygenase (*pmo*) to improve the assignment the nitrogen cycling capabilities of MUCC
1058 MAGs (**Supplementary Table S5**). Specifically, Nxr/Nar and PmoA/AmoA amino acid
1059 reference sequences were downloaded and this set of reference sequences were combined with
1060 amino acid sequences of homologs from the MUCC, aligned separately using MUSCLE
1061 (v3.8.31) (29), and run through an in-house script for generating phylogenetic trees. This resulted
1062 in two phylogenies, one for *Nxr/Nar* genes and one for *Pmo/Amo* genes, with this approach used
1063 to refine the homology-based gene annotations in the MAG database.
1064

1065 For MAGs assigned to the methanogen functional guild, substrate classification was further
1066 characterized, assigning each as hydrogenotrophic methanogen, acetoclastic methanogen, or
1067 methylotrophic (methyl-N, methyl-S, methyl-O) methanogen, with genes required for each
1068 detailed in Table S4. Gene calls were based on DRAM annotations. Specifically, all genes
1069 involved in methanogenesis were inventoried per MAG and then based on those genes that were
1070 specific for a substrate type, substrate calls were made. Archaeal metabolism calls were based on
1071 DRAM annotations with specific genes required for each detailed in **Supplementary Table S7**
1072 (Archaea specific).
1073
1074 **Identification of major environmental axes driving wetlands microbiome variation.** We first
1075 reduced collinearity by dropping a subset of geochemical and climatic variables that were highly
1076 correlated with retained variables (Spearman's r > 0.75; **Fig. S2**) including Ethanol, Isopropanol,
1077 extractable Na, extractable Al, % Sand, extractable K, and temperature. We log-transformed
1078 variables where appropriate and standardized variables prior to analyses. Our final set of
1079 environmental predictors included measured soil properties (Dissolved Oxygen, Acetate, Nitrate,
1080 Nitrite, Sulfate, Fe2, pH, CEC, % Total Organic Carbon, % Silt, % Clay, Hydraulic
1081 Conductivity, extractable Mg), site attributes (ecotype, depth, collection month, collection year),
1082 associated geographic variables (latitude/longitude), and greenhouse gas concentrations (CH_4 and
1083 CO_2 peeper concentrations).
1084
1085 To characterize how the composition and transcription of microbial communities differed across
1086 time, space, and major geochemical gradients, we first assessed calculated dissimilarities with
1087 Bray-Curtis metric on square root transformed relative abundances. We used three metrics of
1088 microbiome communities: taxonomic composition (ASVs), relative mean MAG transcription,
1089 and relative mean metabolic guild transcription, and visualized as ordinations with nonmetric
1090 multidimensional scaling (k=3). We tested the correlation between pairwise distances in bacterial
1091 and archaeal composition/transcription to pairwise distances in our suite of geochemical and
1092 spatiotemporal variables using Mantel tests with 999 permutations. Then, to identify the best
1093 overall explanatory models associated with differences in composition and expression we
1094 conducted multiple regression analyses on distance matrices (**Fig. S7**).
1095
1096 **Genome co-expression network analysis and process-based model generation.** Beyond
1097 individual lineages correlations described above, we were interested in identifying if there were
1098 community-level emergent properties (e.g., clusters of co-transcribed genomes) that were highly
1099 correlated in their relative expression to relate to in situ measurements of greenhouse gas
1100 concentrations (CH_4, CO_2). To test this, we used weighted correlation network analysis
1101 (WGCNA) (30) across all of our metatranscriptomes (n=132, 1 sample was excluded due to
1102 outlier screening within the WGCNA pipeline). A signed adjacency measure for each pair of
1103 features was calculated with the power threshold of 7, identified by plotting mean connectivity as
1104 a function of the soft-thresholding power and selecting the power threshold above 0.80 to
1105 optimize the scale-free topology network fit. The adjacency matrix was used to calculate
1106 topological overall measure (TOM) and derive a co-expression network. We used a signed
1107 hybrid network, with a minimum module size of 50, and a merge height of 0.3. Hierarchical
1108 clustering was employed to resolve subnetworks (i.e., modules), using a distance based on the
1109 calculated TOM. For each subnetwork, we calculate pairwise Pearson correlation coefficients
1110 between the subnetwork principal components and key geochemical measurements (pH, sulfate,

CEC, nitrate, nitrite, DO, % TOC, conductivity, depth, Fe_2, acetate, Mg, and ecotypes) as well as our *in situ* measurements of CO_2 and CH_4 concentrations.

We next assessed the two subnetworks that had the highest significant correlation scores (correlations between module eigenvalues and CO_2/CH_4 concentrations, which we refer to as 'traits') with an intramodular analysis. Our first goal was to identify genomes with a high genome trait significance (the absolute value of the correlation between the genome and the trait) and subnetwork membership (the correlation of the module eigenvalue and the genome expression profile). Second, we also computed VIP scores from sPLS (31) models (in relation to both CH_4 and CO_2) for all genomes within each of the two subnetworks to obtain an additional metric of their individual importance. VIP scores were highly correlated with genome trait significance scores (rho > 0.9 for all). Subnetwork #1 contained 372 genomes (the subnetwork–trait correlation to CH_4 and CO_2 was 0.38 and 0.36 respectively, with $P<0.0001$ for both). Likewise, for each subnetwork, we found that genome significance to be significantly correlated to subnetwork membership both for CH_4 (Rho = 0.54 and 0.52 for subnetworks #1 and 2, respectively; $P<0.001$) and for CO_2 (Rho = 0.54 and 0.29 for subnetworks #1 and 2, respectively; $P<0.001$).

MAGs within the process-based model were based on WGCNA (30) and sPLS (31) analyses, as well as MAG expression profiles. First, MAGs within each of the two subnetworks (brown and turquoise) were filtered to those with a VIP score ≥ 1 in relation to CH_4 and then clustered at the genus level. Then clusters were classified based on expressed traits. For clusters containing more than one MAG, functional guilds were consolidated. If trait assignments disagreed (e.g. aerobe and obligate fermenter) within a cluster, the functional guild from the most complete MAG within a cluster was used. Clusters with no expressed adjectives or without at least one MAG expressing $\geq 20\%$ of its genes were removed from subsequent analyses. Remaining clusters were retained in the process-based model. For MAGs within the remaining clusters, DRAM annotations were used to profile expressed carbon usage. Based on the expressed carbon utilization and functional guild of each cluster, a schematic representation of the process model (**Fig. 8E**) was constructed in Adobe Illustrator. All the statistics and metabolic calls used to construct Fig. 8 are provided in **Supplementary Tables S11 and S12**, including VIP values, expression patterns, functional guilds, and carbon utilization provided in. This analysis workflow is shown in **Fig. S10**.

Figure S1. Experimental Design. (A) Upset plot showing the major data types obtained from samples and the number of samples obtained for each set (top bars) including from left to right: geochemistry, 16S rRNA amplicon sequencing, NMR metabolomics, metatranscriptomics, and metagenomics. The gray bars (right) represent the number of interesting sets across sample types. **(B)** Overview of the number of metagenome-resolved genomes (MAGs) recovered in this study. All 2,502 medium and high-quality (MQHQ) MAGs used in analyses were ≥50% complete, were <10% contaminated. **(C)** Schematic illustration depicting the soil core sampling as well as the greenhouse gas monitoring using a) porewater dialysis peepers for porewater production, b) chambers for emission quantification, and c) EC tower for site wide data. **(D)** Boxplot highlights the amount of sequencing in Giga basepairs (Gbp) for this study in Old Woman Creek (OWC) versus other classified wetlands in the Sequence Read Archive (SRA) for metagenomes and metatranscriptomes. Each point represents a single sample, and boxplots show the median Gbp. **(E)** Overview of the genome-resolved metagenomics (top) and genome-resolved metatranscriptomics (bottom) bioinformatics pipeline we developed to obtain genome-resolved microbial community activity profiles. The experimental data described in this figure can be accessed in **Supplementary Table S1, S3-S4**.

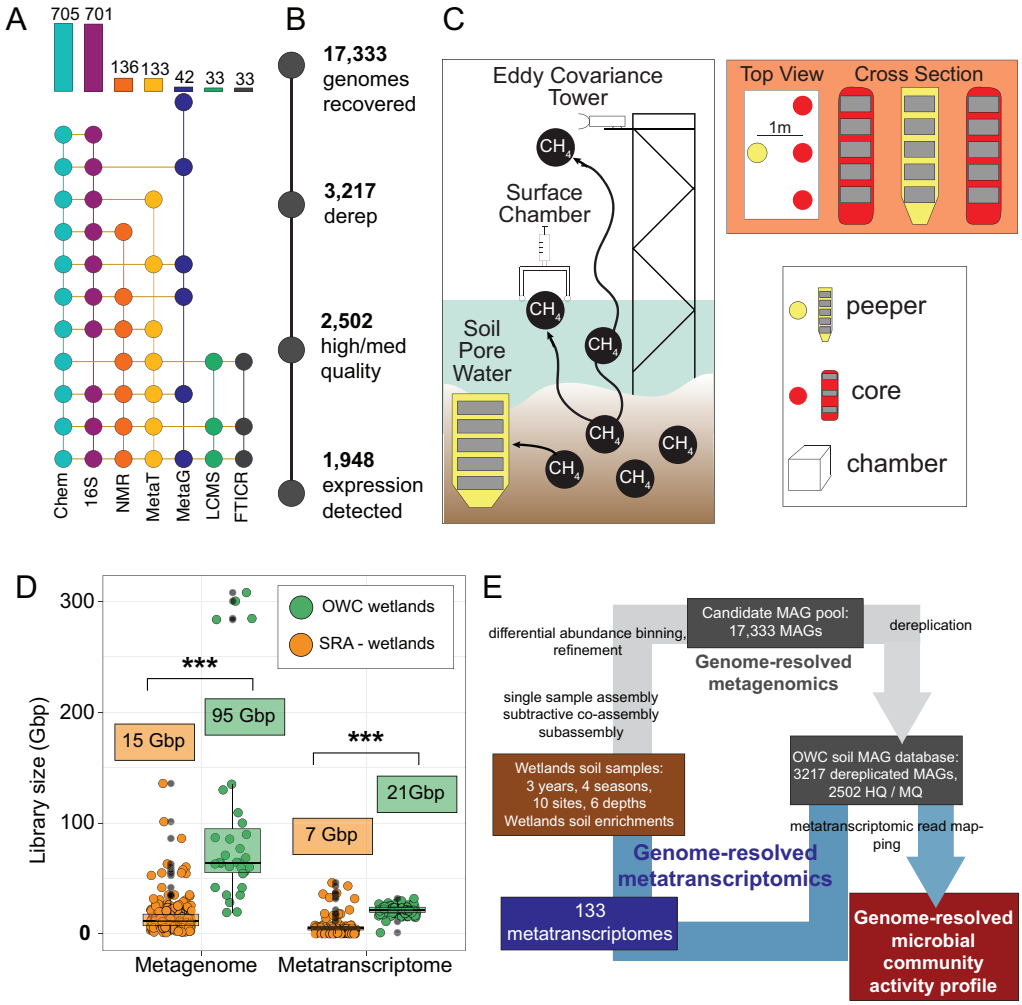


Figure S2. Correlation analysis highlights the soil, porewater, and flux metadata associated with metatranscriptomics and metagenomics samples, and the correlation of these variables to one another (denoted in heatmap legend). The data supporting this figure is included in **Supplementary Table S4** and includes the year, month, land coverage, depth, core, site transect.

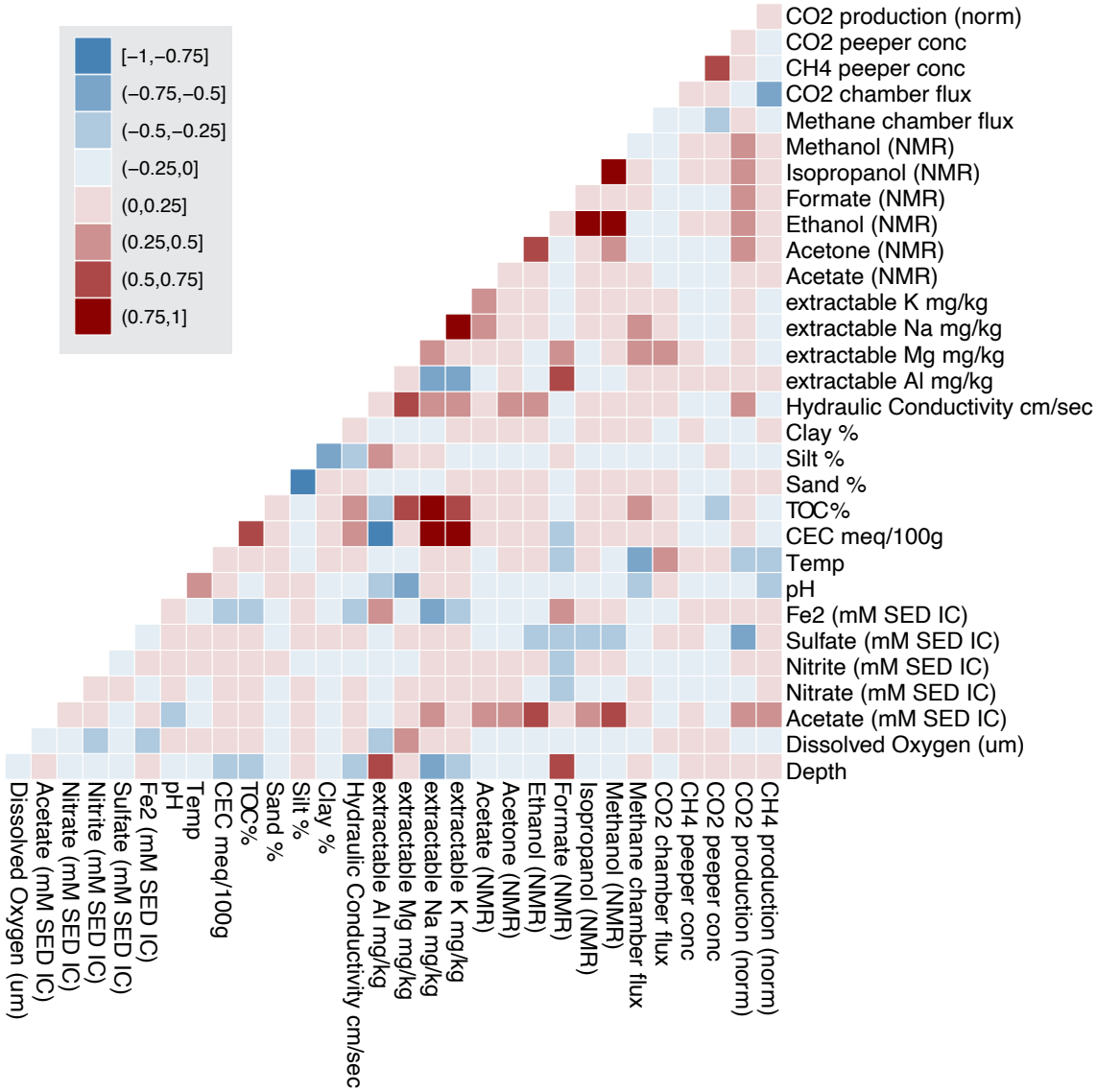


Figure S3. MAGs and ASVs recovered from this study (OWC) represent a large fraction of the microbiome of diverse North American wetland sites. (A) The geographic distribution of wetland sites for which we compiled a North American wetlands database from available 16S rRNA **(B)** A bar plot of the number of samples at each wetland site (ASV samples). The % composition **(C)** and the % of ASV matches **(D)** of the wetlands database covered by OWC ASVs at decreasing threshold cutoffs. OWC ASVs covered 36% and 51% of the composition of the wetlands database, at 99% and 97% threshold cutoffs, respectively.

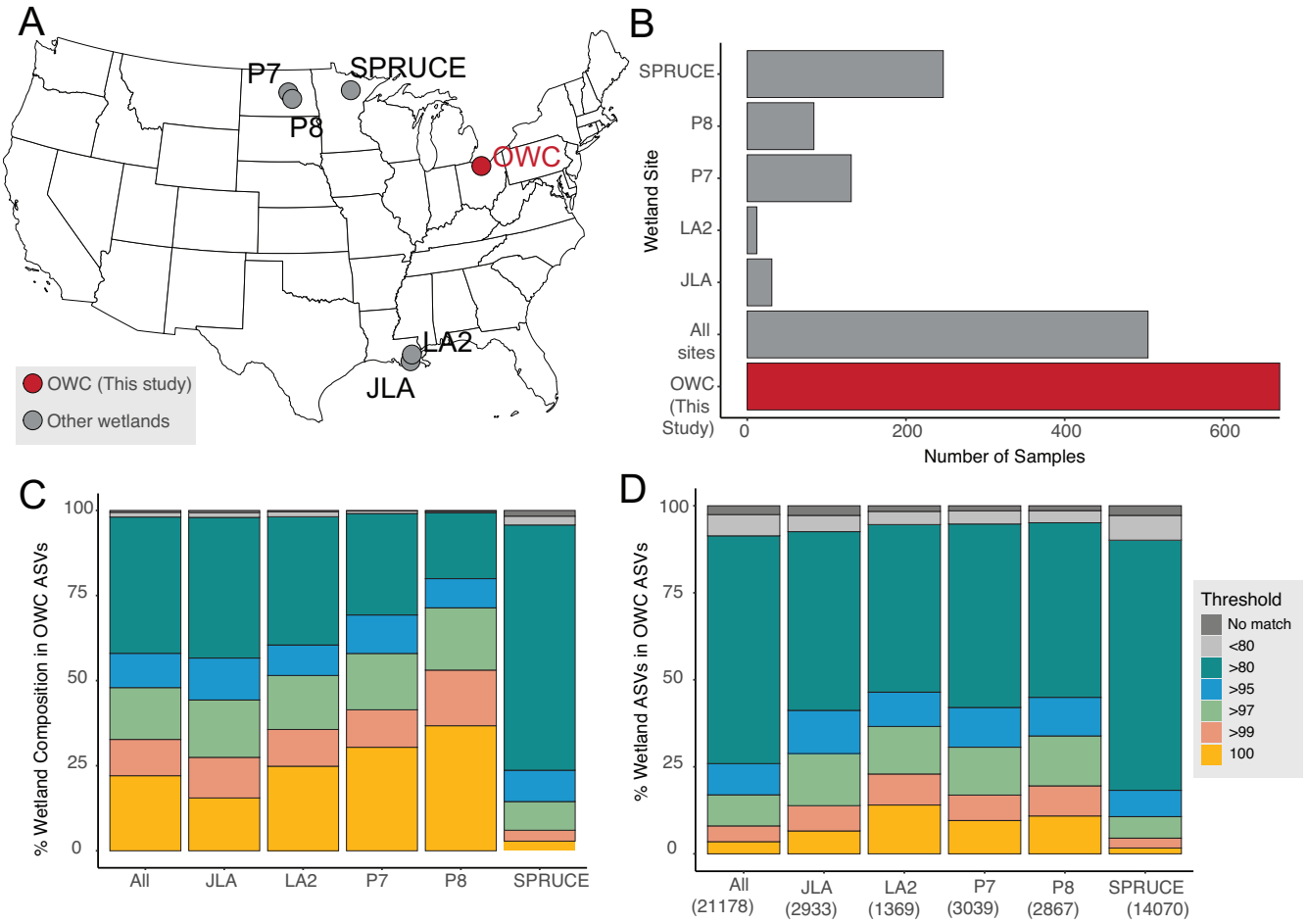
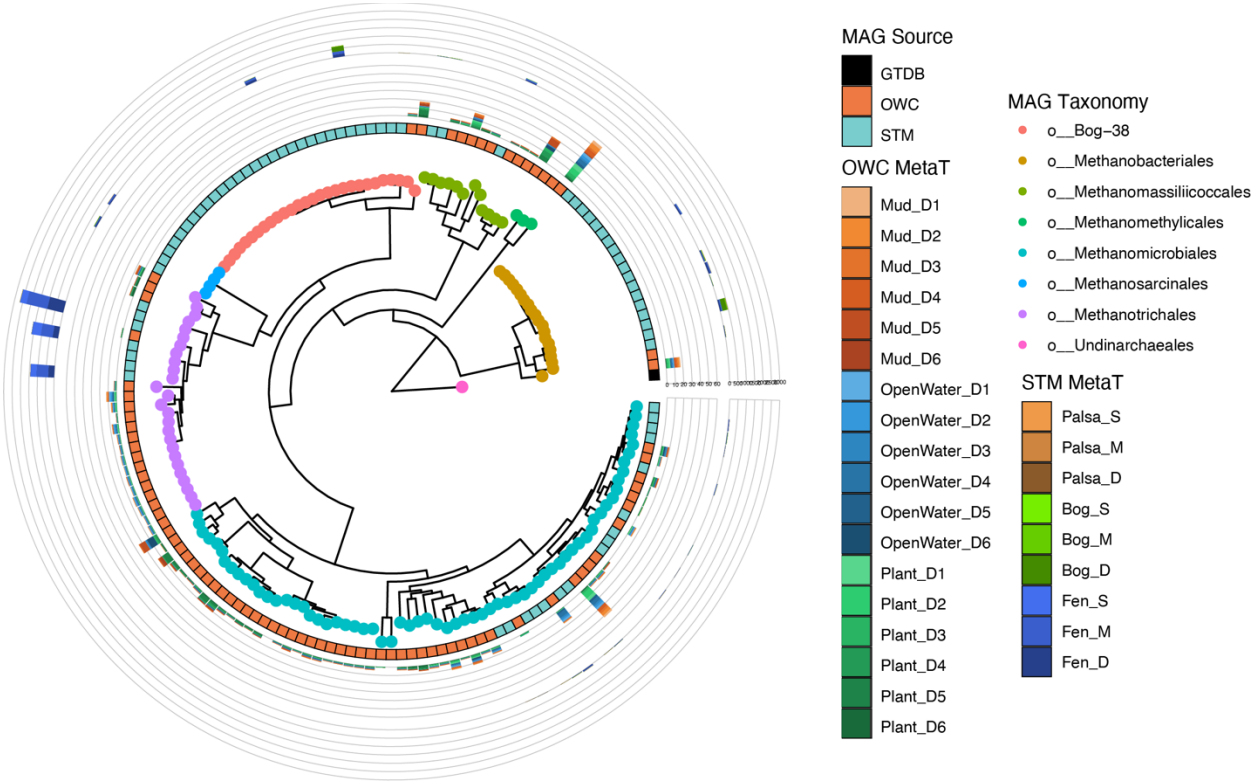
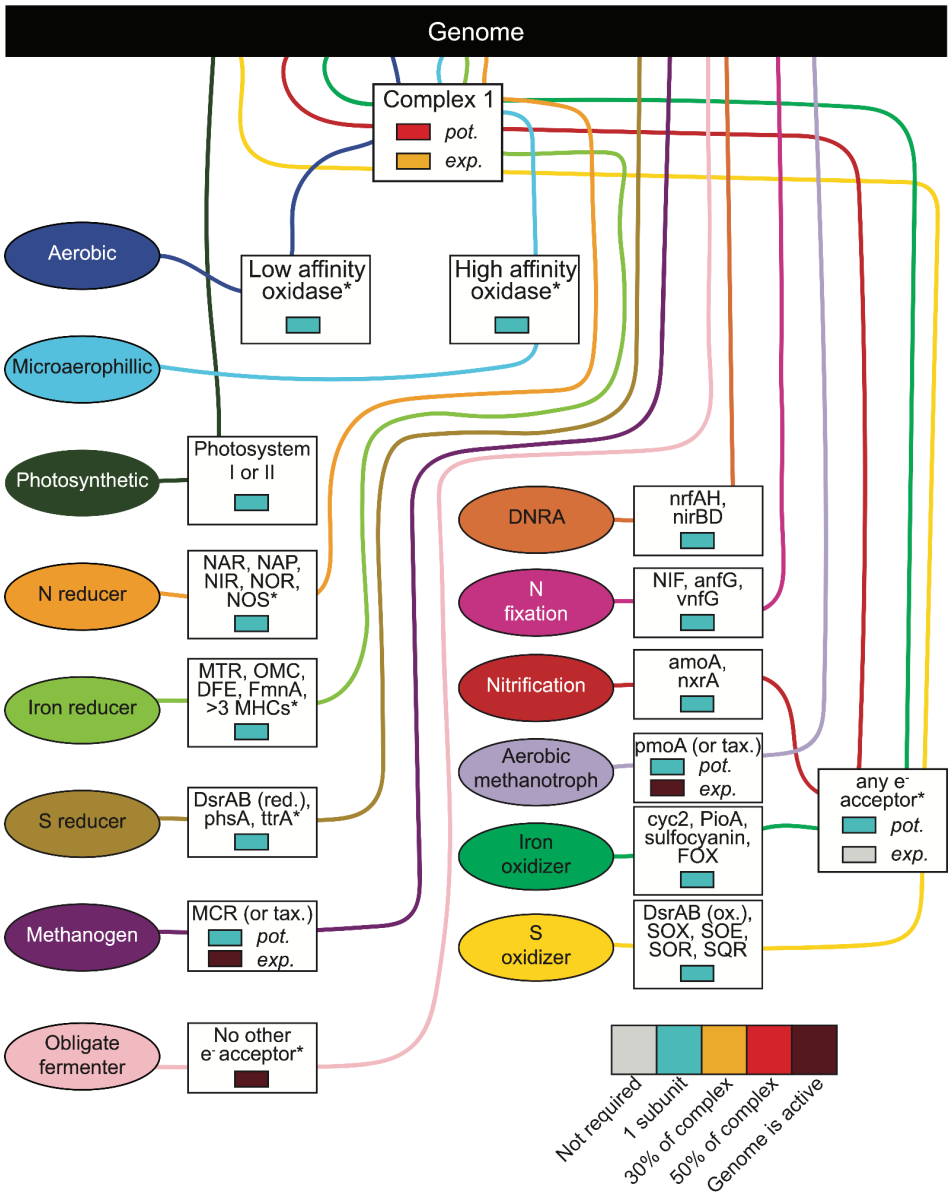


Figure S4. Phylogenetic tree of methanogens recovered from Old Woman Creek and Stordalen Mire shows taxonomic overlap while maintaining site specific MAGs. Phylogenetic tree was inferred using the GTDB-Tk de_novo_wf for methanogen MAGs derived from OWC (n=85) and Stordalen Mire (STM) (n=74), using p__Undinarchaeota as the outgroup. Leaves are colored by MAG methanogen taxonomy at the order level and MAG source is denoted by innermost ring. Two stacked bar charts highlight the average expression per MAG by depth and site in OWC (inner bar chart) and Stordalen Mire (outermost bar chart). The newick file supporting this analysis is accessible on zenodo DOI accompanying this manuscript (see data accessibility).



1212 **Figure S5. Overview of workflow for assigning MAGs to biogeochemical processes** The
1213 2,502 metagenome assembled genomes were characterized for functional potential and
1214 expression of 14 major biogeochemical contributions. Here we show how we used gene content
1215 to classify the aerobic, microaerophilic, photosynthetic, Nitrogen (N) reducer, Iron (I) reducer,
1216 Sulfur (S) reducer, methanogen, obligate fermenter, Dissimilatory Nitrite Reduction to
1217 Ammonium (DNRA), nitrogen (N) fixation, nitrification, aerobic methanotroph, Iron oxidizer,
1218 and Sulfur (S) oxidizer. The path diagram summarizes the gene suites required for a
1219 biogeochemical classification. For the subset of 1,948 transcribed genomes, we also assigned
1220 'expressed' metabolism from transcripts recruited back to MAGs. For some genes (e.g., *pmoA*),
1221 phylogenetic analyses and other non-homology-based methods were used for confirmation.
1222 Additional details are provided in the methods section, **MAG and Gene Based Functional**
1223 **profiling,** while the data derived from this approach is included in **Supplementary Table 5**.
1224



1225

Figure S6. Genome expression across the MAGs reported at phylum levels. Boxplots at the phylum level summarizing the distribution of the percentage of the genome that is transcribed for all 1948 MAGs in which transcription was detected (i.e., the number of genes transcribed over the total number of predicted genes for each MAG). Only phyla composed of at least 5 MAGs recovered are visualized here. Boxplots are ordered by percent of genome transcribed (from top to bottom). Data for this figure can be accessed via the zenodo DOI, file entitled **owc_metat_table_mags.csv**, which is metatranscriptomic expression per MAG across 133 metatranscriptomes (csv table).

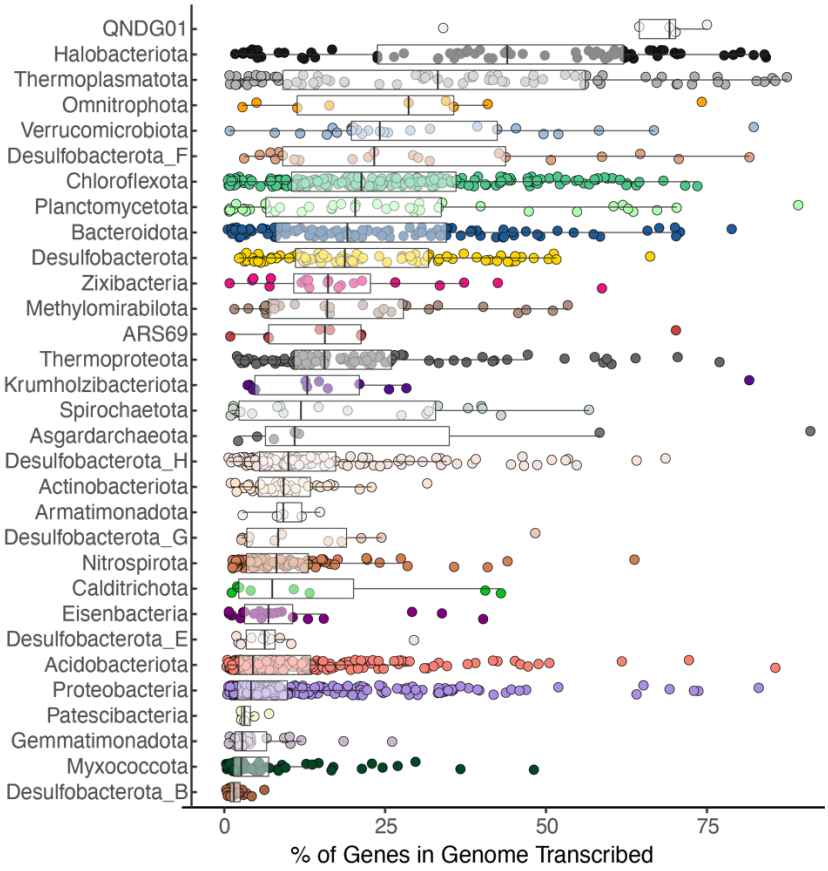


Figure S7. (A) PERMANOVA derived bar plots visualize the relative importance of spatiotemporal predictors included in the models for explaining the overall variation of composition, transcription, and guild microbiome information. **(B)** Mantel based bar plots visualize the mantel correlations of geochemical predictors (see **Supplementary Fig. S2**) included in multiple regression models for explaining overall variation. The model output data from PERMANOVA and mantel tests reported here is included in **Supplementary Table S6**.

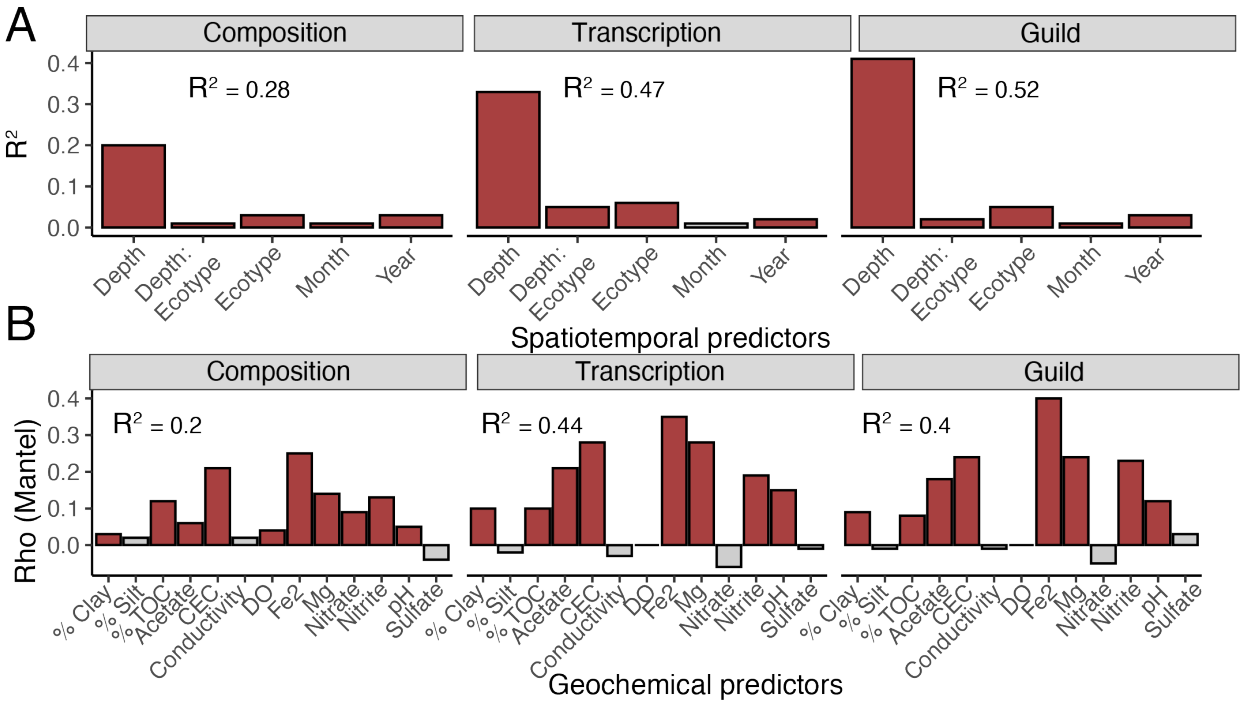


Figure S8. Methanogen substrate concentrations across years and depths. (A) Heatmap displays paired LCMS and NMR metabolite data, with samples ordered from left to right first by patch type (mud, open water) then by depth. Compound text color indicates detection methods including LCMS (black) or NMR (grey). Bar chart shows the number of samples each compound was detected in for all samples analyzed including 109 samples for NMR and 33 samples for LCMS. **(B)** and **(C)** Boxplots show log transformed NMR concentrations of acetate, formate, and methanol by depth and by year, respectively.

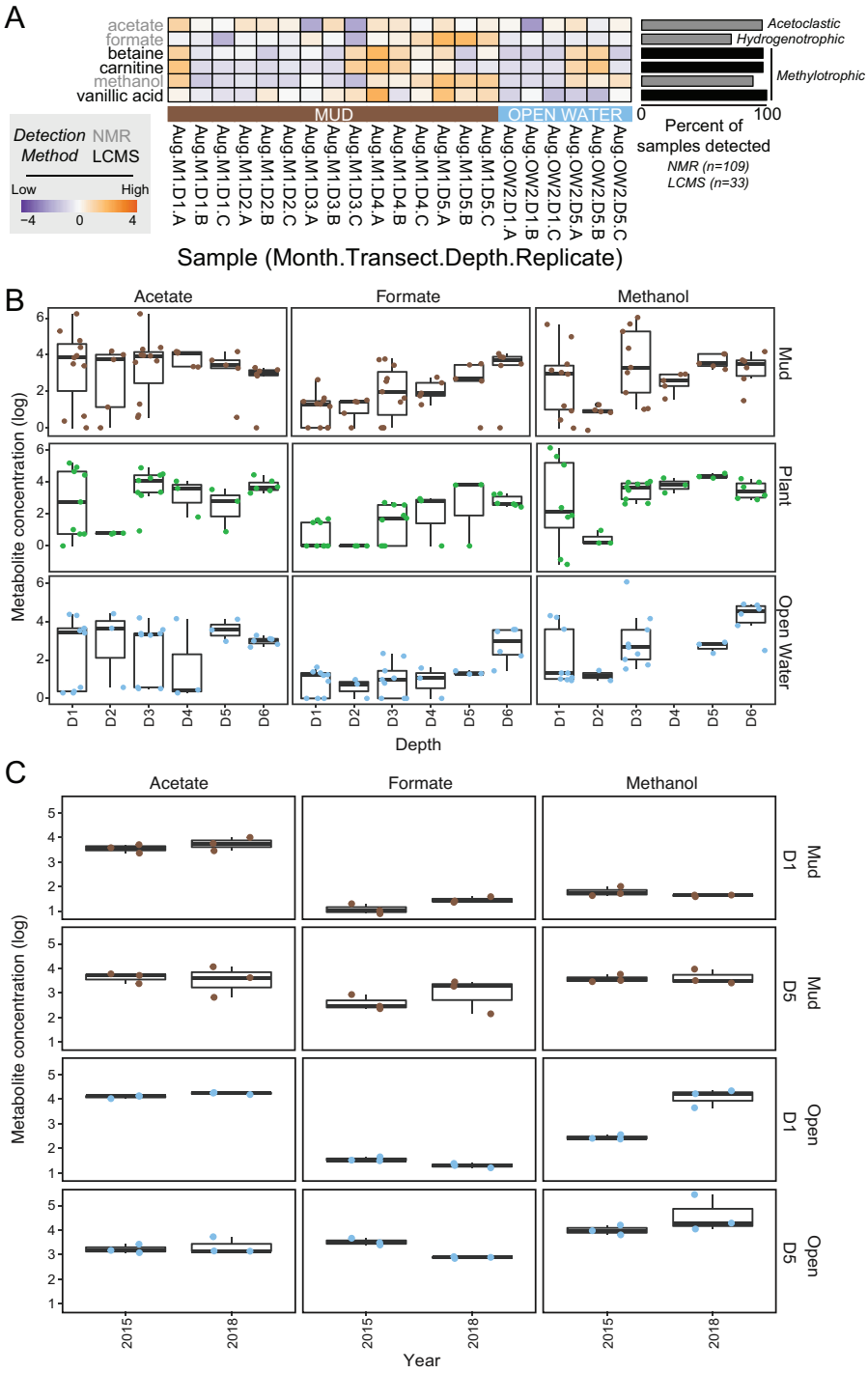


Figure S9. Methane cycling microorganisms show coordinated gene expression patterns, with groups co-localized by depth neighborhoods. By row metatranscriptome data from 3 cores at 3 depths (n=9) for each month (July, August, September) in blocks. Each column is the mean genome transcript for methanotroph or methanogen MAGs, with taxonomy and naming indicated on bottom and energy strategy denoted by box colors. Transcript response patterns are clustered by location, with some methanogens active in all depths in all seasons, some showing stronger surface, surface-middle, deep, deep-middle, and no clear depth signal. The data used to construct this figure is located in Supplementary Table S10 and on the zenodo DOI under file **owc_metat_table_mags.csv**.

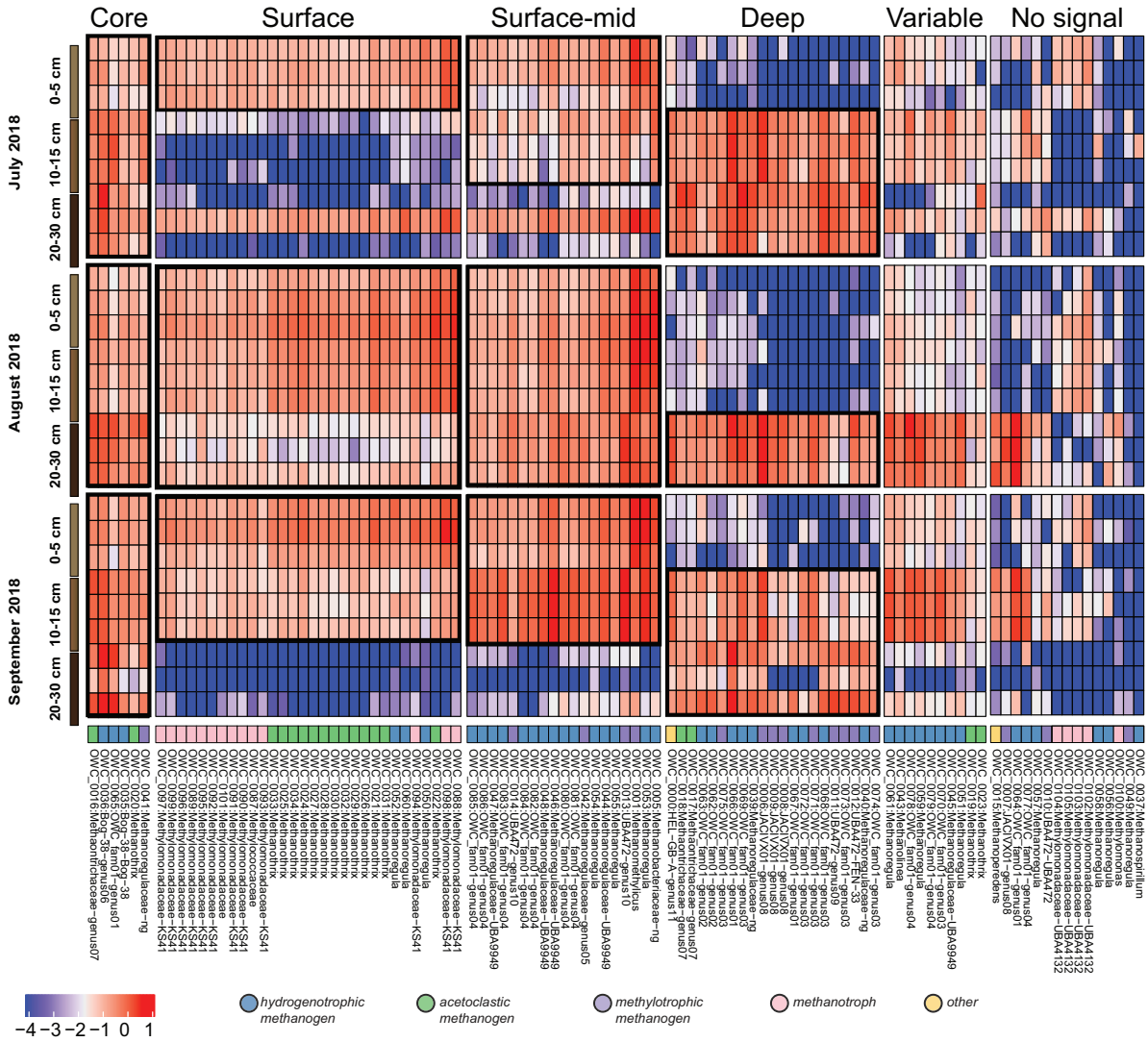
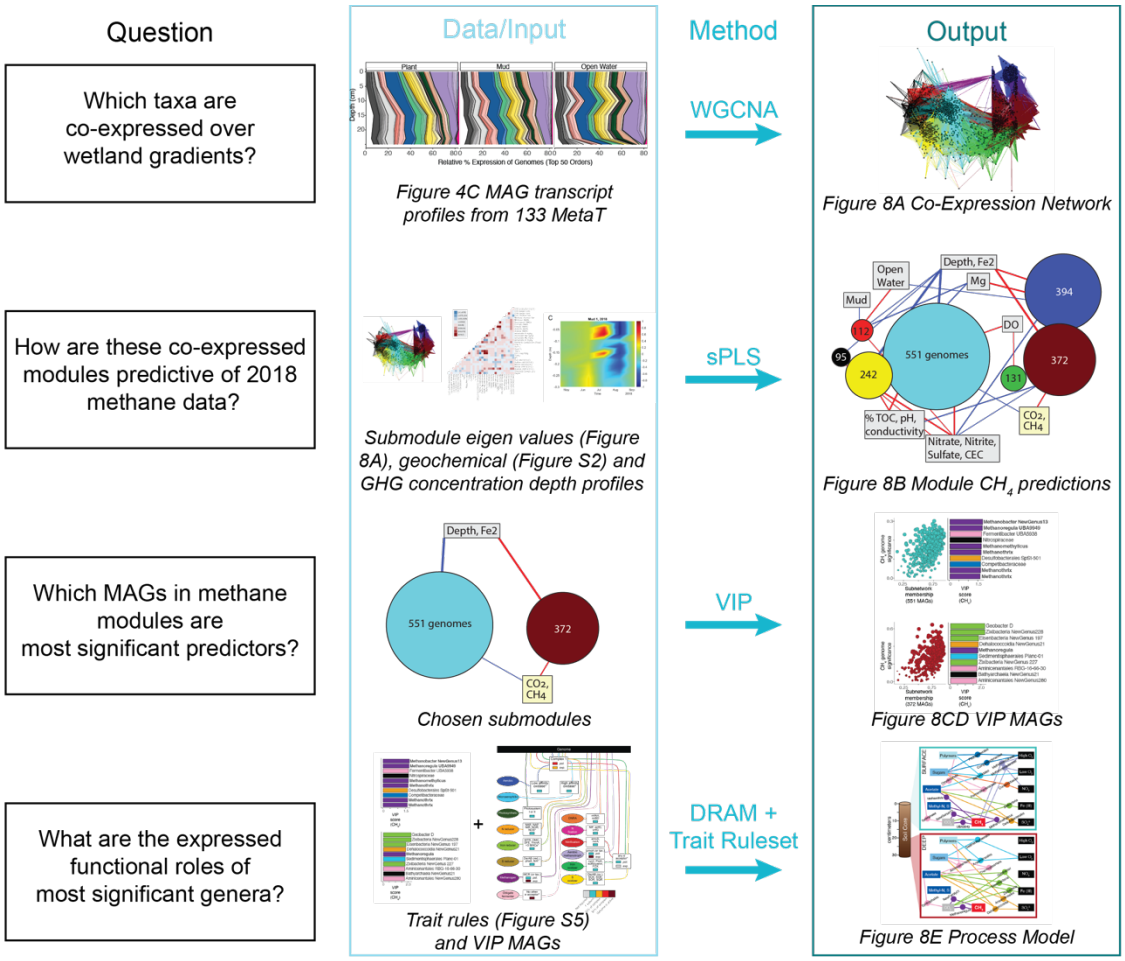


Figure S10. Workflow for data integration to identify functional guilds that predict soil methane. The left most column highlights the biological questions addressed (Questions). The middle column shows the actual data from other parts of this manuscript used to answer the question (Data/Input). The third column/arrow shows the statistical method used (Method). The fourth, most right column, shows the outcome of each analysis (Output). These output figures correspond to main text Figure 8. We denote the main text and supplementary figures referenced in the data inputs and outputs columns. Data supporting this figure is included in **Supplementary Tables S11** and **S12**.



Captions for Supplementary Tables S1-S12. The bolded pre-fix relates to the spreadsheet worksheet name.

Table S1. OWC_Samples: Overview and source details of the Old Woman Creek samples included in this study, including metagenomes, metatranscriptomes, and 16S rRNA gene amplicon sequences. For all metagenomes and metatranscriptomes, the corresponding NCBI accession number and sequencing depth (in Gbp) is provided.

Table S2. 16S: The 16S rRNA amplicon sequencing data from 671 samples. This includes the amplicon sequencing variants (ASV) and relative abundances across space and time. Bacterial and Archaeal taxonomic assignments are listed for each ASV.

Table S3. The bin stats: Data for the 2,502 medium and high-quality MAGs recovered and their attributes, including: the final database ID, a reference bin ID (includes sample bin was recovered from), the RED taxonomy from GTDB (including assigned novelty status), GTDB-Tk assigned taxonomy, contamination, completeness, genome size, number of scaffolds, number of contigs, N50 scaffolds, N50 contigs, % GC, coding density, predicted genes.

Table S4. MetaT metadata: Overview of the metadata associated with metatranscriptomics and metagenomics samples, including year collected, month collected, ecotype code, depth, core, site transect, associated geochemistry (pH, DO, CEC etc.), NMR determined metabolites, as well as peeper concentration and flux measurements.

Table S5. Bin_adjectives: The metabolic guild assignments for 2,502 MAGs in terms of both functional potential (e.g. metagenomic annotations) and transcribed potential (e.g. metatranscriptomic annotations, available for the subset of 1,948 MAGs that met our minimum transcription threshold), corresponding with Fig. 2.

Table S6. Environmental_drivers: Model outputs (mantel tests and PERMANOVA models) assessing the relationships between wetlands microbial communities (including the composition, transcription, and guild level) and spatiotemporal and geochemical variables to explain variation in the distribution of relative abundance. The R^2 values reported refer to PERMANOVA predictors and rho values refer to mantel tests (see Methods for details), with the data reported in Fig. 4 and Supplementary Fig. S7.

Table S7. Archaeal_Annotation_Calls: Functional assignments for Archaeal genomes, and gene-level requirements for assignments corresponding with Fig. 6.

Table S8. Metabolites: Soil metabolite concentrations as detected by LC-MS and NMR that are known and possible methanogen substrates. These data were collected across 2015 and 2018 from surface and deep cores. These data are reported in Supplementary Fig. S8.

Table S9. LCMS Metabolite annotation: Metabolite annotation data including specifics of LC-MS methodology. This information was used to generate the information in Table S8.

Table S10. Methanotroph_and_methanogen annotations: Genome inferred substrate use for methanogens and methanotrophs as defined by gene content. This table includes the gene-based rule set that profiled 80 genes to assign the substrate use categories from metagenome and metatranscriptome data in this study. These data are reported in Figs. 6-8.

Table S11. WGCNA_traits: Correlations between transcription of bacterial and archaeal genomes and geochemical attributes (pH, depth, total organic carbon, ecotype, etc.) detected via WGCNA network construction and analysis for each module detected.

Table S12. Process_model_map: Genomes included in focal WGCNA subnetworks (brown and turquoise) used to build process models (Fig. 8E). For each circle labeled in Fig. 8E, this table includes specific genomes in each transcriptional module, full taxonomy string, figure taxonomy label (column Label Name for Fig. 8), corresponding VIP scores and metabolism (redox guild, consumption, production, C fixation, sugars utilization, polymers, hemi-cellulose, dimers, short chain fatty acids, pyruvate to acetyl-coA).

In addition, we direct readers to the Zenodo DOI 10.5281/zenodo.8194033 (see **Data Availability**). This zenodo data set contains the following files, which are too large or not the appropriate format (excel format) to be included as supplementary table.

1. **MAGs.zip** Dereplicated database of 2,502 MAGs (fasta files)
2. **owc_HQMQ_DB_genes.faa.gz** MAG amino acid gene sequences derived from DRAM gene calls (fasta file)
3. **owc_HQMQ_DB_ANNOTATIONS_20220208.txt.gz** MAG DRAM annotations
4. **owc_metat_table_mags.csv** Metatranscriptomic expression per MAG across 133 metatranscriptomes (csv table)
5. **owc_metat_table_mags_genes.csv** Metatranscriptomic expression per gene across 133 metatranscriptomes (csv table)
6. **owc_metat_table_mags_genes_annotations.csv** corresponding DRAM annotations to #5 for transcribed genes (csv table)
7. **gtdbtk.ar53.decorated.tree** newick file for tree for methanogen based comparisons across sites

SI References

1. J. C. Angle, *et al.*, Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions. *Nat Commun* **8**, 1567 (2017).

2. A. B. Narrowe, *et al.*, High-resolution sequencing reveals unexplored archaeal diversity in freshwater wetland soils. *Environ Microbiol* **19**, 2192–2209 (2017).

3. G. J. Smith, *et al.*, Members of the Genus Methylobacter Are Inferred To Account for the Majority of Aerobic Methane Oxidation in Oxic Soils from a Freshwater Wetland. *mBio* **9**, e00815-18.

4. A. B. Narrowe, *et al.*, Uncovering the Diversity and Activity of Methylotrophic Methanogens in Freshwater Wetland Soils. *mSystems* **4**, e00320-19 (2019).

5. J. B. Ellenbogen, *et al.*, Methylotrophy in the Mire: direct and indirect routes for methane production in thawing permafrost. *mSystems* **0**, e00698-23 (2023).

6. J. G. Caporaso, *et al.*, Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**, 1621–1624 (2012).

7. E. Bolyen, *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852–857 (2019).

8. B. J. Callahan, *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581–583 (2016).

9. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596 (2013).

10. D. H. Parks, *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).

11. M. Borton, *et al.*, Targeted curation of the gut microbial gene content modulating human cardiovascular disease. *mBio* **14**, e0151123 (2023).

12. A. B. Narrowe, *et al.*, Complex Evolutionary History of Translation Elongation Factor 2 and Diphthamide Biosynthesis in Archaea and Parabasalids. *Genome Biology and Evolution* **10**, 2380–2393 (2018).

13. najoshi, sickle - A windowed adaptive trimming tool for FASTQ files using quality. (2023). Deposited 16 June 2023.

14. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

15. Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

16. B. Bushnell, "BBMap: A Fast, Accurate, Splice-Aware Aligner" (Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), 2014).

1401 17. D. D. Kang, *et al.*, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
1402 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

1403 18. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate genomic
1404 comparisons that enables improved genome recovery from metagenomes through de-replication.
1405 *ISME J* **11**, 2864–2868 (2017).

1406 19. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the
1407 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*
1408 **25**, 1043–1055 (2015).

1409 20. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: a toolkit to classify genomes
1410 with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

1411 21. R. M. Bowers, *et al.*, Minimum information about a single amplified genome (MISAG) and a
1412 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731
1413 (2017).

1414 22. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective
1415 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and*
1416 *Evolution* **32**, 268–274 (2015).

1417 23. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and
1418 annotation. *Bioinformatics* **23**, 127–128 (2007).

1419 24. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: an r package for visualization and
1420 annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology*
1421 *and Evolution* **8**, 28–36 (2017).

1422 25. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput
1423 sequencing data. *Bioinformatics* **31**, 166–169 (2015).

1424 26. M. Smid, *et al.*, Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq
1425 data performs similarly in intersample analyses while improving intrasample comparisons. *BMC*
1426 *Bioinformatics* **19**, 236 (2018).

1427 27. B. J. Woodcroft, CoverM. (2023). Deposited 26 June 2023.

1428 28. M. Shaffer, *et al.*, DRAM for distilling microbial metabolism to automate the curation of microbiome
1429 function. *Nucleic Acids Research* **48**, 8883–8900 (2020).

1430 29. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput.
1431 *Nucleic Acids Res* **32**, 1792–1797 (2004).

1432 30. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC*
1433 *Bioinformatics* **9**, 559 (2008).

1434 31. H. Abdi, Partial least squares regression and projection on latent structure regression (PLS
1435 Regression). *WIREs Computational Statistics* **2**, 97–106 (2010).

1436