Targeted metagenomics using probe capture detect a larger diversity of nitrogen and methane cycling genes in complex microbial communities than traditional metagenomics

H.M.P. Siljanen[1,2], L. Manoharan[2a], A.S. Hilts[2], A. Bagnoud[2b], R.J.E. Alves[2c], C.M. Jones[3], M. Kerou[2], F. L. Sousa[2], S. Hallin[3], C. Biasi[1,4], C. Schleper[2]

1. Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland
2. Department of Functional and Evolutionary Ecology, University of Vienna, Vienna, Austria
3. Department of Forest Mycology and Plant Pathology, Sweden University of Agricultural Sciences, Uppsala, Sweden.
4. Department of Ecology, University of Innsbruck, Innsbruck, 6020, Austria

Present address:

a. National Bioinformatics Infrastructure Sweden (NBIS), SciLifeLab, Department of Laboratory Medicine, Lund University, Lund, Sweden.
b. Membratec S, Ecoprac de Daval C 1, CH-3960 Sierre, Switzerland.
c. Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Corresponding authors:

Henri Siljanen, henri.siljanen@uef.fi

Christa Schleper, christa.schleper@univie.ac.at

Keywords: metagenomics, nitrogen cycling, methane cycling, probe hybridization targeted metagenomics, PCR amplicon sequencing, shotgun metagenomics

Manuscript contribution to the field

Metagenomic sequencing often yields limited numbers of sequences of rare microbial taxa or functional genes, preventing in-depth analyses of specific populations and functional groups. Amplicon-based approaches enable the higher diversity coverage of target populations, but the

drawback is the difficulty in designing unbiased primers that cover the highest intra-group diversity. Targeted metagenomics overcomes these challenges and results in similar community structure as traditional amplicon sequencing, while expanding the sequence space in a less biased metagenomic-based approach. Therefore, targeted metagenomics is an invaluable tool for studying the diversity of specific populations within complex natural microbiomes. Here, we present and evaluate a probe library designed for targeted metagenomics of nitrogen and methane cycling genes in complex communities.

Abstract

Microorganisms are key players in the global cycling of nitrogen (N) and carbon (C), controlling their availability and fluxes, including the emissions of the powerful greenhouse gases nitrous oxide (N_2O) and methane (CH_4). Characterizing the microbial functional guilds driving these processes is crucial for understanding ecosystem functioning and predicting their responses to environmental changes. Standard sequence-based characterization methods often reveal only a limited fraction of their diversity in nature because of their low relative abundance, the insufficient sequencing depth of traditional metagenomes of complex communities, and limitations in coverage of PCR-based assays. Here, we developed and tested a targeted metagenomics approach based on probe capture and hybridization to simultaneously characterize the diversity of multiple key metabolic genes involved in inorganic N and CH_4 cycling. We designed comprehensive probe libraries for each of the 14 target marker genes comprising 264,000 unique probes. These probes were used to selectively enrich the target genes in shotgun metagenomic libraries. In validation experiments with the mock communities of known microorganisms, targeted metagenomics yielded gene profiles similar to those of the original communities. Only GC content had a small effect on probe efficiency, as low GC targets were less efficiently detected than those with high GC, within the mock communities. Furthermore, the relative abundances of the marker genes obtained using targeted or traditional shotgun metagenomics from agricultural and wetland soils were significantly correlated, indicating that the targeted approach did not introduce significant quantitative bias. In addition, using archaeal *amoA* genes as a case-study, targeted metagenomics identified substantially higher taxonomic diversity and a larger number of sequence reads per sample, yielding diversity estimates 28 or 1.24 times higher than shotgun metagenomics or

amplicon sequencing, respectively. Notably, shotgun metagenomics detected only three out of the 84 *amoA* gene phylotypes detected using targeted metagenomics. Our results show that targeted metagenomics complements current approaches to characterize key microbial populations and functional guilds in biogeochemical cycles in different ecosystems, enabling more detailed, simultaneous characterization of multiple functional genes.

Introduction

The global nitrogen (N) and carbon (C) cycles are essential processes of the Earth's biosphere and crucial for ecosystem functioning (Zaehle 2013, Gruber and Galloway, 2008; Archer 2010; Falkowski et al., 2000; Steffen et al. 2015). All major N transformation processes (i.e., nitrogen fixation, nitrification, denitrification, dissimilatory nitrate reduction to ammonium (DNRA), and anaerobic ammonium oxidation, or anammox), are performed exclusively by the functional guilds of bacteria, archaea, and eukaryotes (Galloway et al., 2004; Stein and Klotz, 2016; Kuypers et al., 2018; Offre et al. 2013). Their activities regulate N availability for primary producers and microorganisms across ecosystems and control the production and consumption of the potent greenhouse gas nitrous oxide (N_2O) and other gaseous N compounds through processes such as nitrification, denitrification, and non-denitrifier N_2O reduction (Barnard, et al., 2005, Thompson et al., 2012; Jones et al., 2014; Prosser et al. 2020). Methane (CH_4) is another major product of the microbial trophic chain underlying C cycling, and, like the most inorganic N compounds, its production and consumption are also regulated by highly specific groups of microorganisms (Bodelier and Laanbroek 2004, Murrel and Jetten 2009, Conrad 2009). Importantly, CH_4 is the second most potent greenhouse gas after carbon dioxide (CO_2) and, together with N_2O, contributes at least 25% of the total global warming caused by greenhouse gases (Myhre et al., 2013; Friedlingstein et al., 2020; Lan et al., 2023).

Despite their important ecological role, microorganisms participating in inorganic N and CH_4 transformations typically constitute a small fraction of microbial communities in most soil, sediment, and aquatic ecosystems (Håvelsrud, et al., 2011; Palmer et al., 2012; Aanderud et al., 2015, Nelson et al, 2016, Ouygang and Norton 2020). Community profiling based on 16S rRNA

genes only allows for very limited information about microbial functions (Alteio et al. 2021). Even in the rare cases when functions can be inferred from organism phylogeny, the striking genetic and functional diversity within functional guilds remains concealed, such as that among their key metabolic enzymes (e.g., ammonia monooxygenase (*amoA*), Alves et al. 2018; methyl co-enzyme M reductase (*mcrA*), Borrel, 2019). Additionally, taxonomy-based approaches are inadequate in elucidating processes that have a broader taxonomic distribution than originally thought (e.g. Kuypers et al. 2018, Borrel 2019, Li et al. 2023; Saghaï et al., 2024). Shotgun metagenomics is the typical method of choice for obtaining a relatively unbiased picture of the natural microbiome, provided that issues with sample preparation, DNA extraction, library preparation, and sequencing method can be ruled out (Pan et al., 2010, Nnadozie, Lin and Govinden, 2015; Sinha et al., 2017; Sui et al., 2020). Nevertheless, in complex and diverse microbial communities such as those in soils and sediments, even deeply-sequenced metagenomes do not uncover the full diversity of microbial functional guilds and key functional genes owing to their typically low relative abundances within the broader community (Palmer et al., 2012, Orellana et al. 2019, Ouygang and Norton 2020, Zhao et al. 2023). As a low-cost alternative, the characterization of specific functional groups has long relied on the PCR amplification of genes that encode key metabolic enzymes, as for example, of organisms involved in various N and CH_4 cycling pathways (Knief 2015; Brauer et al. 2021; Alves et al. 2018; Romdhane et al. 2021, Clark et al. 2022). However, the gene diversity captured by gene-specific PCR assays is limited by the primers used, which introduce biases that make comparisons among different genes or samples difficult (Hallin and Lindgren, 1999; Throbäck al. 2004; Kolb et al., 2003; Siljanen et al., 2011). Furthermore, in complex microbial communities, target genes of interest are encoded by a large diversity of organisms, and usually include a large fraction of novel gene sequence variants from uncharacterized organisms (Jones et al. 2014; Saghaï et al 2023). Thus the geometric expansion of genomic data has made it increasingly obvious that designing a single PCR primer pair to target the inherent diversity of metabolic genes is highly problematic (Bonilla-Rosso et al. 2016). Therefore, linking ecosystem functions to microbial communities is often only possible through the complex and resource-intensive combinations of meta-omics and isotope labelling approaches (Tveit et al. 2015, Orellana et al. 2019).

126 Targeted high-throughput sequencing approaches, known as probe capture, also called
127 hybridization capture, hybridization-based target enrichment, or captured metagenomics, have
128 been used to study complex eukaryotic samples such as human exons, ancient human genomes,
129 plant transcriptomes, and cancer marker single nucleotide polymorphisms (Skoglund et al. 2014,
130 Ichida et al. 2019, Futema et al. 2012, Gasc et al., 2016, Kamil et al. 2021, Bewicke-Copley et al.
131 2019). A few studies have also used this approach to facilitate the in-depth study of microbial
132 communities using either 16S rRNA genes or other microbial functional genes (Denonfoux et al.,
133 2013; Manoharan et al. 2015, Noyes et al., 2017, Beaudry et al., 2021). The probe capture approach
134 relies on targeting specific short sequences within a broader genomic pool (e.g., full genomes,
135 genome or gene fragments) with biotin-labelled probes. The probes that hybridize with their target
136 regions are then selectively captured from the full genomic library using streptavidin-labelled
137 magnetic beads (Liu et al., 2016). Different from sequence capture, where multiple probes are
138 designed to cover a whole exon, captured metagenomics relies on the large databases of multiple
139 target genes clustered at predefined similarity cut-offs, which are used to design a large number of
140 generic probes that are able to capture an extended sequence space for each of these sequence
141 clusters (Kushwaha et al., 2015, Manoharan et al., 2015).
142
143 To improve detection and the characterization of important, but low abundant microbial guilds
144 involved in N and CH_4 cycling in natural communities, we developed and evaluated a probe-based
145 targeted metagenomics approach to key genes involved in these processes. In addition to the
146 enhanced resolution and coverage of genetic sequence space, this approach allows the parallel
147 analysis of distinct steps within N- and CH_4- cycling pathways, thus combining the advantages of
148 both shotgun metagenomics and amplicon-based approaches. Here, we present a new probe set for
149 targeting 14 distinct functional genes involved in inorganic N cycling (i.e., N fixation, nitrification,
150 denitrification, anammox, and DNRA), and three genes involved in CH_4 production and
151 consumption. We evaluated our approach using mock communities comprising of DNA from
152 microorganisms involved in inorganic N cycling, or CH_4 production or consumption with varying
153 gene GC% content, which is known to affect probe hybridization efficiency. As a proof-of-concept
154 for complex communities, we performed shotgun metagenomics and targeted metagenomics on
155 agricultural and wetland soil samples and compared the diversity of two functional markers, the

156 archaeal *amoA* genes (marker for ammonia oxidizers) and *nosZ* genes (marker for nitrous oxide
157 reduction).
158
159 **Materials and methods**
160
161 **Construction of the target gene databases (TDBs)**
162 Target gene databases (TDBs) were constructed, containing all identifiable variants for the
163 following key genes: the nitrogenase iron subunit (*nifH*), bacterial and archaeal ammonia
164 monooxygenase subunit A (*amoA)*, nitrite oxidoreductase beta subunit (*nxrB)*, hydrazine
165 oxidoreductase A (*hzoA)*, formate dependent nitrite reductase (*nrfA*), periplasmic nitrate reductase
166 alpha subunit (*napA*), respiratory nitrate reductase alpha subunit (*narG*), copper-containing nitrite
167 reductase (*nirK*), cytochrome cd_1 nitrite reductase (*nirS*), nitric oxide reductase subunit B (*norB*),
168 nitrous oxide reductase (*nosZ*), particulate methane monooxygenase subunit A (*pmoA*), soluble
169 methane monooxygenase component A alpha chain (*mmoX*), and methyl-coenzyme M reductase
170 I subunit alpha (*mcrA*).
171
172 Hidden Markov Model (HMM) models were generated in order to identify all variants of the target
173 genes from public databases (Fig. 1). These models were built based on reference sequence
174 alignments from curated databases already available for selected genes, such as *amoA* (Alves et
175 al., 2018), *pmoA* (Knief et al., 2015), *nosZ* (Jones et al., 2013), *nirK, nirS, nor* (Graf et al., 2014);
176 and the Fungene repository (Fish et al., 2013). For target genes where alignments were not
177 available, reference alignments were generated from gene sequence data publicly available on the
178 NCBI WGS-database, from the full length open-reading frame of each subunit, to cover all known
179 diversity in each gene. Structure-based searches of the Genbank nt- and envnt- database were
180 subsequently performed with nhmmer using the generated HMMs for every target gene (Wheeler
181 & Eddy, 2013) on a local supercomputer cluster (Centre of scientific computing CSC, Espoo,
182 Finland) (July 2017). The HMM models are available through Zenodo:
183 https://doi.org/10.5281/zenodo.15752134 (In script: probe-capture/2-
184 selected_outputs/hmmer_profiles.zip). This sequence search selection process generated ~600,000
185 unique sequences across all gene families. The obtained database was clustered to 100% identity
186 in order to remove duplicates sequences in database with cd-hit (Fu et al., 2012) and manually

187 inspected to exclude 16S rRNA genes. The final output comprised the target gene database (TDB;
188 available through Zenodo: https://doi.org/10.5281/zenodo.15752134; in script: probe-capture/2-
189 selected_outputs/all_genes_compiled_ncbi_nt_envnt_fungene_split_file_0.fasta.gz -file (four
190 files ..file_0-3.fasta.gz). The workflow can be seen in Fig 1. Scripts for Target gene database search
191 is available in Zenodo: https://doi.org/10.5281/zenodo.15752134; /probe-capture/tree/main/3-
192 other_scripts/Henri_scripts/Script_probe_capture.sh -folder.
193

194 **Generation of probes for target genes**

195 The MetCap bioinformatics pipeline was used as a protocol to produce probes from highly
196 complex datasets for targeted metagenomics (Kushwaha et al., 2015). Default parameters were
197 used for designing up to six unique 50mer probes for each sequence cluster in the TDBs (clustered
198 with an 80% identity clustering threshold), with melting temperature of 47 °C, resulting in a set of
199 263,111 unique gene-specific probes (Fig 1, and in Zenodo:
200 https://doi.org/10.5281/zenodo.15752134; probe-capture/2-
201 selected_outputs/Final_N_probes.list1.fasta -file. The probes were synthesized by NimbleGen
202 SeqCap EZ (Roche NimbleGen, Inc., Madison, USA) as Custom design (since the MetCap
203 pipeline was used for probe design, probe order was done with Roche HyperDesign tool:
204 www.hyperdesign.com), with biotin labelling to enable retrieval of the hybridized targets using
205 streptavidin coated magnetic beads (Kushwaha et al., 2015; Manoharan et al., 2015).
206

207 **Extraction of DNA from cultures and generation of mock community DNA samples**

208 DNA samples from mock communities encoding the genes of interest were generated by mixing
209 genomic DNA from different organisms in variable proportions, in order to simulate variation in
210 overall GC% content. The pool of DNA in the mock communities included DNA from:
211 *Nitrosospira multiformis*, *Nitrososphaera viennensis*, *Nitrospira defluvii*, Ca. *Kuenenia*
212 *stuttgartiensis* PCR fragment of the full-lenght *hzoA* gene, *Pseudomonas aeruginosa*, *Escherichia*
213 *coli*, *Shigella sonnei*, *Cupriavidus metallidurans*, *Cupriavidus necator*, *Dyadobacter fermentans*,
214 *Pseudomonas stutzeri*, *Rhodobacter sphaeroides*, *Salinibacter ruber*, *Sulfurimonas denitrificans*,
215 *Methylosinus trichosporium* Ob3b, *Methylocella tundrae*, *Methylomicrobium buryatense*,
216 *Methanoregula boonei*, and *Methanolacinia petrolearia*. Strain information, including genome or
217 fragment size and the number of target genes per organism are shown in Supplementary Table S1.

218 DNA from cultured organisms was extracted from 1% SDS-treated cell pellets in CTAB buffer,
219 followed by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation, as described
220 before (Jones et al., 2013). The relative abundances of the microorganisms in the mock
221 communities were multiplied with the median genome GC% contents to generate a weighted GC%
222 content of the pool of samples with the following values: 47, 50, 53, 57, 60, and 63 (Supplementary
223 Table S2). Pooling ratios of microorganisms' DNA were calculated based on the expected number
224 of functional genes in the extracted DNA, the DNA concentration (determined with a Qubit HS
225 dsDNA kit (Thermo)) and genome size. Relative gene abundances in mock communities were
226 calculated and compared to the sequences generated using targeted metagenomics. The relative
227 abundance of each gene was calculated against the sum of abundance values of all 14 genes in the
228 targeted metagenomes of the mock community samples (relative abundance of a specific gene =
229 100 x (gene reads / sum of reads of 14 genes)). The captured reads for mock community targeted
230 metagenomics are shown in Table S4.
231
232 **Annotation of coding sequences from the mock community dataset**
233 Predicted proteomes from the mock communities were retrieved from NCBI to count the number
234 of each functional genes per genome. For *Nitrospira defluvii*; NC_014355.1, genes were predicted
235 from the genome nucleotide sequence using Prodigal (v2.6.3) (Hyatt et al., 2010). Proteins were
236 annotated using KOFAM (HMM database of KEGG Orthologs (KO)). Each protein was assigned
237 the top hitting KO. Where available, cut-offs provided with KOFAMscan (Aramaki et al., 2020)
238 were used to filter spurious results.
239
240 **Extraction of DNA and determination of chemical parameters from environmental samples**
241 Samples from two different environments, an agricultural soil in Hungary (n = 3), and a wetland
242 in Bellefontaine, France (n = 3) (Table S3), were collected in order to assess the effectiveness of
243 the probe set in enriching functional genes in different sample contexts. Extraction of DNA from
244 the environmental samples was performed as previously described by Siljanen et al., (2019).
245 Briefly, samples were homogenized by bead-beating 0.5 g of soil at a speed of 5.5 m/s, for 30 sec
246 with phenol:chloroform:isoamyl alcohol extraction in CTAB buffer, followed by ethanol
247 precipitation. The quality of DNA extracts was assessed using a NanoDrop ND-1000 (Thermo),
248 and DNA concentration was measured with a Qubit HS dsDNA kit (Thermo). The soil chemical

249 parameters soil C/N ratio, pH, Fe II/III-, ammonium- and nitrate content were determined as
250 described earlier (Siljanen et al., 2019; Bagnoud et al, 2020).
251
252 **Targeted metagenomic library preparation, target enrichment of libraries and sequencing**
253 **with Illumina Miseq**
254 To prepare DNA for targeted metagenomics with probe capture, DNA was first fragmented and
255 indexed as follows. For each sample, sequencing indexes and sequencing adapters were provided
256 as commercial service by the Centre for Genomic Research (CGR) -laboratories, at the University
257 of Liverpool, Liverpool, UK. Libraries were produced with KAPA HyperPlus Library Preparation
258 (Roche) kit to produce insert sizes of 630 bp according to the manufacturer's instructions. The
259 protocol is shown in Fig. 1. and described here briefly: 150ng of environmental DNA was
260 fragmented with KAPA fragmentation buffer and KAPA Frag Enzyme at +37°C for 20min.
261 Fragmented DNA was ligated to SeqCap Adapters A and B, +20°C for 15 minutes. Fragments
262 were purified 0.65x volumes of AMPure XP Reagent for 5 min at room temperature in magnetic
263 particle collector (MPC). Then the clear supernatant was removed and the library was washed
264 twice with 200 µl of 80% ethanol. The libraries were removed from MPC by eluting them to 53
265 µl of elution buffer (10 mM Tris-HCl, pH 8.0) by incubated two minutes. Each library was
266 amplified for 7 cycles, with KAPA HyperPlus Library Preparation kit in order to add the indexes
267 for each sample (cycling conditions: +98°C, 15 sec; 60° for 30 sec; +72°C for 30 sec). Amplified
268 libraries were purified with AMPure XP beads as above.
269 The hybridization reactions to the NimbleGen SeqCap EZ Developer Probe set were performed
270 according to the manufacturer's instructions. The hybridization was done in non-binding DNA
271 background 25 µg, on top of 1 µg of each library equally pooled Multiplexed DNA Sample Library
272 were added. The Multiplexed Hybridization Enhancing Oligo Pool (2,000 pmol in 2µl volume),
273 the SeqCap HE Universal Oligo (1,000 pmol in 1µl volume) and (1,000 pmol in 1µl volume)
274 SeqCap HE Index Oligo Pool) was mixed with 1 µg of equally pooled Multiplexed DNA Sample
275 Library. Two volumes of AMPure XP Reagent was added on top of the above mixture and
276 thoroughly mixed. Samples were incubated for 10 minutes to allow the sample library to bind to
277 the beads. Samples were placed on the MPC to capture the beads and the solution allowed to clear.
278 Once clear, the supernatant was discarded carefully, to not disturb the beads. Afterwards 190 µl of
279 80% ethanol was added to the samples containing the bead-bound DNA samples. The samples

280 were left on the MPC during this step. Samples were incubated at room temperature for >30
281 seconds. The 80% ethanol was carefully removed and discarded, without disturbing the beads.
282 Beads were dried at room temperature with the tube lid open for 5 minutes (or until dry).
283 A master mix of the following reagents was prepared, scaling up to reflect the number of captures:
284 7.5 µl of the 2x Hybridization Buffer, 3 µl of Hybridization Component A. Then 10.5 µL of the
285 Hybridization Buffer/Hybridization Component A mix from the previous step was added to the
286 bead-bound DNA sample. Samples were removed from the MPC and mixed thoroughly. It was
287 important that enough mixing was performed at this step to yield a homogeneous mixture. It was
288 left to sit at room temperature for 2 minutes. Sampled were then placed on the MPC. After liquid
289 cleared, 10.5 µL of supernatant (entire volume) was removed and placed in a new tube containing
290 4.5 µl of the SeqCap EZ Developer Probe pool. This was mixed thoroughly. Hybridization
291 incubation was performed in a thermocycler using the following program with heated lid set to 10
292 °C above the block temperature: incubation was 95°C for 5min, and 47°C for 72 hours (SeqCap
293 EZ HyperCap Workflow protocol, Manoharan et al., 2015).
294 To purify the hybridized libraries the SeqCap Pure Capture Bead Kit was used. To prepare the
295 beads for binding the samples, 50 µl of beads (streptavidin coated magnetic beads) was added to
296 the 1.5 ml tube at room temperature for capturing reaction. Tubes were placed on a MPC, the
297 supernatant removed and beads were washed two times with 100 µl of 1x Bead Wash Buffer.
298 Afterwards it was removed from the MPC and mixed by pipetting up and down. Beads were bound
299 on the magnetic particle collector. Once the liquid was clear, the supernatant was removed. In the
300 second wash, the beads were aliquoted with 50 µl of 1X Bead Washing Buffer. Beads were bound
301 and the supernatant again removed. One hybridization sample (15 µl volume) was added to the
302 washed SeqCap Pure Capture Beads. Samples were bound to the beads by placing the samples in
303 a thermocycler at +47°C for 15 minutes (heated lid at +57°C). Then 100 µl of 1x Wash Buffer was
304 added on top of the Capture Beads and bead-bound DNA and mixed thoroughly. The tube was
305 placed on the MPC to capture the beads. The solution was allowed to clear, and the supernatant
306 removed, while being careful not to disturb the beads. Afterwards, 200 µl of 1x Stringent Wash
307 Buffer was added to each capture reaction. Those were mixed by pipetting up and down, when
308 tubes were not in a MPC. Samples were placed in the thermocycler to +47°C for 5 minutes. The
309 wash was repeated with 1x Stringent Wash Buffer. Next, 200 µl of 1x Wash Buffer I was added,

310 vortexed for 10 second, and incubated 1 minute at room temperature. Once the solution was clear,
311 it was removed, again being careful not to disturb the beads. The washing with 1X Wash buffer II
312 was repeated, and then 1x with Washing Buffer III. Samples were removed from the MPC and 53
313 µl of PCR-grade water was added to the bead-bound sample. Sampled were then incubated at room
314 temperature for 2 minutes and placed in the MPC. Then 50 µl of clear solution was collected in a
315 new tube.
316
317 Sequencing for probe hybridized and washed DNA was performed with Illumina MiSeq PE300
318 chemistry in the Centre for Genomic Research (CGR), University of Liverpool, Liverpool, UK,
319 resulting in 198,700-311,500 reads per sample for the environmental samples and up to 2,600,000
320 reads for the mock communities.
321
322 **Targeted metagenomics read processing and mapping**
323 All six possible reading frames of the nucleotide sequence reads generated by the targeted
324 metagenomic sequencing of the mock community samples were translated using transeq from the
325 EMBOSS package (v6.6.0.0) (Rice et al., 2000). These were mapped to the genomes using
326 DIAMOND blastp (v2.0.6.144) (Buchfink et al., 2015), with a minimum percentage identity of
327 60%. This threshold was selected as it resulted in the largest percentage of mapped reads while
328 being stringent enough to prevent spurious mappings. No minimum coverage was used to account
329 for reads that did not fully overlap with the proteins from the mock community dataset. For each
330 direction, up to four matches were retained, to account for cases where a single read spanned
331 neighboring proteins. Multiple matches in the same direction could overlap by up to 15 amino
332 acids to account for possible protein fusions or sequencing errors. The blastx searches against the
333 *refseq* database (Altschul et al., 1990) were performed in order to remove duplicates or reannotate
334 falsely annotated entries due to the presence of homologous gene families in our dataset (*amoA*
335 vs. *pmoA*, *nxrB* vs. *narH* nitrate reductase or *napA*+*narG* vs. *nrfA*). The blastx search was done to
336 evaluate whether the read was correct or not, and the run was made with the default parameters of
337 blast. Each of the (up to four) matches were used to assign KOs to the reads, based on the predicted
338 KOs of the proteins to which they mapped. If, in a given direction, a read was assigned the same
339 KO multiple times, this was collapsed into a single hit, retaining the hit with the best E-value.
340

341 **Evaluation of functional annotation of reads**

342 Protein predictions for each read were performed using nhmmer from the HMMER suite
343 (v.3.3)(Eddy 2011) using the in-house generated HMMs for target genes as well as KOFAM as
344 described above. An inclusion threshold of <0.0001 was used when assigning protein predictions
345 for the reads. For a given read, using all six of the translated frames, up to four protein assignments
346 could be made, evaluated based on KO bit scores or identity filtering. Reads assigned KOs for the
347 function of interest were counted as positively identified hits. Subsequently, reads were mapped
348 to their corresponding coding sequences (CDS) from the mock community organisms to determine
349 which reads constituted "true positives" (TP) or "true negatives" (TN). Reads mapped to a given
350 CDS were assigned functions according to the above-mentioned KOFAM CDS annotation as their
351 "true function". If the HMM assignment (both in-house and from KOFAM) of a read matched the
352 functional assignment of the CDS to which it was mapped, then it was counted as a TP. The TP,
353 false positive (FP), and false negative (FN) values were summed up for all models using all
354 methods, and precision and recall were calculated. Read mapping (described above) was used to
355 determine the total number of possible TPs of the reads (Fig. S6), from which the actual number
356 of TPs was subtracted to calculate the number of FNs. Precision was calculated as the ratio of TPs
357 to total positive detections (i.e., number of TPs and FPs). Recall was calculated as the ratio of TPs
358 to all theoretically obtainable true positives (i.e., sum of TPs and FNs). The script is available on
359 Zenodo: https://doi.org/10.5281/zenodo.15752134; /probe-capture/tree/main/3-
360 other_scripts/Angus_scripts -folder.

361

362 **Shotgun metagenomics library preparation, sequencing with Illumina HiSeq and analysis**

363 Shotgun metagenomic sequencing was performed on environmental DNA (n = 3) from agricultural
364 soil and wetland soil (Table S3), extracted as described above for TDB production. Illumina
365 TruSeq library preparation with an insert size of 250bp was carried out with 1 µg of DNA and
366 sequencing was performed in Illumina HiSeq PE150 lane as a commercial service in Vienna
367 Biocenter Core Facility (VBCF, Austria) and Microsynth AG, Switzerland. In this approach, the
368 previously generated HMM profiles were used to identify the target gene pools (E < 0.001). The
369 predicted functions of each identified sequence were confirmed by tblastx against the *refseq*
370 database, using DIAMOND (Buchfink et al., 2015). Annotation and identity information was
371 further processed with awk in unix and in Rversion 3.5.3 to produce lists of community structures

372 for each target gene. The relative abundance of each gene was normalized against the sum of total
373 reads, Table S4: Relative abundance of gene x = (Abundance of gene x / total reads sequenced) x
374 100.
375
376 **_amoA_ gene amplicon sequencing**
377 Archaeal *amoA* gene amplicons from environmental DNA (n = 3) from the agricultural soil were
378 generated as described earlier by Siljanen et al., (2019), and sequencing was performed with
379 Illumina MiSeq PE250, in LGC, Münich, Germany. In brief, *amoA* gene amplicons were generated
380 using 0.6 ng DNA template with GoTaq DNA polymerase (Promega) and 0.5 µM primers
381 (CamoA_19F [5'-ATGGTCTGGYTWAGACG-3'] and TamoA_632R-4 [5'-
382 GCKGCCATCCATCKRTANGTCCA-3']) with a 5'-prime Illumina sequencing adaptor [5'-
383 CTCTTTCCCTACACGACGCTCTTCCGATCT-3'] for both end of above primers for making
384 the indexing possible in the sequencing service laboratory. Cycling conditions were the following:
385 1min initial denaturation at 95°C, 30sec denaturation at 94°C, 30sec annealing at 60°C and 45sec
386 extension at 72°C, with 35 cycles. PCR reactions were performed in triplicate for each sample (n
387 = 3), and purified with the High Pure PCR product purification kit (Roche). Sequencing indexes
388 were added by LGC, where libraries were equally pooled and sequenced with Illumina MiSeq
389 PE250.
390
391 **Processing of amplicon _amoA_ reads**
392 Archaeal *amoA* (Thaumarchaeal-*amoA*, *TamoA*) amplicon reads processing was carried out by a
393 custom-based Python script available at Zenodo: https://doi.org/10.5281/zenodo.15752134; /probe-
394 capture/tree/main/1-scripts/1-amplicon_seq_script_v1.sh -folder. Briefly, quality trimming of
395 sequencing reads was performed by truncating amplicon forward read R1 reads to 200bp and
396 reverse read R2 180bp, discarding all reads with an expected error greater than 0.5. The *amoA*
397 reads were dereplicated and amplicon sequence variants (ASVs) were assigned to taxonomic bins
398 described in the reference *amoA* gene database (Alves et al. 2018) using a 55% identity cutoff with
399 USEARCH8 (Edgar, 2010). Chimeras were filtered out with UCHIME (Edgar, 2011). For the
400 remaining ASVs, quality controlled R1 and R2 reads were fused together, classified with UCLUST
401 in QIIME (Caparaso et al., 2010) and clustered to OTUs with 97% sequence identity cut-off for
402 each taxonomic bins. Taxonomic bins of each *TamoA* gene cluster were generated using the

403 reference gene database and taxonomy by Alves et al. (2018) (Altschul et al. 1990; Callahan et al.,
404 2016). Relative abundance of *TamoA* taxonomic bins was calculated as follows: Relative
405 abundance of *TamoA* taxonomic bin x = (Abundance of *TamoA* taxonomix bin x / total *TamoA*
406 reads) x 100. In Fig. 4A, the relative abundance of each taxonomic bin (relative to the column
407 sum) was plotted for each sample. The sequencing depth for the *TamoA* gene amplicon library was
408 ~27,000-58,000 PE250 reads per sample up to 31Mb for the agricultural soil (Table S4).
409
410 **Processing of metagenomic *amoA* reads**
411 Archaeal *amoA* (Thaumarchaeal-*amoA*, *TamoA*) reads from targeted- and shotgun metagenomics,
412 were identified with the *TamoA* in-house HMM model. The forward reads were dereplicated and
413 assigned to the reference *amoA* gene database using a 55% identity cutoff with USEARCH8
414 (Edgar 2010). Chimeras were filtered out with UCHIME (Edgar 2011), all reads were clustered
415 with UCLUST with 97% sequence identity in QIIME and assigned to taxonomic bins based on the
416 closest *amoA* genes from the taxonomy database of OTUs (Alves et al. 2018) (Caporaso 2010).
417 Relative abundance was calculated as follows: Relative abundance of *TamoA* taxonomic bin x =
418 (Abundance of *TamoA* taxonomix bin x / total *TamoA* reads) x 100. In the targeted metagenomics
419 approach, multiple detection of the same taxonomic bin is possible because of six probes are used
420 for each sequence cluster. However, when relative abundance for each detected taxonomic bin is
421 calculated for targeted metagenomic, if multiple detection occurs in the relative abundance
422 calculation the community composition is balanced because of equal number of probes per each
423 sequence cluster. Therefore, the relative abundance of different sequencing methods can be
424 compared. The sequencing depth for the *TamoA* gene for three shotgun metagenomics *TamoA*
425 reads out of ~22-58M PE150 reads per sample up to 17 Gb, and for targeted metagenomics 4472-
426 8230 *TamoA* reads out of ~220-311k PE300 reads per sample up to 160 Mb, for the agricultural
427 soil (Table S4). The metagenomic reads processing is described in the custom-made script
428 available at Zenodo: https://doi.org/10.5281/zenodo.15752134; targeted with probe-capture:
429 tree/main/1-scripts/2-probe_capture_script_v1.sh -folder, and shotgun metagenomics:
430 tree/main/1-scripts/3-metagenomics_script_v1.sh -folder.
431
432 **Phylogenetic placement analysis of *nosZ* reads from shotgun and targeted metagenomics**
433 **libraries.**

434 A reference alignment and phylogeny for *nosZ* was generated from full length *nosZ* amino acid
435 sequences obtained from genomes downloaded from the NCBI genome database (accessed
436 October 2019). A profile HMM for *nosZ* was generated from a previously published *nosZ* dataset,
437 as described above (Graf et al, 2022). Reads were converted to amino acid sequences and then
438 aligned using *hmmalign* within the HMMer suite (v.3.1.2)(Eddy 2011), and the homologous
439 sections were identified using FastTree (v. 2.1.11, Price et al., 2010). The reference protein
440 alignment was used for phylogenetic reconstruction using IQ-TREE (v1.6.12) with best model
441 selection and 1000 ultrafast bootstraps. The Le-Gascuel (LG) substitution model with 10 rate
442 categories was selected after automatic model selection. For the phylogenetic placement of *nosZ*
443 reads obtained using either shotgun metagenome sequencing (n=3) or targeted metagenomics
444 (n=3) of the agricultural site, the relevant reads were pooled together and clustered with CD-HIT
445 (Fu et al., 2012) at 90% sequence identity before translating and aligning to the reference *nosZ*
446 alignment using *hmmalign*. The translated reads were then placed in the reference phylogeny using
447 the next generation evolutionary placement algorithm (EPA-NG v0.3.8; Barbera et al., 2019), with
448 the same model parameters as used for constructing the reference phylogeny. The placement
449 positions with the highest likelihood weight ratio were plotted on the reference phylogeny in R
450 using the 'ggtree' package (v3.4.2, Yu et al., 2016). Script available for nosZ phylogenetic
451 placements at Zenodo: https://doi.org/10.5281/zenodo.15752134; /probe-capture/tree/main/3-
452 other_scripts/Chris_Jones_scripts -folder.
453
454 **Data and script availability**
455 The shotgun metagenomic, targeted metagenomic and *amoA* amplicon sequencing data have been
456 deposited in NCBI SRA under the Bioproject numbers PRJNA898102 and PRJNA488558. All
457 scripts are available at Zenodo: https://doi.org/10.5281/zenodo.15752134. The repository contains
458 as well the target gene database https://doi.org/10.5281/zenodo.15752134; probe-capture/2-
459 selected_outputs/all_genes_compiled_ncbi_nt_envnt_fungene_split_file_0.fasta.gz -file (four
460 files ..file_0-3.fasta.gz), and the final probe sequences https://doi.org/10.5281/zenodo.15752134;
461 probe-capture/2-selected_outputs/Final_N_probes.list1.fasta -file.
462
463
464 **Results**

465

466 **Benchmarking gene detection and quantification using targeted metagenomics in mock**
467 **communities**

468 The targeted metagenomic approach was evaluated using mock microbial communities comprising
469 predefined mixtures of genomic DNA from 14 bacterial and three archaeal strains, together
470 containing all 14 functional genes targeted by the probes and mixed in different proportions to
471 produce six different median G+C contents ranging from 47% to 62% (Table S1). The overall
472 relative abundances of target genes in the captured metagenomes across all GC% categories were
473 correlated with their abundance in the original mock communities ($r = 0.78$ ($P < 0.0001$) (Fig. 2).
474 Comparisons between the relative abundance of individual genes in the mock communities as
475 compared to captured metagenomes revealed highly similar values (t = 1.8518e-10, df = 174.83,
476 $P > 0.05$), with the exception of the *nirK* and the *hzoA* genes (Fig. S1a,b). Pearson correlation
477 coefficients across all GC% categories varied between 0.76 and 0.93, although correlations were
478 stronger within higher GC% categories (57%-63%; R = 0.80-0.93) than within lower GC%
479 categories (47%-53%; R = 0.76-0.79) (Fig. S1). The probe capture targeted metagenomics
480 produced 61-72% of target gene sequences from the total sequenced library of the mock-
481 communities (Table S4).

482

483 **Precision of gene identification in mock communities**

484 Precision for the in-house HMMs ranged from 74.1% (*nifH*) to 100% (*mcrA*) (Fig. S5). The
485 average precision of the models was 93.4%, with a median of 99.8% (excluding *amoA* and *pmoA*).
486 The recall ranged from 42.8% (*nxrB*) to 100% (*mcrA*). The average recall was 81.9%, with a
487 median recall of 82.8%. Because KEGG only provides a single model covering the homologous
488 genes *pmoA* and *amoA*, it was not possible to accurately calculate the precision and recall for these
489 genes individually. A precision estimate could still be determined for the two genes, where a "true
490 positive" is defined as a hit for a read identified by the in-house models as *amoA* or *pmoA* mapped
491 to a CDS identified by the HMM for (*a*/*p*)*moA* genes provided by KEGG (K10944). In this case,
492 the precision values for the *amoA* and *pmoA* genes were estimated to be 96.9%, and 100%,
493 respectively. It is important to note that these estimates do not exclude the possibility that the in-
494 house models misclassified reads belonging to these homologous protein families.

495

496 Precision and recall values were similar using KOFAMs instead of in-house HMMs. However, the
497 results for *amoA* and *pmoA* genes were combined due to a single model for the associated KO for
498 both genes, as mentioned above. Precision ranged from 76.5% (*amoA*/*pmoA*) to 100.0% (*mcrA*),
499 while recall varied from 34.3% (*nifH*) to 100% (*mcrA* and *nxrB*). The average and median
500 precision values were 94.5% and 97.7%, while the average and median recall values were 93.39%
501 and 99.03%, respectively. The precision of KOs was lower for *napA* than for other genes when
502 compared with custom DNA HMMs. This could be accounted for by the presence of formate
503 dehydrogenase, a known homolog of *napA*. Formate dehydrogenase (K00123) was indeed
504 identified among the genomes used for the mock community, so it is possible that the custom
505 HMM was misclassifying reads originating from this known ortholog to *napA* (Cerqueira et al.,
506 2015).
507
508 **Higher number of reads for all target functional genes but similar relative abundance**
509 **obtained with targeted compared to shotgun metagenomics from complex environmental**
510 **samples**
511 We investigated the efficiency of the targeted metagenomics approach in natural complex
512 ecosystems by directly comparing this approach with shotgun untargeted metagenomics generated
513 with a nearly 100 times higher amount of sequencing (~66 Mb and ~6.4 Gb per sample,
514 respectively) from wetland soils in Bellefontaine, France, and from an agricultural field in Hungary
515 (Table S4). These soils represent distinct ecosystems with average physicochemical conditions,
516 such as water content and oxygen concentration, and N availability that favors different
517 distributions of functional genes (Jones et al. 2014). Despite the large difference in sequencing
518 depth, targeted metagenomics generated a much larger set of target functional genes (Fig. 3). In
519 both the agricultural and wetland site the targeted metagenomics approach detected 14 out of 15
520 gene-clusters and the shotgun approach 13 out of 15 genes (*mcrA* was not detected in the
521 agricultural site and *hzoA* not in the wetland site). Moreover, up to 60 times as many identified
522 gene sequences were detected using the probe capture compared to shotgun metagenomics (Fig.
523 3, Table S4). The relative abundances of functional genes in captured metagenomes correlated
524 significantly with that in shotgun metagenomes from both the agricultural field (R_{Pearson} = 0.96,
525 d.f.=14, $P < 0.00001$) and the wetland (R_{Pearson} = 0.70, d.f,=14, $P < 0.01$). The relative abundances
526 of target genes were different between the two approaches ($P > 0.00001$) (Fig. 3). Moreover,

527 increased relative abundance was detected with targeted compared to shotgun metagenomics in
528 the agricultural site for *nifH*, *TamoA*, *nrfA*, *napA*, *nirK*, *nirS*, *norB*, *nosZ* and *mmoX* (Fig. 3a) genes
529 and the wetland site for *nifH*, *napA, narG* and *nirS* (Fig. 3b) genes. The targeted approach
530 generated a much higher number of reads for all target genes in both sites than shotgun
531 metagenomics (Target reads in Table S4).
532
533 The different relative abundances of functional genes in the two ecosystems reflected known
534 differences in their environmental conditions: (i) N_2 fixation genes were expectedly higher in the
535 unfertilized wetland; (ii) Genes involved in ammonia oxidation and nitric and nitrous oxide
536 reduction were higher in the agricultural soil, which is likely to be more oxic and also subject to
537 N fertilization which increase inorganic N cycling; (iii) Genes involved in methanogenesis were
538 more abundant in the wetland, where anoxic conditions and methane production typically occur.
539
540 **Comparing diversity recovered for specific target genes from amplicon, targeted and**
541 **shotgun metagenomics from complex environmental samples**
542 In order to compare the diversity of specific functional genes detected by targeted metagenomics,
543 shotgun metagenomics and PCR amplicon sequencing we focused on the archaeal *amoA* gene as
544 a test case, as it is the second most sequenced gene in NCBI indicative of its ecological importance,
545 and for which there is a comprehensive reference database for classification available (Alves et
546 al., 2018). Moreover, archaeal ammonia oxidizers (AOA) typically outnumber their bacterial
547 counterparts in most environments (Leininger et al. 2006), and even though they can be low
548 abundant, their diversity is striking and underlines their distinct metabolic capabilities (Alves et
549 al. 2018; Huang et al. 2021). Additionally, we compared the detection of *nosZ* genes between the
550 targeted and shotgun metagenomics approaches in the agricultural soil, also based on an existing
551 reference database (Graf et al., 2022). The following datasets were compared: (i) targeted and
552 shotgun metagenomes, and PCR amplicons of the archaeal *amoA* gene in the agricultural soil (Fig.
553 4); (ii) *nosZ* gene sequences identified by targeted or shotgun metagenomics in the agricultural
554 soil (Fig. S3).
555
556 Comparison of results from all three methods in agricultural soil samples showed that targeted
557 metagenomics detected a substantially higher diversity of archaeal *amoA* genes than PCR

558 amplicon sequencing, while, as expected, shotgun metagenomics detected a lower diversity than
559 either of the other two methods (Fig. 4A). Specifically, targeted metagenomics detected 71-83
560 distinct taxonomic bins of *amoA* gene, whereas PCR amplicon sequencing and shotgun
561 metagenomics detected 65-68 taxonomic bins and only three bins, respectively. Rarefaction
562 analysis also indicated that targeted metagenomics retrieved a higher alpha diversity of archaeal
563 *amoA* genes than PCR amplicon sequencing (Fig. 4B,C) and it also reaches a plateau much faster
564 and with a lower number of sequences than amplicon sequencing (Fig. 4B).
565
566 Similarly, targeted metagenomes yielded a higher number of taxonomic bins for *nosZ* genes from
567 the agricultural soil (~6,800 bins) than shotgun metagenomics (~1,300 bins) at a 90% identity
568 cutoff, showing that the first approach detected a greater gene diversity with a much lower
569 sequencing depth (Fig. S3). Gene sequence reads from both methods mapped to similar lineages
570 in a comprehensive reference phylogeny of *nosZ* genes (Graf et al., 2022). However, a greater
571 proportion of reads from the targeted metagenomics approach were assigned to deeper nodes of
572 the phylogeny, particularly within the less studied clade II, suggesting that this method also
573 captured more novel *nosZ* variants not represented in the reference phylogeny.
574
575 Targeted metagenomics (sequencing depth 117–160 Mb/sample) yielded 29.1±3.1 *TamoA* reads
576 per 100,000 total reads, while shotgun metagenomics (7–17 Gb/sample) yielded only
577 0.0036±0.0039. The shotgun metagenomics produced only three quality-controlled *amoA*
578 taxonomic bins were detected (Fig. 4A). Conversely, targeted metagenomics generated five times
579 less *TamoA* reads (Fig. 4B) than PCR amplicon sequencing, with a sequencing depth of 15-31 Mb
580 per sample. Nevertheless, the targeted metagenomics approach detected a much higher gene
581 sequence diversity for both *TamoA* (Fig 4A, 4B) and *nosZ* (Fig. S3) than the two other methods in
582 the agricultural soil samples. For *TamoA*, the targeted approach detected 28 times higher alpha-
583 diversity than shotgun metagenomics, while 5 times higher diversity of *nosZ* genes was recovered,
584 importantly covering genes affiliated with the less characterized Clade II. In turn, the targeted
585 approach detected 1.24 times (24% more taxonomic bins) higher diversity of archaeal *amoA* genes
586 than PCR amplicon sequencing in agricultural soils, respectively.
587
588 **Discussion**

589 In this study, we report the development and evaluation of a new probe library for targeted
590 metagenomics of 14 key genes involved in the microbial cycling of inorganic N and CH_4. This
591 included the compilation of databases containing all variants of the target genes from public
592 databases, the development of an extensive set of probes targeting 14 well-established marker
593 genes, and a bioinformatics pipeline to process and annotate the sequence data. This approach has
594 been successfully used earlier to identify genes involved in microbial C metabolism in soil
595 (Manoharan et al., 2015, Manoharan et al., 2017).
596
597 Our custom HMM models could assign reads to the correct genomes based on full protein
598 sequences with higher precision than models based on KEGG Orthologs, especially when used to
599 predict putative functions from short DNA fragments, such as metagenomic reads. This accuracy
600 indicates that the HMMs can reliably classify reads in the mock community and are sensitive
601 enough to detect distantly related genes or novel gene variants. Notably, reads identified as false
602 positives may not actually be incorrect but could represent novel gene variants not captured by
603 KEGG Orthologs.
604
605 In validation experiments with mock communities comprising 18 distinct organisms encoding
606 multiple target genes, we showed that this method successfully reproduced the expected relative
607 abundances of nearly all target genes, with the exception of the *nirK* and *hzoA* genes. Although
608 10.3% of the probes targeted *nirK* genes of nitrifiers, these were detected at a lower relative
609 abundance than that present in the original mock community. However, the relative abundance of
610 *nirK* genes matched the frequency of *nirK* genes only in heterotrophic organisms in the mock
611 community, indicating a bias in the detection of distinct *nirK* gene variants. The sequencing
612 technology used can influence the sequences generated due to known GC biases. In this case, the
613 relative abundance of *nirK* genes may have been biased by the false low coverage of genomic
614 regions with non-optimal GC% content (50-60%) known from MiSeq Illumina sequencing
615 (Browne et al. 2020). Consistent with this observation, there was some variation in probe
616 specificity among mock communities with different genomic GC% content, with high GC%
617 communities having a generally higher target gene detection efficiency. For example, the mean
618 GC% content of nitrifier *nirK* genes is lower than that of heterotrophic *nirK* genes (53% and >60%,
619 respectively) and thus it is plausible that the later are sequenced more efficiently. This effect was

620 also noticeable among *nosZ* genes, as the *nosZ* gene of *D. fermentas*, with GC content of 52.9%
621 (Table S1, Fig. S1), was detected more efficiently than that of *S. denitrificans*, which has a 32.7%
622 GC content. Moreover, variation in detection efficiency may also result from differences in probe
623 hybridization. For instance, probes with higher GC% content are expected to bind more strongly
624 to their targets given the higher binding energy between guanine and cytosine than between
625 adenine and thymine. Although our approach successfully identified *hzoA* genes in the
626 environmental samples, it failed to detect them in mock communities. It is possible that this was
627 due to the PCR-generated genomic fragment used as source material for *hzoA* genes in the mock
628 community, which, unlike other target genes provided in the whole genomic DNA, is more
629 susceptible to degradation or structure-related constraints on probe hybridization.
630
631 Targeted metagenomics also closely reproduced the relative abundances of target genes obtained
632 with shotgun metagenomics in two distinct natural microbial communities from soils. The relative
633 gene abundances were strongly positively correlated between the two methods, indicating that the
634 probe capture step did not introduce a significant quantitative bias. Importantly, this relationship
635 held also true for *nirK* genes, confirming that the lower detection efficiency of nitrifier *nirK* genes
636 was due to a variation in GC content. This observation emphasizes that particular attention should
637 be given to the design of probes for low GC genes, such as by targeting gene regions that are not
638 highly conserved among all gene variants, where higher GC probes are more likely to outcompete
639 lower GC probes, as in the case of *nirK*. Nonetheless, as observed in the mock communities, the
640 relative gene abundances quantified in environmental samples were strongly positively correlated
641 between the two approaches to all GC% categories (Fig. S2). This further supports that the
642 observed variation in detection efficiencies among gene variants with targeted metagenomics has
643 a negligible effect on community profiles.
644
645 The thaumarchaeal *amoA* gene (*TamoA*) and the *nosZ* gene were used as case studies to investigate
646 the differences in the detected diversity of single functional guilds that are known to be highly
647 diverse (Alves et al. 2018, Graf et al. 2022). To determine the community composition of the
648 *TamoA* gene for the different sequencing approaches, the reads are mapped to a reference library,
649 in turn forming the taxonomic bins. Consequently, the same gene from the same organism can be
650 detected multiple times in targeted metagenomics due to multiple probe hybridization. However,

651 when calculation of relative abundance is performed, it balances the detected community because
652 an equal number of probes is used per sequence cluster and therefore the relative abundance of
653 targeted method can be compared to other sequencing tools. At first glance, PCR amplicon
654 sequencing seems to have resulted in 18 more taxonomic bins than targeted metagenomics.
655 However, these sequences were generated by a 35-cycle PCR, known to produce some chimeric
656 sequences. Examples of such errors have been shown even for shorter amplification cycles (Kozich
657 et al., 2013). Problems due to chimeric sequences in public databases from amplicon studies are
658 known and were encountered during the compilation of the thaumarchaeal *amoA* reference
659 database (Alves et al., 2018). Although we used standard pipelines for the detection of chimeric
660 sequences, it is possible that we might have missed some, which may be represented in the 18 bins
661 produced with amplicon sequencing, leading to an inflated number of sequences post-processing.
662 The absence of long amplification steps in targeted metagenomics is expected to preclude the
663 formation of chimeric sequences, which represents a significant advantage over typical PCR-based
664 approaches.
665
666 It should be noted that the efficiency of the targeted metagenomic approach largely depends on the
667 coverage and quality of the database used to generate the probes. We used hmmer profiles to
668 collect all possible genes from the NCBI nucleotide and WGS databases, as well as Fungene and
669 other published databases were used for the target gene library. Only one hmmer profile per gene
670 was used for detection, however, clade specific hmmer profiles could also be used to have a
671 broader outcome for neighbouring clades. We used fairly relaxed conditions when including
672 thresholds for screening the NCBI databases, because we wanted to have closely related organisms
673 and genes in the probe pool. For example, this was the case for the *mmoX* gene family, which also
674 includes butane, propane and toluene monooxygenases. We included these closely related *mmoX*
675 genes in the probes, therefore these closely related *mmoX* genes were found from coniferous trees
676 (Putkinen et al., 2021). However, the database is not constrained to full-length gene sequences
677 from a limited diversity of complete genomes and long metagenomic scaffolds available and the
678 probe/target diversity can be greatly extended through the inclusion of gene fragments generated
679 by PCR or from short metagenomic contigs. Despite its high target precision, this method may
680 nevertheless capture non-target sequences, especially for orthologous gene families who have
681 evolved different functions, as is the case for *napA* genes and formate dehydrogenases. Such cases

682 emphasize the importance of thorough data annotation and filtering procedures. With the target
683 gene database used to generate probes in this study, we managed to detect a higher diversity for
684 the case study of *TamoA* genes than the shotgun sequencing approach, showcasing the advantage
685 of "casting a wider net" with targeted metagenomics. Similar results were obtained in other
686 targeted metagenomics applications, such as for detection of diverse resistome-virulome elements
687 (Noyes et al. 2017) or for improved taxonomic microbial community characterization via 16S
688 rRNA enrichment (Beaudry et al. 2021).
689
690 Despite the technological advances and decline in cost of high-throughput sequencing, shotgun
691 metagenomics remain impractical and prohibitively expensive to capture the diversity of specific,
692 low abundant functional groups in complex environments, such as soils and sediments, especially
693 in longitudinal studies. The current cost of generating the probes for targeted metagenomics is
694 about 20-50 € per sample depending on the probe manufacturer. If the goal is to have a focused,
695 comprehensive view of the diversity of functional guilds involved in inorganic nitrogen or methane
696 cycling in a certain ecosystem, then targeted metagenomics can circumvent the high cost and
697 overabundance of data generated by shotgun metagenomics, as well as provide more quantitative
698 data and more information on the diversity of the genes of interest. In turn, PCR-based assays are
699 limited by the biases associated with any set of primers targeting a large diversity of sequences,
700 assay efficiencies which preclude the detection of highly divergent new sequence diversity as well
701 as known issues with the generation of chimeric sequences during amplification. The targeted
702 metagenomics approach circumvents these limitations by using multiple short probes targeting
703 different regions of each gene, which largely increases the likelihood of capturing sequence
704 variants that elude PCR amplification. Thus, this approach has not only the potential to capture
705 rare and novel gene diversity in complex environments, but also to identify cryptic microorganisms
706 in low-biomass samples or involved in suggested CH_4 metabolisms, such as in the tree
707 phyllosphere (Putkinen et al., 2021) and nitrogen cycling in coral holobiont (Glaze et al., 2021).
708 Moreover, targeted metagenomics can also overcome issues associated with running and
709 comparing multiple independent PCR assays when investigating several distinct targets. In that
710 sense, this approach effectively represents a PCR-independent, multiplex approach to characterize
711 simultaneously and in-depth the distributions of a broad range of functional genes, providing a
712 holistic view of the status of the nitrogen and methane cycles in the studied ecosystems. This is

713 especially advantageous when combined with functional studies, such as the determination of N-
714 transformation rates and *in situ* fluxes, as showcased by a study of N_2O emissions in thawing
715 Yedoma permafrost sites over time (Maruschak et al. 2021). In this study, the application of
716 targeted metagenomics with the N-cycling probe dataset presented here revealed that changes in
717 the N-cycling microbial community composition were responsible for an increase in N_2O
718 emissions in revegetated Yedoma soils, which had undergone thawing a decade prior.
719
720 Conclusively, the targeted metagenomics approach developed here provides an efficient and cost-
721 effective strategy for studying microbial functional guilds that typically represent small fractions
722 of natural microbiomes, and whose diversity is generally underestimated and highly
723 underrepresented in metagenomic datasets. This approach also circumvents the limitations and
724 biases associated with PCR-based methods and has higher potential to capture rare or novel
725 functional gene diversity.
726
727
728
729 **Acknowledgements**
730 H.S. was financially supported by the Atmosphere and Climate Competence Center (ACCC)
731 Academy of Finland project number 337550, and Nitrobiome project 342362, in addition H.S. was
732 supported by Saastamoinen foundation, FEMS Society and Niemi foundation. L.M., A.B., and
733 C.S. were funded by ERC Advanced Grant TACKLE (695192), and R.J.E.A. was funded by
734 project P25369 of the Austrian Science Fund (FWF). A.S.H and F.L.S were funded by ERC
735 Starting Grant EvolPhisiol (grant agreement 803768). S.H and C.M.J were supported by the
736 Swedish University of Agricultural Sciences (senior career grant 2019-2024). We thank Lea
737 Wittorf for providing DNA extracts from isolates for the mock libraries.
738

739 **Authors contribution statement**

740 H.S., C.S., R.A. designed study, H.S. collected the targeted gene library, R.A.,S.H.,C.J. provided
741 reference libraries for *amoA* and *nosZ* genes, L.M. designed the probes, L.M., A.B., H.S. extracted
742 the environmental DNA and did the shotgun metagenomics, H.S. and S.H. extracted and provided

743 the mock sample cultured organisms, A.H., F.S. did the precision and recall analysis between the
744 KEGG and HMM approaches, A.B., H.S. did the *amoA* diversity comparison, S.H.,C.J. did the
745 *nosZ* diversity comparison. H.S. wrote the first version of manuscript and all co-authors further
746 developed the text.

747 **References:**

748

749 Aanderud, Z.T., Jones, S.E., Fiererr, N., Lennon, J.T. 2015. Resuscitation of the rare biosphere
750 contributes to pulses of ecosystem activity. *Front Microbiol* 6: DOI=10.3389/fmicb.2015.00024

751

752 Alteio, L.V., Séneca, J., Canarini, A., Angel, R., Jansa J., Guseva, K., Kaiser, C., Richter, A.,
753 Schmidt, H. 2021. A critical perspective on interpreting amplicon sequencing data in soil
754 ecological research. *Soil Biol Biochem* 160:108357,
755 https://doi.org/10.1016/j.soilbio.2021.108357.

756

757 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. 1990. Basic local alignment
758 search tool. *J. Mol. Biol.* 215: 403-410.

759

760 Alves, R.J.E., Minh, B.Q., Urich, T., von Haeseler, A., Schleper, C. 2018. Unifying the global
761 phylogeny and environmental distribution of ammonia-oxidizing archaea based on amoA genes.
762 *Nat. Comm.* 9: 1517. https://doi.org/10.1038/s41467-018-03861-1.

763

764 Archer, D. 2010. The Global Carbon Cycle. Princeton: Princeton University Press. ISBN
765 9781400837076.

766

767 Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., Ogata, H. 2020.
768 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold.
769 *Bioinform.* 36:2251-2252. Doi: 10.1093/bioinformatics/btz859.

770

771 Barnard, R., Leadley, P. W., and Hungate, B. A. (2005), Global change, nitrification, and
772 denitrification: A review. *Glob. Biogeochem. Cycles* 19, GB1007, doi:10.1029/2004GB002282.

773

774 Beaudry, M.S. et al. (2021) Improved Microbial Community Characterization of 16S rRNA via
775 Metagenome Hybridization Capture Enrichment. *Front. Microbiol.* 12. Available at:
776 https://doi.org/10.3389/fmicb.2021.644662.
777
778 Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., Stamatakis, A. 2019. EPA-ng:
779 Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.* 68: 365-369.
780
781 Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., Korfi, K., Wang, J. 2019. Applications and
782 analysis of targeted genomic sequencing in cancer studies. *Comput. Struct. Biotechnol. J.* 17:1348-
783 1359. Doi: 10.1016/j.csbj.2019.10.004.
784
785 Buchfink, B., Xie, C., Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND.
786 *Nat. methods.* 12: 59.
787
788 Bodelier, P.L.E., Laanbroek, H.J. 2004. Nitrogen as a regulatory factor of methane oxidation in
789 soils and sediments. *FEMS Microbiol. Ecol.* 47: 265–277, https://doi.org/10.1016/S0168-
790 6496(03)00304-0
791
792 Bonilla-Rosso, G., Wittorf, L., Jones, C. *et al.* 2016. Design and evaluation of primers targeting
793 genes encoding NO-forming nitrite reductases: implications for ecological inference of
794 denitrifying communities. *Sci. Rep.* 6: 39208 https://doi.org/10.1038/srep39208
795
796 Borrel, G., Adam, P.S., McKay, L.J. *et al.* Wide diversity of methane and short-chain alkane
797 metabolisms in uncultured archaea. *Nat. Microbiol.* **4**, 603–613 (2019).
798 https://doi.org/10.1038/s41564-019-0363-3
799
800 Browne, P.D. Nielsen. T.K., Kot, W., Aggerholm, A., Thomas, M., Gilbert, P., Puetz, L.,
801 Rasmussen, M., Zervas, A., Hansen L.H. 2020. GC bias affects genomic and metagenomic
802 reconstructions, underrepresenting GC-poor organisms. *GigaScience* 9: giaa008,
803 https://doi.org/10.1093/gigascience/giaa008
804

Bräuer, S.L,, Basiliko, N., Siljanen, H.M.P., Zinder, S.H. 2021. Methanogenic archaea in peatlands. *FEMS Microbiol. Lett*. 367: fnaa172, https://doi.org/10.1093/femsle/fnaa172

Callahan, B.J., McMurde, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13: 581–583. https://doi.org/10.1038/nmeth.3869.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Gonzalez Peña, A., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zeneveld, J., Knight, R., 2010. QIIME allows analysis of Highthroughput community sequencing data. *Nat. Methods* 7: 335–336.

Cerqueira, N.M.F.S.A., Gonzalez, P.J., Fernandes, P.A., Moura, J.J.G., Ramos, M.A. 2015. Periplasmic Nitrate Reductase and Formate Dehydrogenase : Similar Molecular Architectures with Very Different Enzymatic Activities. *Acc. Chem. Res.* 48: 2875-2884 DOI: 10.1021/acs.accounts.5b00333

Clark, D.R., McKew, B.A., Binley, A. et al. 2022. Hydrological properties predict the composition of microbial communities cycling methane and nitrogen in rivers. *ISME Commun.* 2: 5 https://doi.org/10.1038/s43705-022-00087-7

Cleary, B. et al. 2015. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotech.* 33: 1053–1060. https://doi.org/10.1038/nbt.3329.

Conrad, R. 2009. The global methane cycle: recent advances in understanding the microbial processes involved. *Environ. Microbiol. Rep.* 1: 285-292. https://doi.org/10.1111/j.1758-2229.2009.00038.x

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S., Aragão, L. E. O. C., Arneth, A., Arora, V., Bates, N. R., Becker, M., Benoit-Cattin, A., Bittig, H. C., Bopp, L., Bultan, S., Chandra, N., Chevallier, F., Chini, L. P., Evans, W., Florentie, L., Forster, P. M., Gasser, T., Gehlen, M., Gilfillan, D., Gkritzalis, T., Gregor, L., Gruber, N., Harris, I., Hartung, K., Haverd, V., Houghton, R. A., Ilyina, T., Jain, A. K., Joetzjer, E., Kadono, K., Kato, E., Kitidis, V., Korsbakken, J. I., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Liu, Z., Lombardozzi, D., Marland, G., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P. I., Pierrot, D., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Schwinger, J., Séférian, R., Skjelvan, I., Smith, A. J. P., Sutton, A. J., Tanhua, T., Tans, P. P., Tian, H., Tilbrook, B., van der Werf, G., Vuichard, N., Walker, A. P., Wanninkhof, R., Watson, A. J., Willis, D., wiltshire, A. J., Yuan, W., yue, X., and Zaehle, S. 2020. Global Carbon Budget. *Earth Syst. Sci. Data* 12: 3269–3340, https://doi.org/10.5194/essd-12-3269-2020, 2020.

Falkowski, P.; Scholes, R. J.; Boyle, E.; Canadell, J.; Canfield, D.; Elser, J.; Gruber, N.; Hibbard, K.; Högberg, P.; Linder, S.; MacKenzie, F. T.; Moore b, 3.; Pedersen, T.; Rosenthal, Y.; Seitzinger, S.; Smetacek, V.; Steffen, W. 2000. The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System. *Science.* 290: 291–296. Doi:10.1126/science.290.5490.291

Fu, L., Niu, B. Zhu, Z., Wu, S., Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 28: 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Futema, M., Plagnol, V., Whittall, R.A., et al. 2012. Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. *J. Med. Genet.* 49:644-649.

Galloway, J. N., Dentener, F. J., Capone, D. G., Boyer, E. W., Howarth, R. W., Seitzinger, S. P., G. P. Asner, C. C. Cleveland, P. A. Green, Holland, E. A., Karl, D. M., Michaels, A. F., Porter, J. H., Townsend, A. R., & Vörösmarty, C. J. (2004). Nitrogen Cycles: Past, Present, and Future. *Biogeochemistry* 70: 153–226. http://www.jstor.org/stable/4151466

Gasc, C., Peyretaillade, E. and Peyret, P. (2016) Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic. Acids Res.* 44: 4504–4518. https://doi.org/10.1093/nar/gkw309.

871 Graf, D.R., Jones, C.M., Hallin, S. 2014. Intergenomic comparisons highlight modularity of the
872 denitrification pathway and underpin the importance of community structure for N_2O emissions.
873 *PloS one* 9: e114118.
874
875 Graf, D.R.H., Jones, C.M., Zhao, M., Hallin, S. 2022. Assembly of root-associated N2O-reducing
876 communities of annual crops is governed by selection for nosZ clade I over clade II. *FEMS*
877 *Microbiol. Ecol.* 98:1-11
878
879 Gruber, N., Galloway, J. An Earth-system perspective of the global nitrogen cycle. 2008. *Nature*
880 451, 293–296. https://doi.org/10.1038/nature06592
881
882 Glaze, T.D., Erler, D.V., Siljanen, H.M.P. 2021. Microbially facilitated nitrogen cycling in tropical
883 corals. *ISME J.* 16: 68–77 doi: 10.1038/s41396-021-01038-1.
884
885 Eddy, S.R.. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755–763,
886 https://doi.org/10.1093/bioinformatics/14.9.755
887
888 Eddy S.R. 2011. Accelerated profile HMM searches. *PLOS Comp. Biol.*, 7:e1002195, 2011
889
890 Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *BMC*
891 *Bioinformatics* 26: 2460–2461
892
893 Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves
894 sensitivity and speed of chimera detection. *BMC Bioinformatics* 27: 2194–2200.
895
896 Fish, J. A., Chai, B., Wang, Q., Sun, Y., Brown, C. T., Tiedje, J. M., Cole, J. R. 2013.
897 FunGene: the Functional Gene Pipeline and Repository. *Front. Microbiol.* 4: 291.
898
899 Hallin S, Lindgren PE. 1999. PCR detection of genes encoding nitrite reductase in denitrifying
900 bacteria. *Appl Environ Microbiol.* 65:1652-7. doi: 10.1128/AEM.65.4.1652-1657.1999.

Hoang, D.T., Chernomor, O., Von Haeseler, A. et al. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 35:518

Huang, L., Chakrabarti, S., Cooper, J., Perez,A., John, S.M., Daroub, S.H., Martens-Habbena, W. 2021. Ammonia-oxidizing archaea are integral to nitrogen cycling in a highly fertile agricultural soil. *ISME Comm.* 1:19 ; https://doi.org/10.1038/s43705-021-00020-4

Håvelsrud, O.E., Haverkamp, T.H., Kristensen, T. *et al.* A metagenomic study of methanotrophic microorganisms in Coal Oil Point seep sediments. *BMC Microbiol* **11**, 221 (2011). https://doi.org/10.1186/1471-2180-11-221

Hyatt, D., Chen, G.L., LoCascio, R.F. *et al.* 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119 https://doi.org/10.1186/1471-2105-11-119

Ichida, H., Morita, R., Shirakawa, Y., Hayashi, Y., Abe, T. 2019. Targeted exome sequencing of unselected heavy-ion beam-irradiated populations reveals less-biased mutation characteristics in the rice genome. *Plant J.* 98:301-314. Doi: 10.1111/tpj.14213.

Jones, C.M., Graft, D.R.H., Bru, D., Philippot, L., Hallin, S. 2013. The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *ISME J* 7: 417-426.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K. et al. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–9.

Kamil, G., Yoon, J.Y., Yoo, S. et al. 2021. Clinical relevance of targeted exome sequencing in patients with rare syndromic short stature. *Orphanet J Rare Dis* 16: 297 https://doi.org/10.1186/s13023-021-01937-8

931 Katoh, K., Rozewicki, J., Yamada, K.D. 2019. MAFFT online service: multiple sequence
932 alignment, interactive sequence choice and visualization. *Brief Bioinform* 20: 1160–66.
933
934 Knief, C. 2015. Diversity and Habitat Preferences of Cultivated and Uncultivated Aerobic
935 Methanotrophic Bacteria Evaluated Based on *pmoA* as Molecular Marker. *Front Microbiol* 6:1346.
936 Doi: 10.3389/fmicb.2015.01346
937
938 Kushwaha, S.K., Manoharan, L., Meerupati, T., Hedlund, K., & Ahrén, D. 2015. MetCap: a
939 bioinformatics probe design pipeline for large-scale targeted metagenomics. *BMC bioinformatics*
940 16: 65.
941
942 Kuypers, M., Marchant, H. & Kartal, B. 2018. The microbial nitrogen-cycling network. *Nat Rev*
943 *Microbiol* 16: 263–276. https://doi.org/10.1038/nrmicro.2018.9
944
945 Kolb S, Knief C, Stubner S, Conrad R. 2003. Quantitative detection of methanotrophs in soil by
946 novel pmoA-targeted real-time PCR assays. *Appl Environ Microbiol.* 69:2423-9. Doi:
947 10.1128/AEM.69.5.2423-2429.2003.
948
949 Lan, X., K.W. Thoning, and E.J. Dlugokencky: Trends in globally-averaged CH4, N2O, and SF6
950 determined from NOAA Global Monitoring Laboratory measurements. Version 2023-09,
951 https://doi.org/10.15138/P8XG-AA10
952
953 Letunic, I, Bork, P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new
954 developments. *Nucleic Acids Res* 47:W256–9
955
956 Li, J., Liu, T., McIlroy, S.J. et al. 2023. Phylogenetic and metabolic diversity of microbial
957 communities performing anaerobic ammonium and methane oxidations under different nitrogen
958 loadings. *ISME Commun.* 3, 39. https://doi.org/10.1038/s43705-023-00246-4
959
960 Liu, F., Zhang, J. & Mei, Y. 2016. The origin of the cooperativity in the streptavidin-biotin system:
961 A computational investigation through molecular dynamics simulations. *Sci. Rep.* 6, 27190.
962 https://doi.org/10.1038/srep27190

Ludwig, W., Strunk, O., Westram, R., et al. 2004. ARB: a software environment for sequence data. *Nucleic Acids Res.* 32: 1363-1371.

Manoharan, L., Kushwaha, S. K., Hedlund, K., Ahrén, D. (2015). Captured metagenomics: large-scale targeting of genes based on 'sequence capture' reveals functional diversity in soils. *DNA Res.* 22: 451-460.

Manoharan, L., Kushwaha, S.K., Ahrén, D., Hedlund, K. 2017. Agricultural land use determines functional genetic diversity of soil microbial communities. *Soil Biol Biochem* 115: 423-432

Morris A., Meyer K., Bohannan B. 2020. Linking microbial communities to ecosystem functions: what we can learn from genotype–phenotype mapping in organisms. *Phil. Trans. R. Soc.* B3752019024420190244. https://doi.org/10.1038/s41467-021-21400-3

Murrell, J.C., Jetten, M.S.M. 2009. The microbial methane cycle. *Environ Microb Rep* 1: 279-284. https://doi.org/10.1111/j.1758-2229.2009.00089.x

Myhre, G.G.D., Shindell, F.M., Bréon, W., Collins, J., Fuglestvedt, J., Huang, J., Koch, D.,Lamarque, J.F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., Zhang, H., 2013. Anthropogenic and natural radiative forcing. In:Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A.,Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013 the Physical ScienceBasis. Cambridge University Press, Cambridge, United Kingdom and New York, NY,USA Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Nelson, M.B., Martiny , A.C., Martiny, J.B.H. 2016. Global biogeography of microbial nitrogen-cycling traits in soil. *PNAS* 113 (29) 8033-8040. https://doi.org/10.1073/pnas.1601070113

Nguyen, L.T., Schmidt, H.A., von Haeseler, A. et al. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–74.

995 Nnadozie, C.F., Lin, J. and Govinden, R. (2015) 'Selective isolation of bacteria for metagenomic
996 analysis: Impact of membrane characteristics on bacterial filterability'. *Biotechnol Prog.* 31: 853–
997 866. https://doi.org/10.1002/btpr.2109.
998
999 Noyes, N.R. et al. (2017) Enrichment allows identification of diverse, rare elements in
1000 metagenomic resistome-virulome sequencing. *Microbiome* 5. 142.
1001 https://doi.org/10.1186/s40168-017-0361-8.
1002
1003 R Core Team 2018. R: A language and environment for statistical computing. R Foundation for
1004 Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
1005
1006 Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software
1007 Suite. Trends Genet. 16: 276-7. doi: 10.1016/s0168-9525(00)02024-2. PMID: 10827456.
1008
1009 Romdhane, S., Spor, A., Aubert, J. *et al.* 2021. Unraveling negative biotic interactions determining
1010 soil microbial community assembly and functioning. *ISME J.* https://doi.org/10.1038/s41396-021-
1011 01076-9
1012
1013 Palmer, K., Biasi, C., Horn, M.A. 2012. Contrasting denitrifier communities relate to contrasting
1014 N_2O emissions patterns from acidic peat soils in arctic tundra. *ISME J.* 6:1058-1077.
1015
1016 Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the
1017 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25:
1018 1043-1055.
1019
1020 Pan Y, Bodrossy L, Frenzel P, Hestnes AG, Krause S, Lüke C, Meima-Franke M, Siljanen H,
1021 Svenning MM, Bodelier PL. Impacts of inter- and intralaboratory variations on the reproducibility
1022 of microbial community analyses. *Appl Environ Microbiol.* 2010 Nov;76(22):7451-8. doi:
1023 10.1128/AEM.01595-10
1024

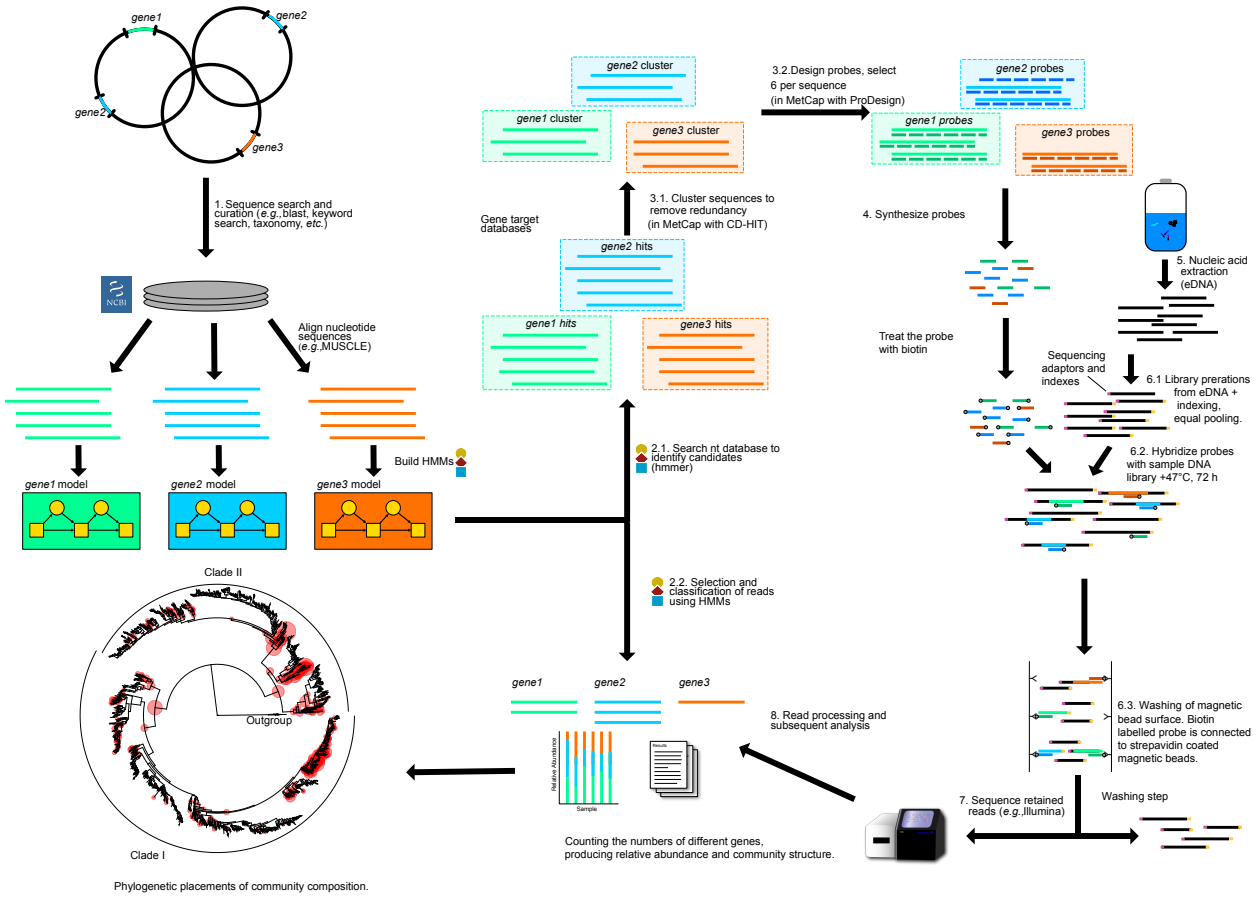
1025 Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for
1026 large alignments. *PLoS ONE* 5: e9490.
1027
1028 Prosser, JI, Hink, L, Gubry-Rangin, C, Nicol, GW. Nitrous oxide production by ammonia
1029 oxidizers: Physiological diversity, niche differentiation and potential mitigation strategies. Glob
1030 Change Biol. 2020; 26: 103– 118. https://doi.org/10.1111/gcb.14877
1031
1032 Putkinen, A., Siljanen, H.M.P., Laihonen, A., Paasisalo, I., Porkka, K., Tiirola, M., Haikarainen,
1033 I., Tenhovirta, S. and Pihlatie, M. 2021. New insight to the role of microbes in the methane
1034 exchange in trees: evidence from metagenomic sequencing. *New Phytol* 231: 524-536.
1035 https://doi.org/10.1111/nph.17365
1036
1037 Saghaï, A., Pold, G., Jones, C.M., Hallin, S. 2023. Phyloecology of nitrate ammonifiers and their
1038 importance relative to denitrifiers in global terrestrial biomes. *Nature Communications*, 14:8249
1039
1040 Skoglund, P., et al. 2014. Separating endogenous ancient DNA from modern day contamination
1041 in a Siberian Neandertal. *Proc Natl Acad Sci USA* 111: 2229–2234 .

1042 Siljanen, H.M.P., Alves, R.J.E., Ronkainen, J.G., Lamprecht, R.E., Bhattarai, H.R., Bagnoud, A.,
1043 Marushchak, M.E., Martikainen, P.J. Schleper, C., Biasi, C. 2019. Archaeal nitrification is a key driver of
1044 high nitrous oxide emissions from arctic peatlands. *Soil Biol Biochem* 137: 107539
1045
1046 Siljanen, H.M.P., Saari, A., Krause, S., Lensu, A., Abell, G.C.J., Bodrossy, L., Bodelier, P.L.E.,
1047 Martikainen, P.J. 2011. Hydrology is reflected in the functioning and community composition of
1048 methanotrophs in the littoral wetland of a boreal lake. *FEMS Microbiol Ecol*. 75:430–445,
1049 https://doi.org/10.1111/j.1574-6941.2010.01015.x
1050
1051 Sinha, R. et al. 2017. Assessment of variation in microbial community amplicon sequencing by
1052 the Microbiome Quality Control (MBQC) project consortium. *Nature Biot* 35: 1077–1086.
1053 https://doi.org/10.1038/nbt.3981.
1054

1055 Stamatakis, A. 2014 RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large
1056 phylogenies. *Bioinformatics* 30:1312–13.
1057
1058 Stein, L.Y., Klotz, M.G. 2016. The nitrogen cycle. *Curr Biol.* 26: R94-R98,
1059 https://doi.org/10.1016/j.cub.2015.12.021.
1060
1061 Sui, H.Y. et al. (2020) Impact of DNA Extraction Method on Variation in Human and Built
1062 Environment Microbial Community and Functional Profiles Assessed by Shotgun Metagenomics
1063 Sequencing. *Front Microb.* 11: . Available at: https://doi.org/10.3389/fmicb.2020.00953
1064
1065 Trifinopoulos, J., Nguyen, L.T., von Haeseler, A., Minh, B.Q. 2016. W-IQ-TREE: a fast online
1066 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44(W1):W232-5. doi:
1067 10.1093/nar/gkw256.
1068
1069 Thomson R.A.J., Giannopoulos, G., Pretty, J., Baggs, E.M., David J. 2012. Biological sources
1070 and sinks of nitrous oxide and strategies to mitigate emissions. *Phil. Trans. R. Soc. B* 2012 367,
1071 doi: 10.1098/rstb.2011.0415, published 26 March 2012
1072
1073 Throbäck, I.N., Enwall, K., Jarvis, Å. and Hallin, S. 2004. Reassessing PCR primers targeting nirS,
1074 nirK and nosZ genes for community surveys of denitrifying bacteria with DGGE. *FEMS Microbiol*
1075 *Ecol* 49: 401-417. https://doi.org/10.1016/j.femsec.2004.04.011
1076
1077 Orellana, L.H., Hatt, J.K., Iyer, R. *et al.* Comparing DNA, RNA and protein levels for measuring
1078 microbial dynamics in soil microcosms amended with nitrogen fertilizer. *Sci Rep.* **9**, 17630 (2019).
1079 https://doi.org/10.1038/s41598-019-53679-0
1080
1081 Ouyang Y., Norton J.M. 2020. Nitrite Oxidizer Activity and Community Are More Responsive
1082 Than Their Abundance to Ammonium-Based Fertilizer in an Agricultural Soil. *Front Microbiol.*
1083 11: DOI=10.3389/fmicb.2020.01736
1084

1085 Wheeler, T.J., Eddy, S.R. 2013. nhmmer: DNA homology search with profile HMMs.
1086 *Bioinformatics* 29: 2487-2489.
1087
1088 Yu, G.C., Smith, D.K., Zhu, H.C., Guan, Y., Lam, T.T.Y. 2017. GGTREE: an R package for visualization
1089 and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecol Evol.*
1090 8: 28-36.
1091
1092 Zaehle S. 2013.Terrestrial nitrogen–carbon cycle interactions at the global scale. Phil. Trans. R.
1093 Soc. B3682013012520130125 http://doi.org/10.1098/rstb.2013.0125
1094
1095 Zhao, J., Huang, L., Chakrabarti, S. et al. 2023. Nitrogen and phosphorous acquisition strategies
1096 drive coexistence patterns among archaeal lineages in soil. *ISME J.* 17, 1839–1850
1097 https://doi.org/10.1038/s41396-023-01493-y
1098
1099

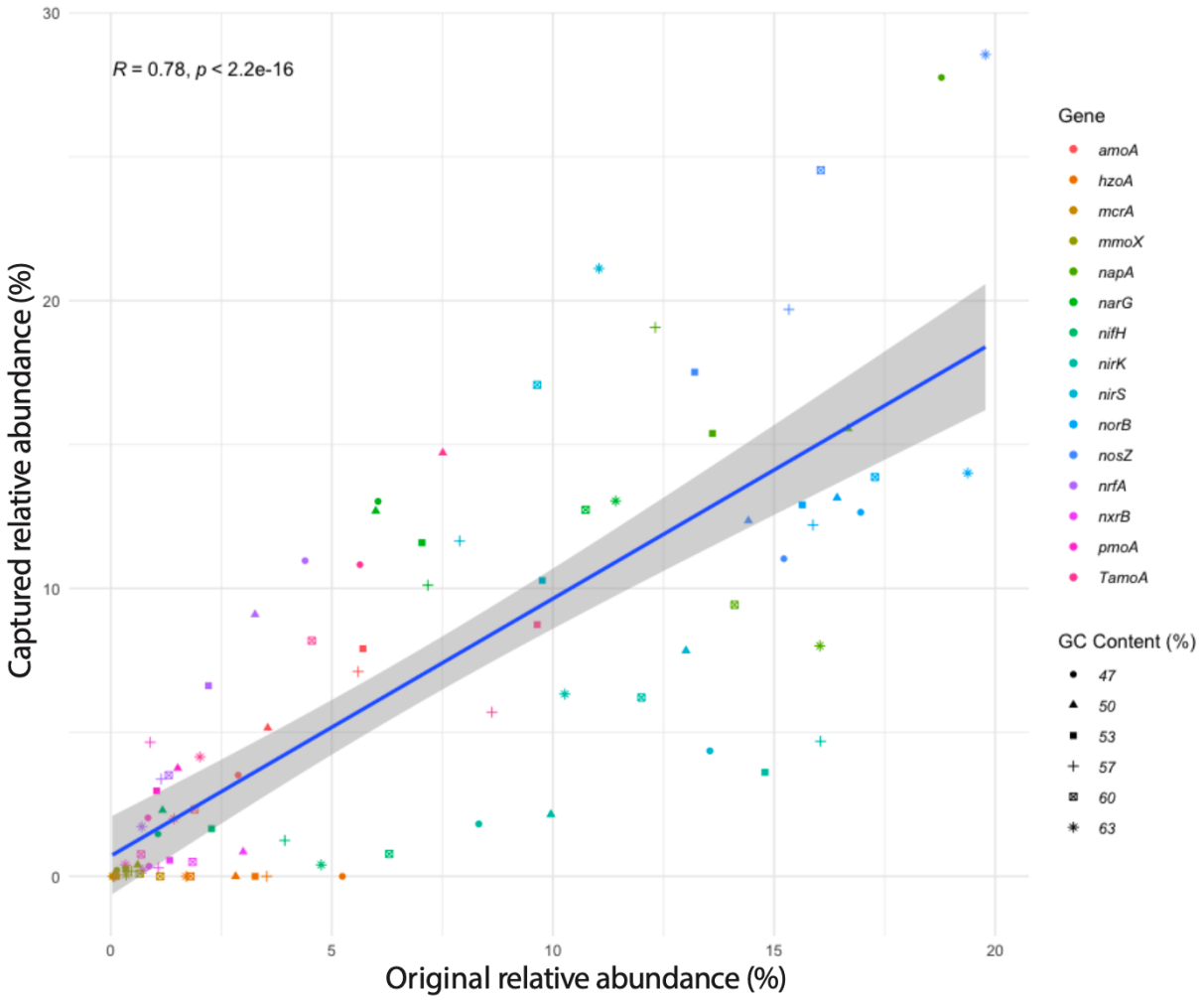
1100 **Figures:**



1101

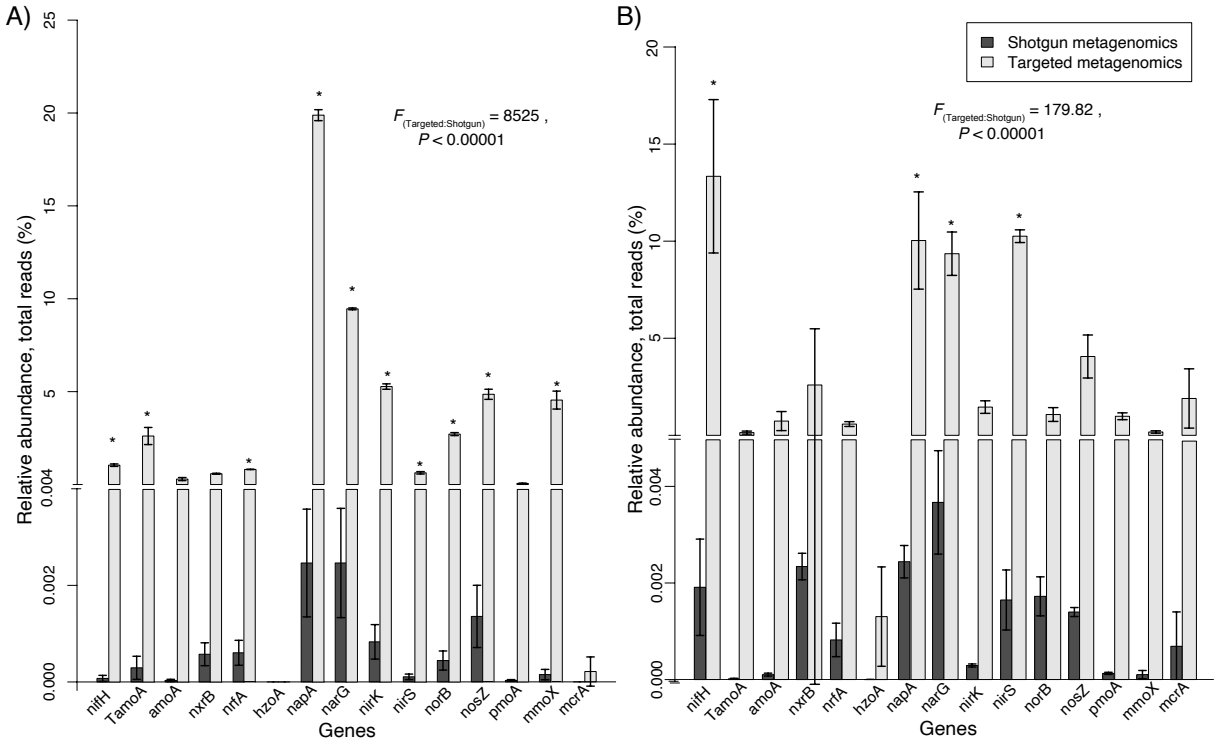
1102 Fig. 1 Illustration of the pipeline for the generation of HMM models for each gene, and how these
1103 models are used to study community composition with targeted metagenomics: 1. Steps followed
1104 for the generation of sequence databases for the genes of interest and subsequent building of the
1105 respective HMM models for each gene; 2.1. The produced HMM models were used to recruit more
1106 gene variants from the NCBI nt database and generate a target gene database for probe design; 2.2.
1107 The same HMM models were also used to classify the reads captured downstream; 3.1. Clustering
1108 of the target gene database with CD-HIT (in MetCap pipeline); 3.2. Design and selection of six
1109 50-mer probes for each sequence cluster (in MetCap pipeline); 4. Synthesis and biotinylation of
1110 the designed probes by Roche; 5. DNA extraction of the sample of interest; 6.1. Sequencing
1111 libraries preparation and addition of sequencing adaptors and indexes; 6.2. Hybridization of the
1112 probes to indexed DNA libraries for 72 hours at +47°C; 6.3. Purification of the DNA hybridized
1113 to the probes with streptavidin coated magnetic beads; 7. Sequencing of the hybridized libraries

1114 with Illumina MiSeq; 8. Read processing with HMM profiles and taxonomy assignment for the
1115 genes of interest.
1116
1117



1118
1119
1120 Fig. 2. Comparison between the original composition (DNA of strains applied into the mock
1121 sample) of the mock community and the relative abundance for each functional gene detected
1122 using probe captured metagenomics for all GC% combinations combined. The relative abundance
1123 in the mock community was calculated based on the amount of each functional gene and organism
1124 as well as genome size. Color of the symbols depict gene and shape depicts the category of GC
1125 content (%).
1126

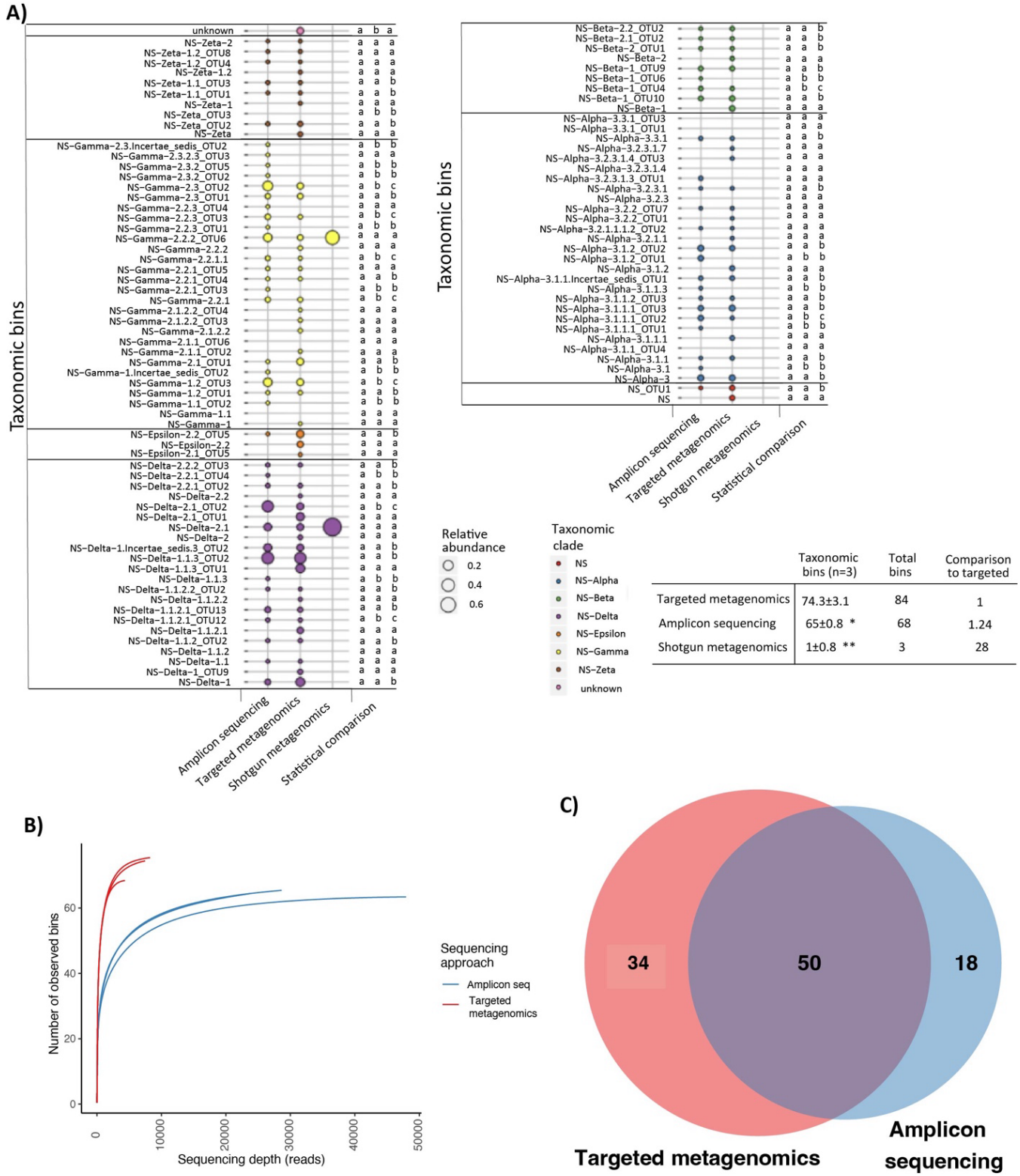
1127



1128

1129 Fig. 3. The relative abundance of functional genes (calculated against total sequencing reads)
1130 involved in inorganic nitrogen and methane cycling obtained by targeted metagenomics and
1131 shotgun metagenomics from A) agricultural soil; and B) wetland soil. Relative abundance was
1132 retrieved with HMMER searches of each functional gene. Statistically significant differences
1133 between shotgun and targeted relative abundances according to pairwise comparisons with
1134 ANOVA are shown with asterisk (P < 0.05, n = 3). The comparison of shotgun vs. targeted
1135 metagenomic measured with ANOVA is shown as F values and significance in the inset for both
1136 samples.

1137



	Taxonomic bins (n=3)	Total bins	Comparison to targeted
Targeted metagenomics	74.3±3.1	84	1
Amplicon sequencing	65±0.8 *	68	1.24
Shotgun metagenomics	1±0.8 **	3	28

1142
1143 Fig. 4. A) The relative abundance of archaeal *amoA* taxonomic bins produced from amplicon
1144 sequencing with *amoA* amplicon sequencing (*amoA* gene PCR amplicons sequenced with Illumina
1145 MiSeq, about 22000-47000 *amoA* reads per sample) (n = 3), metagenomics reads of targeted
1146 metagenomics (genomic DNA sequenced with Illumina MiSeq, after target enrichment, about
1147 4500-8500 *amoA* reads out of 220000-310000 targeted reads per sample)(n = 3), or shotgun
1148 metagenomics (genomic DNA sequenced with Illumina HiSeq, about 22-57 million reads per
1149 sample, of which only two were archaeal *amoA*)(n = 3). In all cases, the analysis pipeline consisted
1150 of USEARCH comparison on *amoA* reference database (Alves et al., 2018) and taxonomy were
1151 analyzed with QIIME1 analytical pipeline. Sample having largest diversity is shown (n = 3).
1152 Bubble symbol size is proportional to the relative abundance of each taxonomic bin produced with
1153 sequencing approach in each sequencing method. Color indicates Nitrosospheria (NS)-clade of
1154 taxonomic bin. Statistically significant differences in abundance of each taxonomic bins between
1155 sequencing approaches based on Pairwise-Wilcox comparison are indicated with different letters
1156 ($P < 0.05$, n=3), and the significant difference of comparison for taxonomic bins is shown an
1157 asterisk, * $P<0.05$, ** $P<0.01$. B) The rarefaction analysis of archaeal *amoA* obtained by amplicon
1158 sequencing and captured metagenomics. C) Venn-diagram of archaeal *amoA* sequences obtained
1159 by amplicon sequencing and targeted metagenomics.
1160

1161