

1 **Metabolic overlap in environmentally diverse microbial communities**

2 Eric R. Hester^{1*}, Mike S.M. Jetten¹, Cornelia U. Welte¹, Sebastian Lücker¹

3 ¹ Department of Microbiology, Radboud University, Nijmegen, The Netherlands

4 * Correspondence:

5 Eric R. Hester

6 ericokh@gmail.com

7

8 ***Abstract***

9 The majority of microbial communities consist of hundreds to thousands of species, creating a
10 massive network of organisms competing for available resources within an ecosystem. In
11 natural microbial communities it appears that many microbial species have highly redundant
12 metabolisms and seemingly are capable of utilizing the same substrates. This is paradoxical,
13 as theory indicates that species requiring a common resource should outcompete one another.
14 To better understand why microbial species can co-exist, we developed Metabolic Overlap
15 (MO) as a new metric to survey the functional redundancy of microbial communities at the
16 genome scale across a wide variety of ecosystems. Using metagenome-assembled genomes,
17 we surveyed over 1200 studies across ten ecosystem types. We found the highest MO in
18 extreme (i.e., low pH/high temperature) and aquatic environments, while the lowest MO was
19 observed in communities associated with animal hosts, or the built/engineered environment.
20 In addition, different metabolism subcategories were explored for their degree of metabolic
21 overlap. For instance, overlap in nitrogen metabolism was among the lowest in Animal and
22 Engineered ecosystems, while the most was in species from the Built environment. Together,
23 we present a metric that utilizes whole genome information to explore overlapping niches of
24 microbes. This provides a detailed picture of potential metabolic competition and cooperation
25 between species present in an ecosystem, indicates the main substrate types sustaining the
26 community and serves as a valuable tool to generate hypotheses for future research.

27

28 ***Introduction***

29 Microorganisms drive global biogeochemical cycles, but they do not work or live in isolation.
30 In order for any living species to survive they must engage in competition for space and
31 resources with other organisms that share similar nutritional requirements. The concept of loss
32 of species less adapted relative to their competitors is known as competitive exclusion (Gause
33 1934). When one species cannot sufficiently persist in a habitat, they become locally extinct.
34 Through selection of traits that reduce the dependence on a common resource, populations
35 may shift towards coexistence. This is known as niche partitioning, whereby competition is
36 avoided through the utilization of different resources (Schoener 1974). Evidence that these
37 ecological and evolutionary forces shape microbial communities is prevalent in literature;
38 however, the strength of these forces varies with the availability of resources (reviewed in
39 (Nemergut et al. 2013).

40 Describing a niche of an organism has remained challenging ever since the concept first
41 emerged (Hutchinson 1957). Typically, closely related species are thought to share similar
42 niches, assuming their evolutionary relatedness is reflected in their nutritional requirements.
43 Recently, neutral genetic markers have emerged as a proxy to measure species' divergence on
44 an evolutionary timescale; however, these phylogenetic markers (i.e., 16S rRNA genes) are
45 unsuitable to evaluate differences in the biochemical capacity of the organisms. Whole
46 genomes contain information relevant to the metabolic capacity of a species, which is
47 essential to describe the putative niches a microbial species may occupy. If one were to ask
48 about the overlap of two microorganisms' niches, it is conceivable that this is akin to asking
49 how similar the two are on a genomic level.

50 With the continued advancement in high-throughput DNA sequencing, large amounts of
51 genomic data are frequently released and available for public use. Several recent publications
52 have reported thousands of novel bacterial and archaeal metagenome-assembled genomes
53 (MAGs; Anantharaman et al. 2016; Delmont et al. 2018; Parks et al. 2017; Tully, Graham,
54 and Heidelberg 2018). The sequencing data originated from hundreds of studies investigating
55 different ecosystems, such that these genomes represent a diverse set of taxa from ecosystems
56 around the globe. This presents an opportunity to address the following important questions:
57 how variable is niche overlap in microbial communities across different ecosystems and does
58 the nature of the overlap (i.e., abundance of genes involved in nitrogen cycling) change based
59 on habitat?

60 In the current study, we surveyed niche overlap in microbial communities by searching for
61 shared pathways in the metabolic reaction network of species within these communities,
62 which we refer to as ‘metabolic overlap’ (MO). This approach was used to investigate two
63 main questions. First, does the degree of niche overlap in microbial communities vary
64 between ecosystems (i.e., do some communities have more species that utilize the same
65 substrates)? Second, how do these microbial communities vary in the degree of overlap of
66 different metabolic categories (i.e., nitrogen or sulfur metabolism)?

67 We observed patterns of overlap in microbial community members’ metabolism across
68 different ecosystems, which were largely consistent with literature reports. For instance, a low
69 degree of MO was found in microorganisms involved in highly specialized animal host-
70 microbe associations, while aquatic microbes displayed a cosmopolitan repertoire of strategies
71 for nutrient acquisition. These variations seem to be driven by different categories of
72 metabolism, depending on the ecosystem. In addition, we addressed the question of how
73 much the phylogenetic relationship of microbes corresponds to their metabolic overlap. We
74 found that phylogenetic distance between microorganisms was indeed a good predictor for the
75 degree of MO. The strength of this relationship, however, varied between different
76 ecosystems. Generally, survey-based metrics like MO enable observations of global trends
77 and prompt fundamental questions about the biology and ecology of microorganisms.

78

79

80

81 **Results**

82 *Definition of metabolic overlap.*

83 We defined metabolic overlap (MO) as the number of compounds (i.e., reactants) that can be
84 utilized by two organisms based on their shared metabolic network (Figure 1). For example,
85 an organism (Org₁) that can perform all steps of denitrification from nitrate (NO₃⁻) to nitrogen
86 gas (N₂, four reactions in total) shares two reactants with a partially denitrifying organism
87 (Org₂) that only reduces NO₂⁻ to N₂O. This then results in a MO = 2 (ignoring the rest of their
88 metabolism). Conceivably, identifying MO allows a broad identification of species with
89 overlapping niches by counting the compounds that link complimentary metabolic pathways.
90 As the metabolic routes used to degrade certain substrates can vary between organisms,
91 counting the number of shared reactants will reveal MOs that would not be uncovered by
92 shared reactions only. Furthermore, as the number of reactants can vary between reactions,
93 this approach is more sensitive in identifying weak metabolic similarities between organisms.

94 We acknowledge that previous efforts to predict microbe-microbe interactions within
95 microbial communities have been made with similar logic to the current approach. In
96 particular, the NetCooperate software, utilizing the NetSeed framework, is a method to
97 identify putative interactions in a community. It does so by using genome information to
98 predict auxotrophies of the organisms present, based on the incompleteness of certain
99 biosynthesis pathways leading to a dependency of the respective organism to external sources
100 of the lacking metabolite (Levy *et al.*, 2015; Carr and Borenstein, 2012). Thus, the
101 NetSeed/NetCooperate approach predicts complementarity between species, which
102 consequently occupy distinct niches, while the goal of our MO approach is to identify to what
103 extent two species fill a common niche.

104 *Metabolic overlap of microbial communities in different ecosystems.*

105 In order to survey the degree of MO in various ecosystems from around the globe, thereby
106 identifying the degree in which microbial species within the community overlap in the niches
107 they fill, the set of Uncultivated Bacteria and Archaea (UBA) MAGs published by Parks and
108 colleagues was utilized (Parks *et al.*, 2017). The average predicted genome completeness of
109 these MAGs ranged from 50-100%. A completion-based inclusion threshold of MAGs was
110 found to have a negligible impact on the average MO of communities (Supplemental Figure
111 1). In contrast, the number of MAGs included drastically decreased as a result of a more
112 stringent threshold on genome completeness, resulting in ecosystems poorly or not at all

113 represented (Supplemental Figure 1). Thus, we included all 7903 MAGs from the Parks et al.
114 dataset, representing 1248 studies. Studies were classified into their respective ecosystems of
115 origin based on information included in the submission to the public repository or by manual
116 curation if this information was insufficient. This resulted in ten ecosystem categories, with
117 studies that could not be reasonably identified classified as “Other” (Table 1).

118 In a given ecosystem, metabolic overlap and the predicted average genome sizes of MAGs
119 were strongly correlated (Supplemental Figure 2; $p < 0.01$). In addition, average genome sizes
120 significantly varied between ecosystems (Supplemental Figure 3; ANOVA; $F = 88$; $p <$
121 0.001). The average predicted genome sizes were the highest in studies from the built
122 environment (4Mbp +/- 0.65Mbp) and lowest in extreme environments (2Mbp +/- 0.96Mbp;
123 Table 2). The number of MAGs in a given community (grouped per study) negatively
124 correlated with the average MO of the community (Figure 2; Kendall’s tau = -0.38; $p <$
125 0.001). As we were interested in investigating how MO varied between ecosystems,
126 irrespective of the differences in genome sizes between ecosystems, we normalized MO to the
127 average genome size of the respective study. Furthermore, the values were scaled so that the
128 average MO of all ecosystems combined was 0 (Figure 3).

129 To evaluate how the MO of microbial communities varied between ecosystems, we
130 determined how the average MO of a single ecosystem differed from the average MO of all
131 ecosystems. Communities from Animal, Built, and Engineered ecosystems had significantly
132 lower MO than average (t-test; $p < 0.01$; Table 3; Figure 3). On the contrary, those from
133 Extreme, Freshwater and Marine ecosystems had significantly higher MO than average (t-test;
134 $p < 0.01$; Table 3; Figure 3).

135 *Breakdown of MO scores across different ecosystems to different levels of metabolism*

136 To investigate how metabolic overlap varied between ecosystems within different categories
137 of metabolism (SEED subsystems), the MO within these subcategories was determined for
138 each ecosystem and compared to the average value of all ecosystems (Table 4). Animal, Built
139 and Engineered ecosystems were generally below the average MO for the majority of
140 subcategories of metabolism with a few exceptions (t-test; $p < 0.01$; Table 4). Communities
141 from Engineered ecosystems had an above average MO in Protein and Nucleotide sugar
142 metabolism, as did communities from Animal ecosystems. In addition, communities from the
143 Animal ecosystem had an above average MO in Nucleotide metabolism. While most
144 subcategories of metabolism from the Built environment were below the average MO, these

145 communities contained higher MO in Nitrogen and Sulfur metabolism (Table 4). In contrast
146 to the above communities, which were dominated by lower than average MO scores, Extreme,
147 Freshwater, and Marine ecosystems had higher than average MO scores in the majority of the
148 categories of metabolism (Table 4).

149 The nitrogen metabolism was used to further investigate the influence of incomplete pathways
150 on the MO. Therefore, the ratios of complete to incomplete denitrifiers were calculated for all
151 ecosystems (i.e., complete denitrifiers encoding all proteins required for NO_3^- , NO_2^- , NO , and
152 N_2O reduction; incomplete denitrifiers missing at least one gene; Figure 4A). The Built
153 environment showed the largest MO in nitrogen metabolism and also had the highest ratio of
154 complete to incomplete denitrifiers compared to all other ecosystems (Figure 4B). Contrary,
155 the Animal ecosystem, which by far had the lowest MO in this category also contained mostly
156 incomplete denitrifiers.

157 *Phylogenetic relationship of organisms and its relationship to the metabolic overlap.*

158 In order to determine if the evolutionary relatedness between MAGs was correlated with MO,
159 the UBCG pipeline was utilized to infer a phylogenetic tree based on a concatenated
160 alignment of 92 universal bacterial marker genes (Na et al. 2018). A significant negative
161 correlation was observed between phylogenetic distance and metabolic overlap for all
162 ecosystems (Figure 5; $r = -0.33$; $p < 0.001$), however the strength of this association varied.
163 Phylogenetic distance and MO had the strongest association in Plant ($r = -0.64$), Built ($r = -$
164 0.53) and Marine ecosystems ($r = -0.47$), whereas the lowest associations were seen in
165 Animal ($r = -0.16$), Extreme ($r = -0.19$) and Fresh Water ecosystems ($r = -0.21$; Figure 6).

166 **Discussion**

167 In the current study a new metric termed MO, which describes how similar two species'
168 metabolisms are, was developed in the context of a genome-based survey of microbial
169 communities from diverse ecosystems. High MO between two species suggests that they have
170 the capacity to perform similar metabolic reactions, thus have similar growth requirements
171 and fill similar niches. In contrast, low MO suggests that the two species in question may
172 compete for fewer resources. We determined the average metabolic overlap of all community
173 members (i.e., the average MO of all pairwise species comparisons) for a given study, which
174 were grouped into distinct ecosystems based on their origin for comparison (Figure 3; Table
175 3). The average MO of a community can be similarly interpreted as the pairwise species
176 comparisons. In the case of high average overlap, many community members are overlapping
177 in their biochemistry and could in theory compete for a similar niche, whereas a low average
178 MO would suggest the opposite.

179 *Ecological and evolutionary drivers of metabolic overlap*

180 There are several well studied ecological forces that shape microbial community
181 structure. Community diversity is maintained via dispersion (immigration and emigration) as
182 well as speciation and extinction. In studying patterns of microbial biogeography, dispersion
183 limitations were seen as one of the driving forces in structuring microbial community patterns
184 in salt marshes and rice paddies, and likely have an influence on the genomic adaptations of
185 marine microorganisms (Kelly et al. 2014; Lüke et al. 2014; Martiny et al. 2006). Microbial
186 biogeography theory has also been applied to help understanding compartmentalized host-
187 associated microbial communities such as microbes in the human lungs (Whiteson et al.
188 2014). In this study, we observed major ecosystem-dependent differences in the MO of
189 microbial community members (Figure 3; Table 3). This variation may in part be attributed to
190 dispersion limitations inherent to each ecosystem, where ecosystems in which the dispersion
191 of microbial community members is limited would have less overlap than open homogenous
192 ecosystems. Accordingly, the highest MO was observed in aquatic ecosystems, namely
193 communities from the marine open ocean environment, while animal host-associated
194 communities contained some of the lowest MO (Figure 3; Table 3). Ecosystems such as the
195 ocean are likely to not have as strong dispersal limitations as ecosystems like the animal gut
196 or human lungs, and these differences may be a driving force in structuring the MO of their
197 respective microbial communities.

198 In addition to dispersion as an ecological force, disturbances to ecosystems can also
199 play a large role for species diversity, driving extinction or speciation within the community
200 (Buckling et al. 2000; Connell 1978). Varying degrees of disruption would impart some
201 signature on the metabolic pathways represented in the microbial community. A higher
202 frequency of disturbance would contribute to the extinction of species and reduce the number
203 of redundant metabolisms in a given system. For example, disturbances associated with the
204 marine ecosystem (high MO) such as storms or temperature anomalies are likely less frequent
205 and intense than the regular consumption of foodstuff or intermittent bouts of inflammation in
206 animal guts (low MO) (David et al. 2014; Kashyap et al. 2013; Reese et al. 2018).

207 *Substrate spectrum as a possible driver of metabolic overlap in ecosystems.*

208 The availability of resources, both in quality and quantity, drives which species can thrive in a
209 given system. In the open ocean, the input of labile organic matter is a major factor
210 controlling microbial activity in the photic zone, where phototrophs fix large quantities of
211 inorganic carbon, making new organic matter available to heterotrophic organisms (Aylward
212 et al. 2015; Hansell and Carlson 2002). It is understood that differences in the composition of
213 dissolved organic matter (DOM) enrich for different clades of microorganisms and that the
214 composition of the community is highly influential on the capacity to degrade this carbon
215 (Nelson et al. 2013; Solden et al. 2018). It would follow that a higher substrate selection
216 would drive diversity in the microbial community, and the higher diversity of substrates
217 would then lead to more diverse microbial metabolisms. In the current study, a negative
218 relationship between the richness of a community (number of genomes in a given sample) and
219 their average MO was observed, which suggests that in more diverse communities there is
220 less metabolic overlap (Figure 2). Indeed, there are many studies that report species-specific
221 differences in the composition of host-associated microbial communities ranging from plants
222 to animal hosts (Berg et al. 2014; E.R. Hester et al. 2016; Reese et al. 2018). These
223 differences are in part attributed to the selection of organic compounds that are shared from
224 host to symbiont (Lee et al. 2016; Sasse, Martinoia, and Northen 2018; Zhalnina et al. 2018).

225 In addition to the quality of substrates, the quantity of organic matter also drastically differs
226 between ecosystems. The concentration of DOC can vary greatly in aquatic systems, with
227 around 40 $\mu\text{mol l}^{-1}$ DOC in groundwater and 5000 $\mu\text{mol l}^{-1}$ in swamps and marshes
228 (Søndergaard and Thomas 2004). Likewise, variations in animal's diet influence the
229 availability of different substrates for microorganisms. In particular the diet of an animal

230 influences the availability of nitrogen to microbes in animal guts (Reese et al. 2018). Equally,
231 N availability has a strong impact on plant-soil feedbacks, influencing the abundance and
232 metabolism of microorganisms in the rhizosphere (Eric R Hester et al. 2018). If substrates are
233 available in high enough concentrations, the effect of competition may be reduced, potentially
234 leading to a higher number of species consuming a common substrate (i.e., higher MO). In the
235 current study, we observe microbial communities from animal ecosystems had the lowest
236 overlap in categories of metabolism involved in nitrogen and amino acid metabolism, which
237 corresponds to the idea of N limitations in the animal gut and known auxotrophies (Table 4;
238 Reese *et al.*, 2018; Zengler and Zaramela, 2018). In contrast, microbial communities from the
239 built environment tend to have higher overlap in nitrogen and sulfur metabolism, though the
240 built environment is a loosely defined ecosystem with limited literature detailing nutrient
241 fluxes through the system (Table 4; Adams et al. 2015). This stark contrast of nitrogen
242 metabolism overlap between the Built and Animal ecosystems, which both generally
243 displayed a lower than average MO, corresponded to the observed number of species capable
244 of complete denitrification. The Built ecosystem had the highest nitrogen metabolism MO,
245 which largely was attributed to the highest proportion of microbial species capable of
246 complete denitrification (Figure 4). This was contrasted by the low number of complete
247 denitrifiers in the animal system. While the differences here could be due to nutrient
248 availability, one should also consider possible differences in life strategies for persisting in a
249 particular environment (i.e., detoxification versus energy conservation).

250 *Influence of phylogenetic relationship on metabolic overlap.*

251 Populations that become isolated and diverge on an evolutionary timescale do so as a result of
252 being exposed to different environments and thus different selection pressures on specific
253 traits, although some mechanisms exist that make this divergence less clear (i.e., convergent
254 evolution, horizontal gene transfer, etc.). In the current study, a relationship was observed
255 between the MO of species and their relatedness (Figure 5), with a reduction of MO with
256 increasing taxonomic distance. While this corresponds well to theory, the strength of the
257 relationship between phylogenetic relatedness and MO varied between ecosystems,
258 suggesting that ecological differences between these ecosystems influence this relationship.

259 The dominant taxonomic groups often vary between different ecosystems as a result of the
260 underlying nutrient profiles or physical properties of those ecosystems. This may be a result
261 of stronger selection pressures in a given ecosystem for traits specific to a few select

262 monophylogenetic groups (i.e., methanogenesis, ammonia and nitrite oxidation), as opposed
263 to traits that are more widespread (i.e., denitrification). Phylogenetic groups may vary in the
264 number of traits (i.e., some groups are more metabolically versatile than others), and MO is
265 determined by the number of reactions a given pair of species share. For example,
266 Zimmerman et al., found that a set of phylogenetically diverse Bacteria and Archaea had the
267 potential to produce a subset of three extracellular enzymes (Zimmerman, Martiny, and
268 Allison 2013). Specifically, the ability to produce these enzymes was non-randomly
269 distributed phylogenetically. It follows that ecosystems which have strong selection pressures
270 for metabolically diverse phylogenetic groups would have a weaker relationship between the
271 phylogenetic relatedness and metabolic overlap.

272 *Caveats and limitations of genetic predictions of metabolic overlap.*

273 The emergence of vast amounts of sequence data has allowed the assembly of genomes of
274 microorganisms from fragmented DNA isolated from the environment. The degree of
275 information in whole genomes compared to that from marker genes (both phylogenetic and
276 metabolic) is likely to provide significant advances in our understanding of the genetic
277 organization of microorganisms. In addition, knowing that a certain set of genomes were
278 physically in the same sample is advantageous in addressing fundamental questions about the
279 ecology and evolution of microbial communities from natural settings. Unfortunately, there
280 are still significant limitations when dealing with metagenome-assembled genomes.
281 Specifically, the amount of information lost in the process of genome assembly and binning
282 reduces our understanding of population-level genetic variation. Current sequencing depths do
283 not provide sufficient coverage for the metagenomic assembly of low abundance organisms'
284 genomes, narrowing our view of genetic linkages between species towards the highly
285 abundant species. However, these are mainly technological limitations, with solutions like
286 long read sequencing becoming increasingly more available. Additionally, there is a
287 significant lack of information about the environments in which samples were taken in the
288 public archives, limiting what can be assessed with metrics such as metabolic overlap, and
289 calling for an urgent need to provide as much metadata on samples as possible.

290 In addition to the technical limitations mentioned above, there are also limitations in
291 methods such as MO, which rely heavily on accurate automated annotation of genetic
292 elements in genomes. Specifically, database quality is a key driver in the accuracy of survey
293 studies such as the one presented here. A major issue is the inability to assign functions to

294 many genes, even in the genomes of the most well studied microorganisms (35% hypothetical
295 proteins in *E. coli* genome; Ghatak et al. 2019). Apart from the limitations to automatic
296 annotation methods, there are different levels of biology associated with niches that are not
297 captured in genome-level information. These limitations include a lack of information of
298 whether a gene is transcribed, whether the transcript is translated to a functional product and
299 ultimately variations in affinity and activity of this protein. The variation in transport
300 efficiency and regulatory mechanisms certainly contributes to the competitive advantage of an
301 organism and thus the niche this organism fills. These complexities are not easily derived
302 from genomic information. Idealistically, as emphasized by (Bowers et al. 2017), in order to
303 improve discovery-based approaches that rely on machine readable formats of public
304 repositories, additional information should accompany MAG submissions. This set of
305 information would not only help assess the quality of the genome but aid in associating the
306 genetic information to the biology and ecology of the organism. Ideally, such information
307 should include conditions of the environment from which the species' genome was obtained
308 (i.e., pH and temperature), and if the species was cultivated, any physiological parameters that
309 may have been measured (i.e., growth rate, substrate usage profile and affinities, etc.).

310 **Conclusions**

311 The observation of variation in MO across different ecosystems begs several questions about
312 the nature of microbial community metabolism. Specifically, what drives metabolic versatility
313 in microbial communities? Are there generalizable rules that can be deduced? Survey-based
314 studies enriched with additional information, such as those highlighted above, may shed
315 additional light on important factors that drive MO. In addition, there is a severe need to
316 complement predictions based on the genetics of microorganisms with phenotypic data.
317 Ultimately, understanding drivers of microbial community metabolism will lead to a better
318 ability to predict and engineer microbial communities for industrial or conservational
319 purposes.

320

321 **Methods**

322 *Data origin and Annotation of Ecosystems*

323 Metagenome-assembled genomes (MAGs) utilized in the current study comprised the set
324 published by Parks et al. (Parks et al. 2017). The Uncultured Bacterial and Archaeal (UBA)

325 MAGs were downloaded from the author's repository
326 (https://data.ace.uq.edu.au/public/misc_downloads/uba_genomes/). The accompanying data
327 from the UBA MAG set, including CheckM metrics of predicted genome completeness and
328 size, was obtained from the publication (Parks et al. 2017). Each study in the UBA set of
329 MAGs was manually sorted into a set of nine ecosystems and an unclassifiable category
330 called 'Other'.

331 *Metabolic overlap calculation*

332 All MAGs were subsequently annotated using a custom pipeline based on the SEED API
333 (Aziz et al. 2008; Overbeek et al. 2005). In brief, protein encoding genes (pegs) were called
334 from the assemblies using `svr_call_pegs`
335 (<http://servers.nmpdr.org/sapling/server.cgi?pod=ServerScripts>). Each of these proteins was
336 then assigned to a figfam with `svr_assign_using_figfams`. The association of a protein to a
337 biochemical reaction was then made with `svr_roles_to_reactions`. A custom script
338 (`rxn_expandinfo`) associated reactions with compounds from the reaction database which is
339 found on the ModelSEED git repository (<https://github.com/ModelSEED>). Finally, the
340 number of compounds shared between two organism's set of biochemical reactions is
341 calculated to create a pair-wise MO score, and a distance matrix was constructed to store this
342 information. This was made using the custom python scripts `rxn_to_connections` and
343 `lists_to_matrix`, respectively (<https://github.com/ericHester/metabolicOverlap>). The distance
344 matrix represents the MO of all organisms within a single community and the average MO of
345 all of these organisms is utilized in comparisons in this study.

346 In addition to an overall MO score for a community, the above approach was used to calculate
347 the MO of various sub-categories of metabolism for the respective community. In addition to
348 the above, an additional step was performed where pegs were assigned to their respective
349 SEED subsystems and filtered with a custom script utilizing `svr_roles_to_subsys`. With pegs
350 assigned to these metabolic categories, the above pipeline was used to identify reactions and
351 compounds shared between pairs of organisms, subsequently resulting in a distance matrix
352 similar to that above. In this case, the distance matrix stores the MO of the community
353 pertaining to a specific category of metabolism. Matrices and accompanying data were further
354 analyzed in R (R Core Team 2016).

355 *Relating phylogenetic distances of MAGs to their MO within communities*

356 In order to associate the phylogenetic distance of assembled genomes to their MO, the UBCG
357 pipeline was utilized (Na et al. 2018). This pipeline extracts 92 conserved phylogenetic
358 marker genes and builds multiple alignments for each gene. The resulting alignments are
359 concatenated and a maximum likelihood tree is inferred. This tree was imported into R
360 utilizing the *ape* package and distances were extracted from the tree object with the
361 *cophenetic* function (Paradis, Claude, and Strimmer 2004). The result is a distance matrix
362 containing phylogenetic distances between each pair of MAGs. Subsequently, this
363 phylogenetic distance matrix and the distance matrix storing MO scores were correlated using
364 the *mantel.test* function from the *ape* package. The Spearman's rank correlation coefficient
365 was calculated for each ecosystem subset.

366

367 Acknowledgements

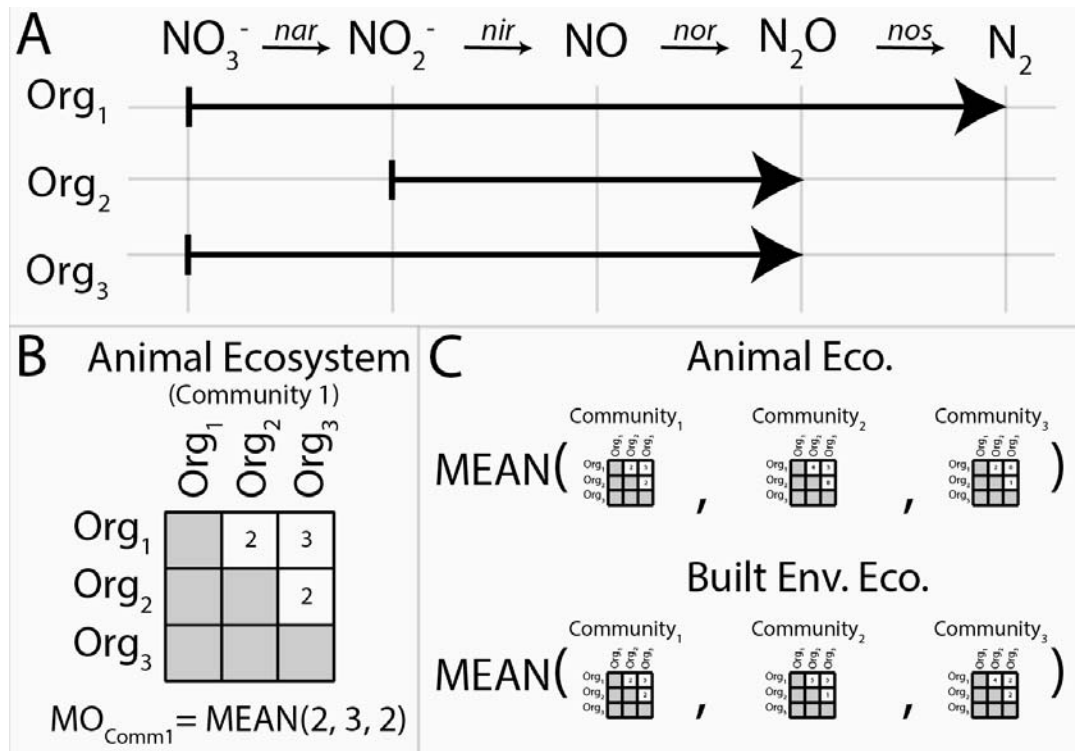
368 *We would like to acknowledge Michiel van den Heuvel for contributing to the metabolic*
369 *overlap pipeline. Funding was provided by the European Research Council (ERC AG*
370 *ecomom 339880) and the Netherlands Organisation for Scientific Research (NWO;*
371 *Gravitation Grant SIAM 024.002.002, VIDI grant 016.Vidi.189.050).*

372 **References**

- 373 Adams, Rachel I., Ashley C. Bateman, Holly M. Bik, and James F. Meadow. 2015.
374 "Microbiota of the Indoor Environment: A Meta-Analysis." *Microbiome* 3(1): 49.
- 375 Anantharaman, Karthik et al. 2016. "Thousands of Microbial Genomes Shed Light on
376 Interconnected Biogeochemical Processes in an Aquifer System." *Nature*
377 *Communications* 7: 13219.
- 378 Aylward, Frank O et al. 2015. "Microbial Community Transcriptional Networks Are
379 Conserved in Three Domains at Ocean Basin Scales." *Proceedings of the National*
380 *Academy of Sciences of the United States of America* 112(17): 5443–48.
- 381 Aziz, Ramy K et al. 2008. "The RAST Server: Rapid Annotations Using Subsystems
382 Technology." *BMC Genomics* 9(1): 75.
- 383 Berg, Gabriele, Martin Grube, Michael Schloter, and Kornelia Smalla. 2014. "The Plant
384 Microbiome and Its Importance for Plant and Human Health." *Frontiers in Microbiology*
385 5: 1.
- 386 Bowers, Robert M et al. 2017. "Minimum Information about a Single Amplified Genome
387 (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea."
388 *Nature Biotechnology* 35(8): 725–31.
- 389 Buckling, Angus, Rees Kassen, Graham Bell, and Paul B. Rainey. 2000. "Disturbance and
390 Diversity in Experimental Microcosms." *Nature* 408(6815): 961–64.
- 391 Carr, Rogan, and Elhanan Borenstein. 2012. "NetSeed: A Network-Based Reverse-Ecology
392 Tool for Calculating the Metabolic Interface of an Organism with Its Environment."
393 *Bioinformatics (Oxford, England)* 28(5): 734–35.
- 394 Connell, J.H. 1978. "Diversity in Tropical Rain Forests and Coral Reefs." *Science*.
- 395 David, Lawrence A. et al. 2014. "Diet Rapidly and Reproducibly Alters the Human Gut
396 Microbiome." *Nature* 505(7484): 559–63.
- 397 Delmont, Tom O. et al. 2018. "Nitrogen-Fixing Populations of Planctomycetes and
398 Proteobacteria Are Abundant in Surface Ocean Metagenomes." *Nature Microbiology*
399 3(7): 804–13.
- 400 Gause, G. F. 1934. *The Struggle for Existence*, by G. F. Gause. Baltimore,: The Williams &
401 Wilkins company,.
- 402 Ghatak, Sankha, Zachary A King, Anand Sastry, and Bernhard O Palsson. 2019. "The Y-Ome
403 Defines the 35% of *Escherichia Coli* Genes That Lack Experimental Evidence of
404 Function." *Nucleic Acids Research* 47(5): 2446–54.
- 405 Hansell, Dennis A., and Craig A. Carlson. 2002. *Biogeochemistry of Marine Dissolved*
406 *Organic Matter*. Academic Press.
- 407 Hester, E.R. et al. 2016. "Stable and Sporadic Symbiotic Communities of Coral and Algal
408 Holobionts." *ISME Journal* 10(5).
- 409 Hester, Eric R et al. 2018. "Linking Nitrogen Load to the Structure and Function of Wetland

- 410 Soil and Rhizosphere Microbial Communities.” *mSystems* 3(1): e00214-17.
- 411 Hutchinson, G. E. 1957. “Concluding Remarks.” *Cold Spring Harbor Symposia on*
412 *Quantitative Biology* 22(0): 415–27.
- 413 Kashyap, Purna C. et al. 2013. “Complex Interactions Among Diet, Gastrointestinal Transit,
414 and Gut Microbiota in Humanized Mice.” *Gastroenterology* 144(5): 967–77.
- 415 Kelly, Linda W et al. 2014. “Local Genomic Adaptation of Coral Reef-Associated
416 Microbiomes to Gradients of Natural Variability and Anthropogenic Stressors.”
417 *Proceedings of the National Academy of Sciences of the United States of America*
418 111(28): 10227–32.
- 419 Lee, Sonny T. M., Simon K. Davy, Sen-Lin Tang, and Paul S. Kench. 2016. “Mucus Sugar
420 Content Shapes the Bacterial Community Structure in Thermally Stressed Acropora
421 Muricata.” *Frontiers in Microbiology* 7: 371.
- 422 Levy, Roie et al. 2015. “NetCooperate: A Network-Based Tool for Inferring Host-Microbe
423 and Microbe-Microbe Cooperation.” *BMC Bioinformatics* 16(1): 164.
- 424 Lüke, Claudia et al. 2014. “Macroecology of Methane-Oxidizing Bacteria: The β -Diversity of
425 *PmoA* Genotypes in Tropical and Subtropical Rice Paddies.” *Environmental*
426 *Microbiology* 16(1): 72–83.
- 427 Martiny, Jennifer B. Hughes et al. 2006. “Microbial Biogeography: Putting Microorganisms
428 on the Map.” *Nature Reviews Microbiology* 4(2): 102–12.
- 429 Na, Seong-In et al. 2018. “UBCG: Up-to-Date Bacterial Core Gene Set and Pipeline for
430 Phylogenomic Tree Reconstruction.” *Journal of Microbiology* 56(4): 280–85.
- 431 Nelson, Craig E et al. 2013. “Coral and Macroalgal Exudates Vary in Neutral Sugar
432 Composition and Differentially Enrich Reef Bacterioplankton Lineages.” *The ISME*
433 *Journal* 7(5): 962–79.
- 434 Nemergut, Diana R et al. 2013. “Patterns and Processes of Microbial Community Assembly.”
435 *Microbiology and molecular biology reviews* □: *MMBR* 77(3): 342–56.
- 436 Overbeek, R. et al. 2005. “The Subsystems Approach to Genome Annotation and Its Use in
437 the Project to Annotate 1000 Genomes.” *Nucleic Acids Research* 33(17): 5691–5702.
- 438 Paradis, E., J. Claude, and K. Strimmer. 2004. “APE: Analyses of Phylogenetics and
439 Evolution in R Language.” *Bioinformatics* 20(2): 289–90.
- 440 Parks, Donovan H. et al. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes
441 Substantially Expands the Tree of Life.” *Nature Microbiology* 2(11): 1533–42.
- 442 R Core Team. 2016. “R: A Language and Environment for Statistical Computing.” *R*
443 *Foundation for Statistical Computing, Vienna, Austria*. <https://www.r-project.org/>.
- 444 Reese, Aspen T. et al. 2018. “Microbial Nitrogen Limitation in the Mammalian Large
445 Intestine.” *Nature Microbiology* 3(12): 1441–50.
- 446 Sasse, Joelle, Enrico Martinoia, and Trent Northen. 2018. “Feed Your Friends: Do Plant
447 Exudates Shape the Root Microbiome?” *Trends in Plant Science* 23(1): 25–41.

- 448 Schoener, T. W. 1974. "Resource Partitioning in Ecological Communities." *Science*
449 185(4145): 27–39.
- 450 Solden, Lindsey M. et al. 2018. "Interspecies Cross-Feeding Orchestrates Carbon Degradation
451 in the Rumen Ecosystem." *Nature Microbiology* 3(11): 1274–84.
- 452 Søndergaard, M, and D N Thomas. 2004. "Dissolved Organic Matter (DOM) in Aquatic
453 Ecosystems." *A study of European catchments and coastal waters, The DOMAINE*
454 *project*.
- 455 Tully, Benjamin J., Elaina D. Graham, and John F. Heidelberg. 2018. "The Reconstruction of
456 2,631 Draft Metagenome-Assembled Genomes from the Global Oceans." *Scientific Data*
457 5: 170203.
- 458 Whiteson, Katrine L. et al. 2014. "The Upper Respiratory Tract as a Microbial Source for
459 Pulmonary Infections in Cystic Fibrosis. Parallels from Island Biogeography." *American*
460 *Journal of Respiratory and Critical Care Medicine* 189(11): 1309–15.
- 461 Zengler, Karsten, and Livia S. Zaramela. 2018. "The Social Network of Microorganisms —
462 How Auxotrophies Shape Complex Communities." *Nature Reviews Microbiology* 16(6):
463 383–90.
- 464 Zhalnina, Kateryna et al. 2018. "Dynamic Root Exudate Chemistry and Microbial Substrate
465 Preferences Drive Patterns in Rhizosphere Microbial Community Assembly." *Nature*
466 *Microbiology* 3(4): 470–80.
- 467 Zimmerman, Amy E, Adam C Martiny, and Steven D Allison. 2013. "Microdiversity of
468 Extracellular Enzyme Genes among Sequenced Prokaryotic Genomes." *The ISME*
469 *Journal* 7(6): 1187–99.
- 470
- 471

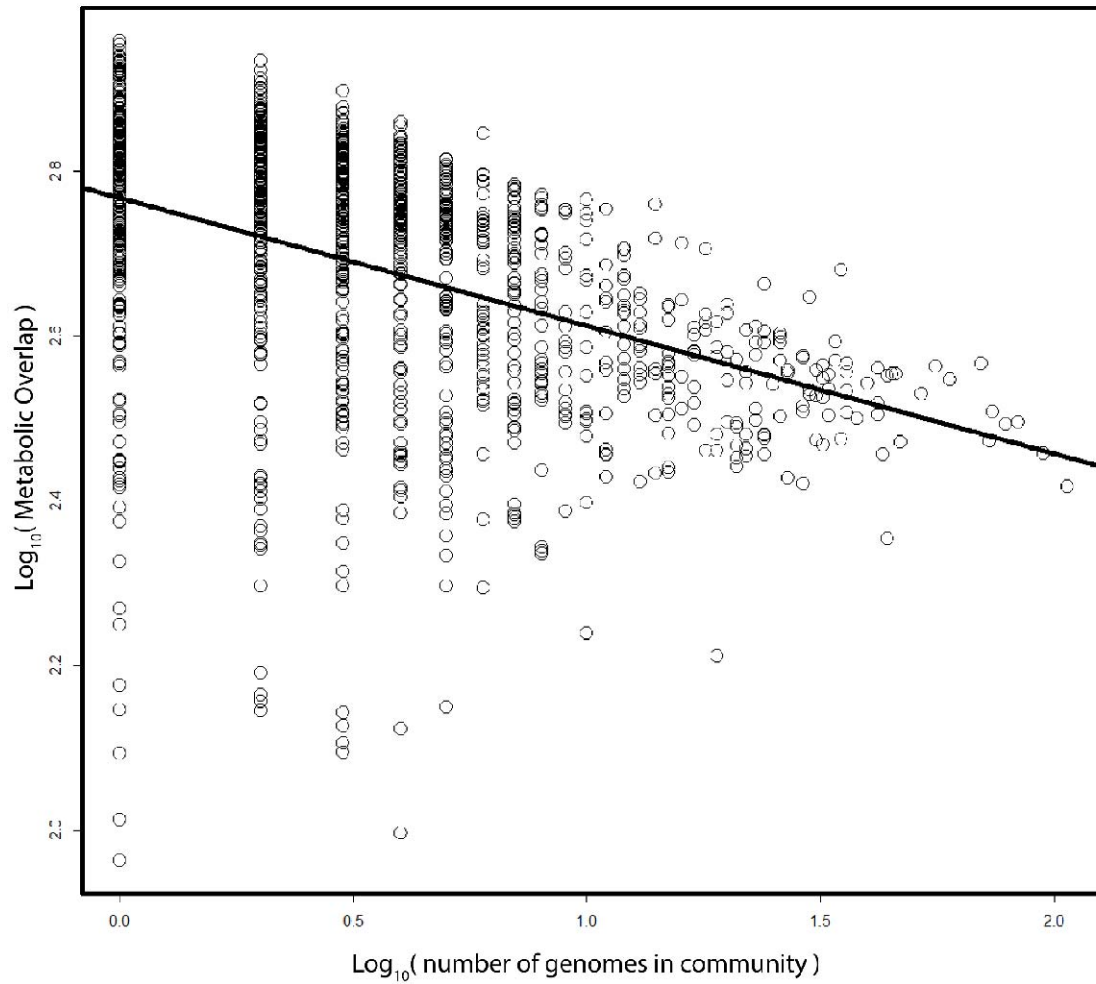


472

473 **Figure 1.** Metabolic overlap is a metric that compares the overlap in the metabolism of two organisms
 474 by calculating the number of reactants these species can utilize in common. This is determined by
 475 establishing their shared biochemical pathways (A). The number of substrates shared between a set of
 476 organisms is represented in a matrix (B), typically a symmetrical distance matrix. The average
 477 metabolic overlap of all communities from a given ecosystem are calculated and can be then compared
 478 to other ecosystems as seen in the current study (C).

479

Relationship of metabolic overlap and community richness

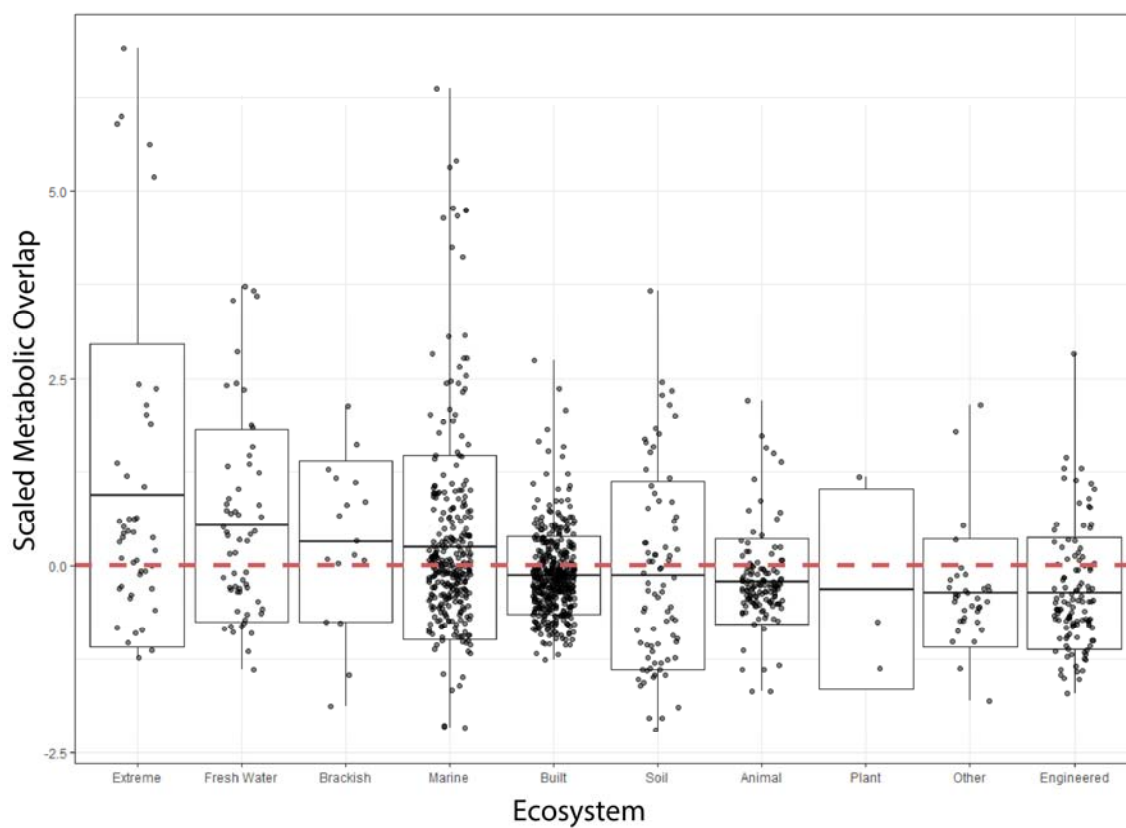


480

481

482 **Figure 2.** Relationship between metabolic overlap and the number of genomes in a community. Each
483 circle represents one of the 1248 studies. The x-axis depicts the total number of MAGs in a given
484 study, the y-axis the mean metabolic overlap of that study.

485

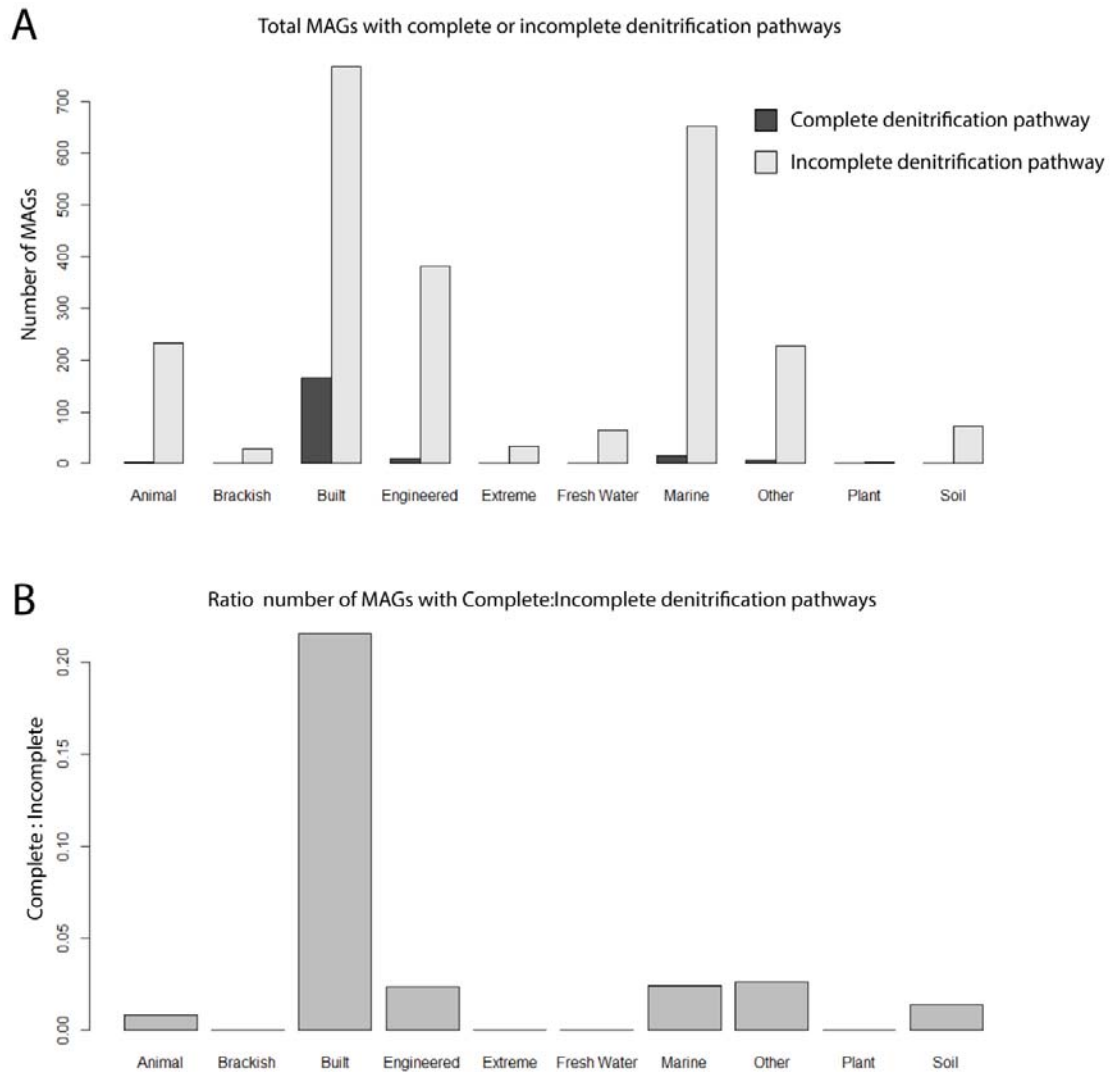


486

487 **Figure 3.** Metabolic overlap across all ecosystems. Boxplots are plotted with the black bar
488 representing the mean, the box is the 25% and 75% quartiles, and the whiskers are the extreme values.
489 A horizontal red dashed line was plotted to indicate 0, which corresponds to the average MO of all
490 ecosystems combined. Each point represents the mean metabolic overlap of all MAGs from a given
491 study.

492

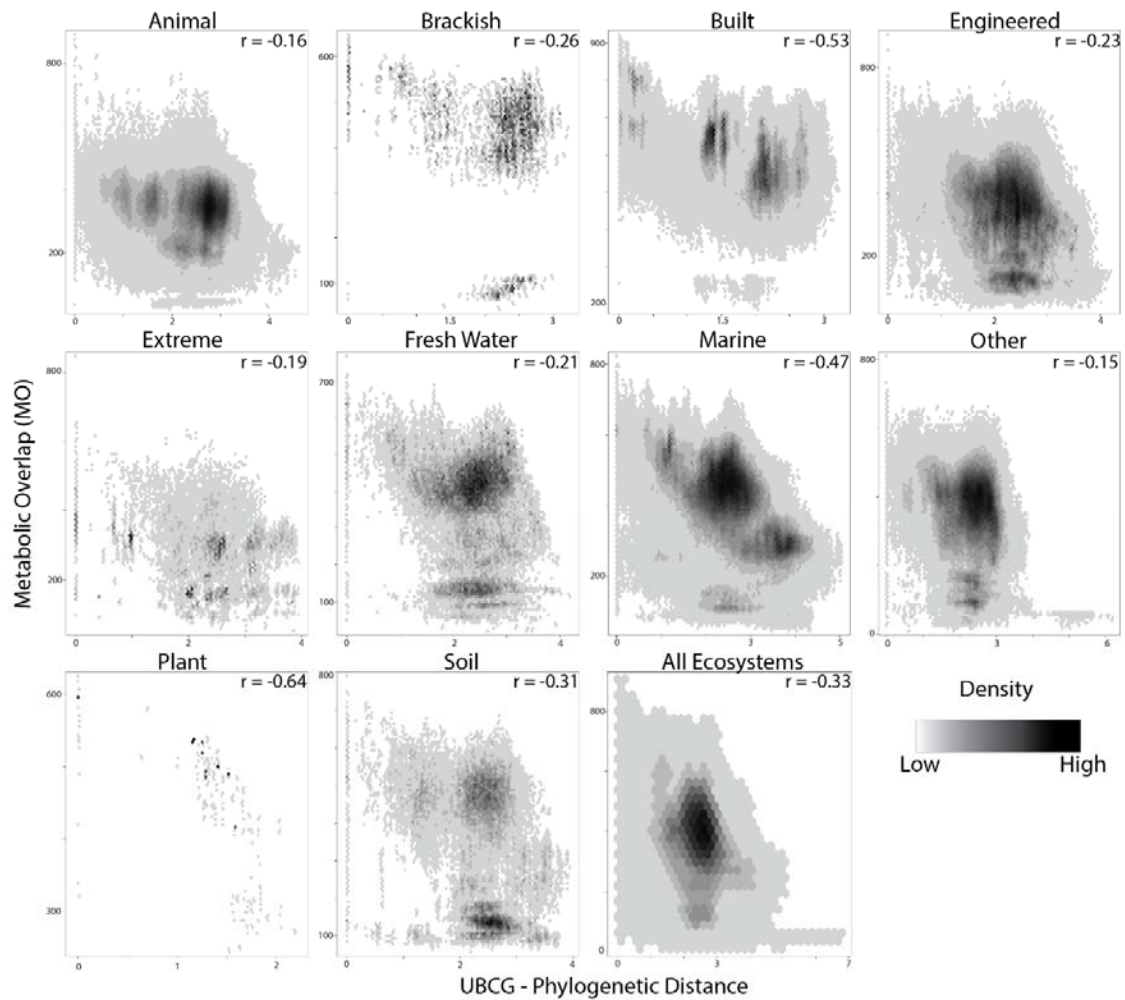
493



494

495 **Figure 4.** Proportion of complete to incomplete denitrification pathways across different ecosystems.
496 (A) Number of MAGs encoding all proteins to reduce NO_3^- to N_2 (complete denitrifiers) compared to
497 the number of MAGs with one or more of the respective genes missing. (B) Ratio of complete to
498 incomplete denitrification pathways.

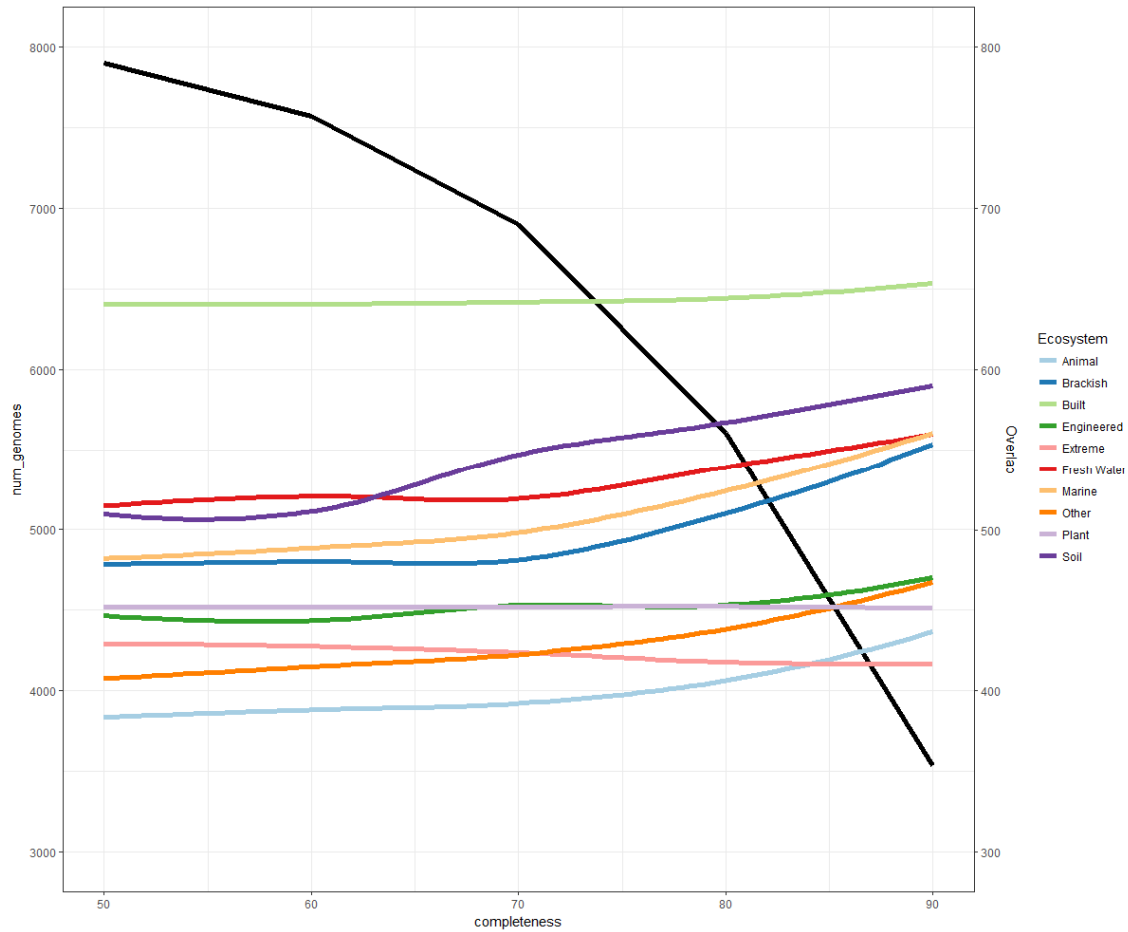
499



500

501 **Figure 5.** Relationship between metabolic overlap and phylogenetic distance of MAGs. Each point
502 represents a pairwise comparison between two MAGs. The density of points is represented by a black
503 and white gradient. The Spearman's correlation coefficient is indicated in the upper left-hand corner of
504 each plot.

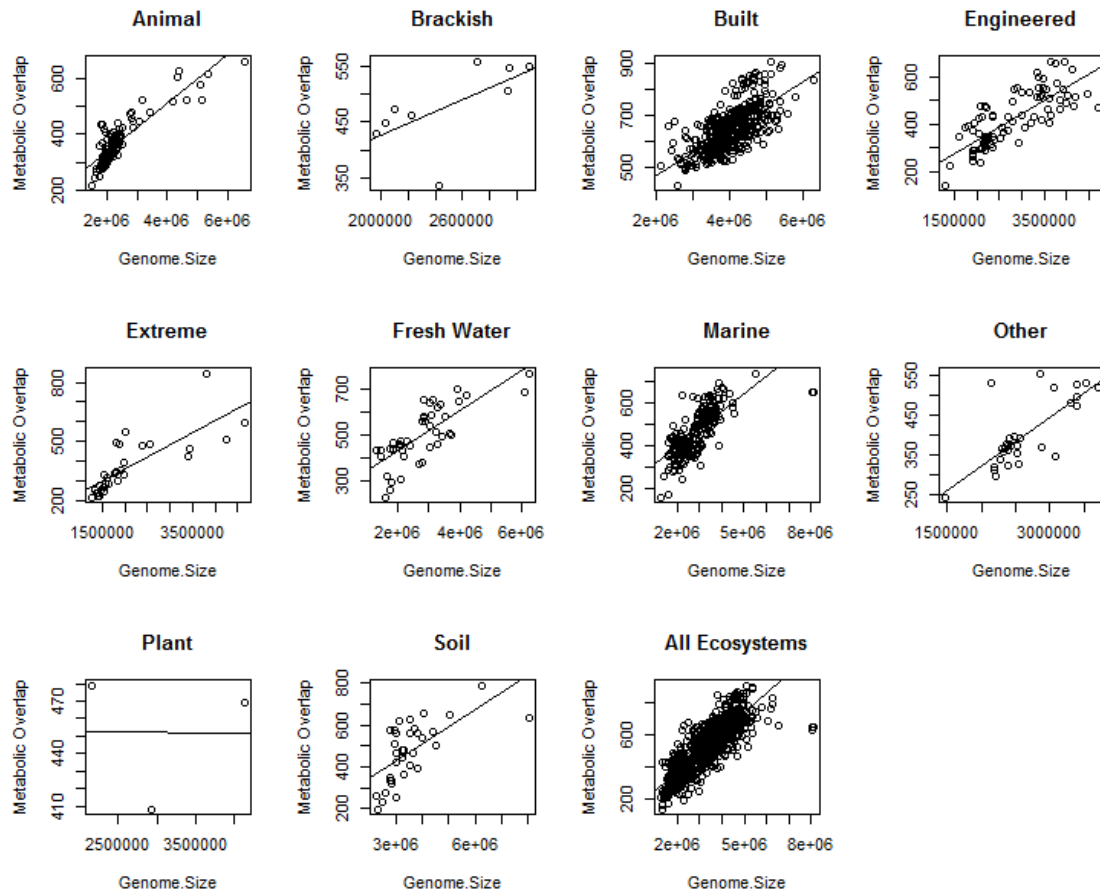
505



506

507 **Supplemental Figure 1.** Relationship between the genome completeness and the average metabolic
508 overlap observed (colored lines, right axis). The number of MAGs retained at the different
509 completeness cutoffs is indicated by the black line (left axis).

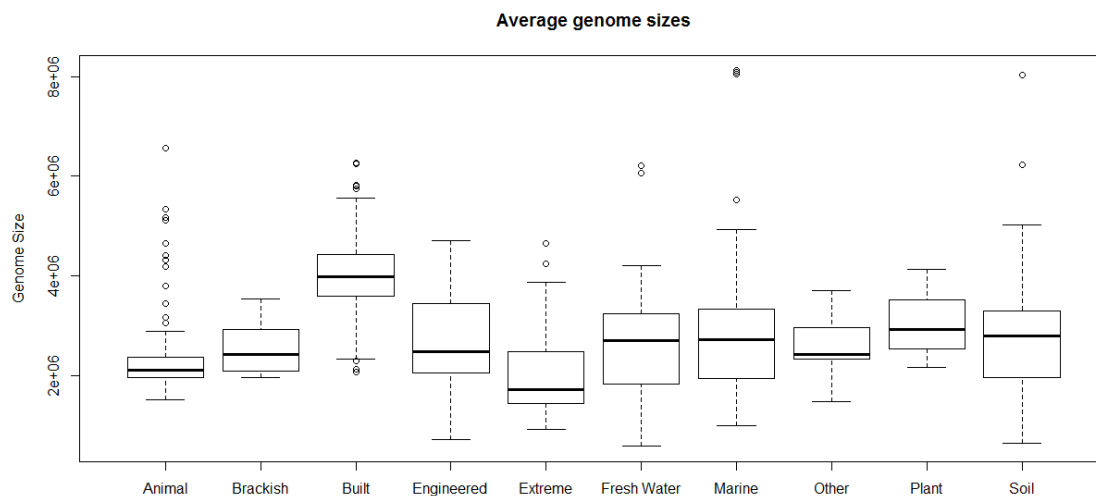
510



511

512 **Supplemental Figure 2.** Relationship between metabolic overlap and genome size. Each circle
513 represents one study. The y-axis indicates the average metabolic overlap of all MAGs in one study,
514 and on the x-axis the average genome size for all MAGs in this study.

515



516

517 **Supplemental Figure 3. Average genome sizes across ecosystems.** The black bar of the boxplot
518 indicates the median, the box edge represents the upper and lower quartiles, whiskers denote extreme
519 values, and individual points are outliers.

520

	Number of studies	Number of metagenomes
Animal	130	1823
Brackish	17	66
Built	446	1275
Engineered	122	1374
Extreme	44	156
Fresh Water	59	231
Marine	311	1811
Other	35	928
Plant	3	16
Soil	81	223
Total	1248	7903

521 **Table 1. Number of studies and metagenomes within each ecosystem.**

522

Ecosystem	Predicted Genome Size (bp)	
	Mean	s.d.
Extreme	2051727	966129
Animal	2340102	784483
Brackish	2530756	516589
Other	2608893	539528
Fresh Water	2631460	1103610
Engineered	2651621	920438
Marine	2709229	987460
Soil	2760719	1226428
Plant	3074667	990199
Built	4000017	658349

523 **Table 2. Mean genome size in each ecosystem.**

524

Ecosystem	t	p.value
Animal	-4.25	< 0.001
Brackish	1.23	0.23
Built	-5.32	< 0.001
Engineered	-5.31	< 0.001
Extreme	3.08	0.003
Fresh Water	3.19	0.002
Marine	3.5	< 0.001
Other	-3	0.005
Plant	-0.41	0.72
Soil	-0.93	0.35

525 **Table 3. Metabolic overlap statistics in each ecosystem.**

526

Statistic (t)	Animal	Brackish	Built	Engineered	Extreme	Fresh Water	Marine	Other	Plant	Soil
Amino Acid Metabolism	-18.808	1.0396	-7.0897	-8.0706	4.3911	3.0921	7.8829	-2.8291	-0.1535	0.0306
Aromatic Metabolism	-14.002	-0.2595	-4.3238	-2.9174	-1.4082	1.0256	7.0192	-0.793	-2.968	1.4364
Carbohydrate Metabolism	-10.015	0.5472	-4.4021	-7.0678	3.4987	1.9514	6.0391	-4.0708	-0.6076	-0.1959
Cofactor Metabolism	-23.996	1.7968	-8.3088	-9.3914	3.4899	3.6088	8.5258	-3.1922	-0.1503	0.5693
Fatty Acid Metabolism	-6.2286	3.0776	-7.0429	-5.054	2.5558	2.4734	6.5712	-1.5586	-0.5669	-2.5669
Nitrogen Metabolism	-15.495	1.7972	6.9147	-4.0639	2.3216	2.2063	-0.9558	-0.4684	-0.4381	-1.4858
Nucleotide Metabolism	4.6226	-0.4585	-16.574	0.0291	3.241	1.4079	4.8181	-1.164	-0.6831	-0.8228
Nucleotide Sugar Metabolism	3.073	1.4103	-12.295	2.9834	3.4546	3.6569	-1.2075	0.6512	-0.0622	2.3715
Phosphorous Metabolism	-3.8878	0.8176	-8.3411	-1.0578	0.7587	2.2691	4.6001	-1.064	-0.5048	-0.8658
Protein Metabolism	3.7253	1.8878	-33.882	2.2077	5.5824	3.9273	4.4379	-0.6715	0.2008	2.5578
Respiration	-15.021	3.4966	-3.8864	-5.1202	2.8844	2.5968	9.519	-1.5019	-1.0597	-1.8178
Sulfur Metabolism	-24.636	-0.3785	14.3902	-10.8816	-0.6835	0.5794	1.7196	-7.2878	-1.7582	-2.4058
Secondary Metabolism	-0.41	1.005	-15.753	-1.161	3.738	2.241	4.644	0.609	NA	2.243

527 **Table 4. Metabolic overlap in different categories of metabolism**