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 results in limited numbers of sequences of rare taxa or functional genes, which prevents in-depth analyses of specific populations and functional groups. Amplicon-based approaches provide more sequences, but the drawback is the difficulty in designing primers that cover the diversity of the target group. Targeted metagenomics overcomes these challenges, as it results in similar community structure as traditional amplicon sequencing does but broadens the sequence space in a less biased metagenomic-based strategy. It is thus a useful tool when studying the diversity of specific populations in complex natural communities. We present and evaluate here a probe library designed for targeted metagenomics of nitrogen and methane cycling genes in complex communities.

37

Abstract

Microbes are the key players in the global cycling of nitrogen (N) and carbon (C), controlling the availability and fluxes of C and N in ecosystems, as well as being responsible for losses through the emissions of the powerful greenhouse gasses nitrous oxide (N_2O) and methane (CH_4). Thus, characterization of microbial functional guilds involved in these processes is high on the scientific agenda. Yet, standard sequence-based characterization methods often reveal only a minor fraction of their diversity in nature due to their frequent low relative abundance, insufficient sequencing depth of traditional metagenomes of complex communities, and limitations in coverage and efficiency of PCR-based assays. Here, we developed and tested a targeted metagenomic 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 14 target marker genes, comprising 264,000 unique probes in total. These probes were used to selectively enrich the target genes in shotgun metagenomic libraries. In validation experiments with mock communities of cultured microorganisms, the target gene profiles were similar to those of the original community when sequenced with targeted metagenomics. Furthermore, relative abundances of the marker genes obtained by targeted and shotgun metagenomics from agricultural and wetland soils correlated positively, indicating that the targeted approach did not introduce a significant quantitative bias. However, targeted metagenomics generated substantially higher diversity in terms of taxonomic coverage, and a larger number of sequence reads per sample, which allowed 41 or 1.2 times higher diversity estimates than when using shotgun metagenomics or targeted PCR amplification, respectively. Thus, targeted metagenomics complements current approaches by enabling a targeted, more detailed characterization of the diversity of key functional genes involved in N and CH_4 cycling within and between ecosystems.

Introduction

The global nitrogen (N) and carbon (C) cycles are essential processes of the Earth's biosphere (Zaehle 2013, Gruber and Galloway, 2008; Archer 2010; Falkowski et al., 2000; Steffen et al. 2015). All major N transformation processes, crucial for ecosystem functioning (i.e., nitrogen fixation, nitrification, denitrification and anaerobic ammonium oxidation, or anammox), are performed exclusively by functional guilds comprising 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, but also result in emissions of the potent greenhouse gas nitrous oxide (N_2O) and other volatile N compounds mainly through nitrification and denitrification (Barnard, et al., 2005, Thompson et al., 2012 Prosser et al. 2020). Methane (CH_4) is also a major product of the microbial trophic chain driving C cycling, and, like most inorganic N compounds, its production and consumption is 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).

Microbes participating in inorganic N and CH_4 cycling typically constitute a small fraction of the total microbial biomass and are usually rare biosphere organisms 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). Traditional 16S rRNA gene-based community analyses suffer from a number of limitations (Alteio et al. 2021), only in rare cases can be used to infer function from phylogeny, and even then, they can conceal the striking diversity of functional capabilities of these microorganisms and of the respective functional enzymatic variants (e.g. *amoA,* Alves et al. 2018; *mcrA*, Borrel 2019). Conversely, they are inadequate in elucidating processes where the reaction of interest can have a broader taxonomic distribution than originally thought (e.g. Kuypers et al. 2018, Borrel 2019, Li et al. 2023). Shotgun metagenomics have traditionally been the method of choice for an unbiased picture of the natural microbiome, if errors due to sample preparation, DNA extraction, library preparation, and sequencing method are 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 involved in inorganic N and CH_4 cycling, or capture the diversity of sequence variants and properties of key genes, which could be attributed to their low relative abundances precluding their assembly into MAGs (Palmer et al., 2012, Orellana et al. 2019, Ouygang and Norton 2020, Zhao et al. 2023), in addition to the high financial and computational costs of the method. As a low-cost alternative, characterization of specific functional groups has long relied on PCR amplification of genes that encode key metabolic enzymes, as is well exemplified for 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 one PCR primer pair per functional gene is limited and potentially biased by the primers used, rendering quantitative assessments and comparisons between genes and 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

106 gene sequence variants from uncharacterized organisms (Jones et al. 2014). The geometric expansion of
107 genomic data has made it increasingly obvious that designing a single PCR primer pair to target the inherent
108 diversity of metabolic genes is highly problematic (Bonilla-Rosso et al. 2016). Therefore, especially with
109 respect to nitrogen and methane cycling, often only resource-intensive combinations of
110 metagenomics/metatranscriptomics and isotope labelling or metabolic profiling studies can contribute to
111 linking ecosystem functions to the observed community composition (Tveit et al. 2015, Orellana et al.
112 2019), which is not always a scalable solution. In most cases though, going beyond correlations and
113 investigating the mechanistic links between microbial diversity and ecosystem function has been
114 challenging (Morris et al 2020), and usually opens new questions (e.g. Ouyang et al. 2021).
115 Targeted high-throughput sequencing approaches known as probe capture, also called hybridisation capture
116 or hybridisation-based target enrichment, have been used to study complex eukaryotic samples such as
117 human exons, ancient human genomes, plant transcriptomes, and cancer marker single nucleotide
118 polymorphisms (Skoglund et al. 2014, Ichida et al. 2019, Futema et al. 2012, Gasc et al., 2016, Kamil et
119 al. 2021, Bewicke-Copley et al. 2019), but also to facilitate the in-depth study of microbial communities
120 using either 16S rRNA probes or other microbial functional gene probes (Denonfoux et al., 2013;
121 Manoharan et al. 2015, Noyes et al., 2017, Beaudry et al., 2021). Different from sequence capture, where
122 multiple probes are designed to cover a whole exon, captured metagenomics relies on large databases of
123 multiple target genes clustered at predefined similarity cut-offs, which are used to design a large number of
124 generic probes that are able to capture an extended sequence space for each of these sequence clusters
125 (Kushwaha et al., 2015, Manoharan et al., 2015).
126
127 To improve detection and characterization of critical, but low abundant microbial guilds involved in N and
128 CH_4 cycling in natural communities, we developed and tested a probe-based targeted metagenomics
129 approach for key genes involved in these processes. In addition to the enhanced resolution provided by
130 designing specific probes for each gene in order to extend the sequence space, the added value of this
131 approach consists of the ability to analyze the distinct N- and CH_4- cycling sub-pathways as a whole,
132 combining the advantages of both shotgun metagenomics and gene marker amplicon approaches. Here, we
133 present a new probe set for targeting 14 distinct functional genes involved in inorganic N cycling (i.e., N
134 fixation, nitrification, denitrification, anammox, and dissimilatory nitrate reduction to ammonium
135 (DNRA)), and genes involved in CH_4 production and consumption. We evaluated our approach using mock
136 communities comprising of DNA from organisms involved in inorganic N cycling, or CH_4 production or
137 consumption with varying gene GC% content, which is known to affect probe hybridization efficiency. As
138 a proof-of-concept for complex communities, we performed shotgun metagenomics and targeted

139 metagenomics on agricultural and wetland soil samples and compared the diversity of archaeal *amoA* genes
140 (marker for ammonia oxidizers) and *nosZ* genes (marker for nitrous oxide reduction).
141
142 **Materials and methods**
143
144
145 **Construction of the target gene databases (TDBs)**
146 Target gene databases (TDBs) were constructed, containing all identifiable variants for the following key
147 genes: the nitrogenase iron subunit (*nifH*), bacterial and archaeal ammonia monooxygenase subunit A
148 (*amoA)*, nitrite oxidoreductase beta subunit (*nxrB)*, hydrazine oxidoreductase A (*hzoA)*, formate dependent
149 nitrite reductase (*nrfA*), periplasmic nitrate reductase alpha subunit (*napA*), respiratory nitrate reductase
150 alpha subunit (*narG*), copper-containing nitrite reductase (*nirK*), cytochrome cd_1 nitrite reductase (*nirS*),
151 nitric oxide reductase B subunit (*norB*), nitrous oxide reductase (*nosZ*), particulate methane
152 monooxygenase subunit A (*pmoA*), soluble methane monooxygenase component A alpha chain (*mmoX*),
153 and methyl-coenzyme M reductase I subunit alpha (*mcrA*).
154
155 Firstly, Hidden Markov Model (HMM) models were generated in order to identify all variants of the target
156 genes from public databases (Fig. S1). These models were built based on reference sequence alignments
157 from curated databases already available for selected genes, such *amoA* (Alves et al., 2018), *pmoA* (Knief
158 et al., 2015), *nosZ* (Jones et al., 2013), *nirK, nirS, nor* (Graf et al., 2014); and the Fungene repository (Fish
159 et al., 2013). For target genes where alignments were not available, reference alignments were generated
160 from gene sequence data publicly available in the NCBI WGS-database, from the full length open-reading
161 frame of each subunit, to cover all known diversity in each gene. Structure-based searches were
162 subsequently performed over the whole Genbank nt-database with nhmmer using the generated HMMs for
163 every target gene (Wheeler & Eddy, 2013) on a local supercomputer cluster (Centre of scientific computing
164 CSC, Espoo, Finland) (July 2017). The specificity of HMM searches was evaluated by comparing the
165 results to the latest curated sequence databases of selected genes (Alves et al., 2018; Graf et al., 2022,
166 primary data was obtained directly from the respective authors), and results were further evaluated based
167 on annotation and gene structure (Supplementary Data sheet 2). Tblastx/diamond searches against the
168 *refseq* database (Altschul et al., 1990) were finally performed in order to remove duplicates or reannotate
169 falsely annotated entries due to the presence of homologous gene families in our dataset (*amoA*/ *pmoA*,
170 *nxrB*/ /nitrate reductase, *napA*/*narG* / formate reductase). This sequence selection process generated
171 ~500,000 unique sequences across all gene families. The final output comprised the target gene database
172 (TDB). The workflow can be seen in Fig S1.

173

174 **Generation of probes for target genes**

175 To aid in the design of probe libraries to capture the diversity of N and CH_4 cycling genes, the MetCap
176 bioinformatics pipeline was used as a protocol to produce probes from highly complex datasets for targeted
177 metagenomics (Kushwaha et al., 2015). The pipeline with default parameters was used for designing up to
178 six unique 50mer probes for each sequence cluster in the TDBs (clustered with an 80% identity clustering
179 threshold), resulting in a set of 263,111 unique gene-specific probes (Fig S1., Supplementary Data sheet
180 2). The probes were produced by NimbleGen SeqCap EZ (Roche NimbleGen, Inc., Madison, USA), with
181 biotin labelling to enable retrieval of the hybridized targets using streptavidin coated magnetic beads
182 (Kushwaha et al., 2015; Manoharan et al., 2015).

183

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

185 DNA samples from mock communities encoding the genes of interest were generated by mixing genomic
186 DNA from different organisms in variable proportions, in order to simulate variation in overall GC%
187 content. The mock communities dataset included: *Nitrosospira multiformis*, *Nitrososphaera viennensis*,
188 *Nitrospira defluvii*, Ca. *Kuenenia stuttgartiensis* genome fragment, *Pseudomonas aeruginosa*, *Escherichia*
189 *coli*, *Shigella sonnei*, *Cupriavidus metallidurans*, *Cupriavidus necator*, *Dyadobacter fermentans*,
190 *Pseudomonas stutzeri*, *Rhodobacter sphaeroides*, *Salinibacter ruber*, *Sulfurimonas denitrificans*,
191 *Methylosinus trichosporium* Ob3b, *Methylocella tundrae*, *Methylomicrobium buryatense*, *Methanoregula*
192 *boonei*, *Methanolacinia petrolearia*. Strain information, including genome size and the number of target
193 genes per organism are shown in Supplementary Table S1. DNA from cultured organisms was extracted
194 from 1% SD-treated cell pellets in CTAB buffer, followed by phenol:chloroform:isoamyl alcohol extraction
195 and ethanol precipitation, as described before (Jones et al., 2013). The relative abundances of the
196 microorganisms in the mock communities were multiplied with the median genome GC% contents to
197 generate a weighted GC% content of the pool of samples with the following values: 47, 50, 53, 57, 60, and
198 63 (Supplementary Table S2). Pooling ratios of microorganisms' DNA were calculated based on the
199 expected number of functional genes in the extracted DNA, the DNA concentration and genome size.
200 Relative gene abundances in mock communities were calculated and compared to the sequences generated
201 using targeted metagenomics.

202

203 **Annotation of coding sequences from the mock community dataset**

204 Proteomes from the mock communities were retrieved from NCBI, to count the number of each functional
205 genes per genome. For *Nitrospira defluvii*; NC_014355.1, genes were predicted from the genome
206 nucleotide sequence using Prodigal (v2.6.3) (Hyatt et al., 2010). Proteins were annotated using KOFAM

207 (HMM database of KEGG Orthologs). Each protein was assigned the top hitting KO. Where available,
208 cut-offs provided with KOFAMscan (Aramaki et al., 2020) were used to filter spurious results.
209
210 **Extraction of DNA and determination of chemical parameters from environmental samples**
211 Samples from different environments were collected in order to assess the effectiveness of the probe set in
212 enriching functional genes in different sample contexts. Environmental samples were collected from
213 agricultural soil in Hungary (n = 3), and from a wetland in Bellefontaine, France (Table S3). Extraction of
214 DNA from environmental samples was performed as previously described by Siljanen et al., (2019).
215 Samples were homogenized by bead-beating 0.5 g of soil at a speed of 5.5.m/s, for 30 sec with
216 phenol:chloroform:isoamyl alcohol extraction in CTAB buffer, followed by ethanol precipitation. The
217 quality of DNA extracts was assessed using a NanoDrop ND-1000 (Thermo), and DNA concentration was
218 measured with a Qubit HS dsDNA kit (Thermo). The soil chemical parameters were determined as
219 described earlier (Siljanen et al., 2019; Bagnoud et al, 2020).
220
221 **Library preparation, target enrichment of libraries and sequencing with Illumina Miseq**
222 For each sample, sequencing indexes and sequencing adapters were produced in the Centre for Genomic
223 Research (CGR), University of Liverpool, Liverpool, UK. Libraries were produced with KAPA HyperPlus
224 library preparation kits and an insert size of 630bp. For the targeted metagenomics, after quality assessment
225 with an Agilent Bioanalyzer, hybridization reactions to the NimbleGen SeqCap probe set were performed
226 according to the manufacturer's instructions at 47°C for 72 hours in CGR as previously described
227 (Manoharan et al., 2015). Sequencing was performed with Illumina MiSeq PE300 chemistry, resulting in
228 198,700-311,500 reads per sample for environmental samples and up to 2,600,000 reads for mock
229 communities.
230
231 **Targeted metagenomics read processing and mapping**
232 Quality trimming of sequencing reads was performed with the *seqtk* program in the DADA2 pipeline by
233 truncating amplicon reads to 200bp and discarding all reads with an expected error greater than 2, for all
234 amplicon and metagenomic sequencing reads.
235
236 All six possible reading frames of the nucleotide sequence reads generated by the targeted metagenomic
237 sequencing of the mock community samples were translated using transeq from the EMBOSS package
238 (v6.6.0.0). These were mapped to the genomes using DIAMOND blastp (v2.0.6.144), with a minimum
239 percentage identity of 60%. This threshold was selected as it resulted in the largest percentage of mapped
240 reads while being stringent enough to prevent spurious mappings. No minimum coverage was used to

241 account for reads that did not fully overlap with the proteins from the mock community dataset. For each
242 direction, up to four matches were retained, to account for cases where a single read spanned neighboring
243 proteins. Multiple matches in the same direction could overlap by up to 15 amino acids to account for
244 possible protein fusions or sequencing errors. Each of the (up to four) matches were used to assign KOs to
245 the reads, based on the predicted KOs of the proteins to which they mapped. If, in a given direction, a read
246 was assigned the same KO multiple times, this was collapsed into a single hit, retaining the hit with the best
247 E-value.
248
249 **Evaluation of functional annotation of reads**
250 Protein predictions for each read were performed using nhmmer from the HMMER suite (v.3.3) using the
251 in-house generated HMMs for target genes as well as KOFAM as described above. An inclusion threshold
252 of 0.0001 was used when assigning protein predictions for the reads. For a given read, using all six of the
253 translated frames, up to four protein assignments could be made, evaluated based on KO bit scores or
254 identity filtering. Reads assigned KOs for the function of interest were counted as positively identified hits.
255 Subsequently, reads were mapped to their corresponding coding sequences (CDS) from the mock
256 community organisms to determine which reads constituted "true positives" (TP) or "true negatives"(TN).
257 Reads mapped to a given CDS were assigned functions according to the above-mentioned KOFAM CDS
258 annotation as their "true function". If the HMM assignment (both in-house and from KOFAM) of a read
259 matched the functional assignment of the CDS to which it was mapped, then it was counted as a TP. The
260 TP, false positive (FP), and false negative (FN) values were summed up for all models using all methods,
261 and precision and recall were calculated. Read mapping (described above) was used to determine the total
262 number of possible TPs of the reads (Fig. S7), from which the actual number of TPs was subtracted to
263 calculate the number of FNs. Precision was calculated as the ratio of TPs to total positive detections (i.e.,
264 number of TPs and FPs). Recall was calculated as the ratio of TPs to all theoretically obtainable true
265 positives (i.e., sum of TPs and FNs).
266
267 **Shotgun metagenomics sequencing with Illumina HiSeq and analysis**
268 Shotgun metagenomic sequencing was performed on environmental DNA (n = 3) from agricultural soil and
269 wetland soil (Table S3), extracted as described above for TDB production. Illumina TruSeq library
270 preparation with an insert size of 250bp was carried out with 1 μg of DNA and sequencing was performed
271 in one Illumina HiSeq PE150 lane at the Vienna Biocenter Core Facility (VBCF, Austria) and Microsynth
272 AG, Switzerland. In this approach, the previously generated HMM profiles were used to identify the target
273 genes in the assemblies ($E < 0.001$). The predicted functions of each identified sequence were confirmed
274 by tblastx against the *refseq* database, using DIAMOND (Buchfink et al., 2015). Annotation and identity

275 information was further processed with awk in unix and in Rversion 3.5.3 to produce lists of community
276 structures for each target gene. The relative abundance of each gene was calculated against the sum of the
277 summed abundance values of all 14 genes present in the shotgun or targeted metagenomics samples (a total
278 reads and captured reads for all approaches is in Table S4). Equation of relative abundance: Relative
279 abundance gene x = (Abundance of gene x / Sum of abundances of all target genes)*100.
280
281 ***amoA* gene amplicon sequencing and processing of *amoA* reads**
282 Archaeal *amoA* gene amplicons from environmental DNA (n = 3) from the agricultural soil were generated
283 as described earlier by Siljanen et al., (2019), and sequencing was performed with Illumina MiSeq PE250,
284 in LGC, Münich, Germany. In brief, *amoA* gene amplicons were generated using 0.6 ng DNA template
285 with GoTaq DNA polymerase (Promega) and 0.5 µM primers (CamoA_19F [5'-
286 ATGGTCTGGYTWAGACG-3'] and TamoA_632R-4 [5'-GCKGCCATCCATCKRTANGTCCA-3'])
287 with a 5' Illumina sequencing adaptor [5'-CTCTTTCCCTACACGACGCTCTTCCGATCT-3']. Cycling
288 conditions were the following: 1min initial denaturation at 95°C, 30sec denaturation at 94°C, 30sec
289 annealing at 60°C and 45sec extension at 72°C, with 35 cycles. PCR reactions were performed in triplicate
290 for each sample (n = 3), and purified with the High Pure PCR product purification kit (Roche). Sequencing
291 indexes were added by LGC, where libraries were equally pooled and sequenced with Illumina MiSeq
292 PE250. The sequencing depths per sample for a comparison of amplicon, targeted and shotgun
293 metagenomics were up to 31Mb, 160 Mb and 17 Gb, respectively, for the agricultural soil.
294
295 Archaeal *amoA* reads from amplicons, targeted and shotgun metagenomics, as identified with the *TamoA*
296 in-house HMM model, were processed as described earlier (Siljanen et al., 2019; script available at:
297 https://github.com/alex-bagnoud/arctic-nitrifiers). Briefly, *amoA* reads were classified into
298 amplicon/metagenome sequence variants (A/MSVs) by DADA2 and assigned to the reference *amoA* gene
299 database (Alves et al. 2018) using a 55% identity cutoff with USEARCH8. Chimeras were filtered out with
300 UCHIME, and the remaining ASVs were further inspected with BLASTN against the NCBI nt database.
301 Finally, the resulting *amoA* gene sequences were classified with UCLUST as implemented in QIIME using
302 the reference gene database and taxonomy by Alves et al. (2018) (Altschul et al. 1990; Edgar, 2010; Edgar
303 et al., 2011; Caporaso et al., 2010; Callahan et al., 2016). The sequencing depth for the *TamoA* gene
304 amplicon library was ~27,000-58,000 PE250 reads per sample, for shotgun metagenomics 3-162 *TamoA*
305 reads out of ~22-58M PE150 reads per sample, and for targeted metagenomics 4472-8230 *TamoA* reads
306 out of ~220-311k PE300 reads per sample.
307
308 **Phylogenetic placement analysis of *nosZ* reads from shotgun and targeted metagenomics libraries.**

309

310 A reference alignment and phylogeny for *nosZ* was generated from full length *nosZ* amino acid sequences
311 obtained from genomes downloaded from the NCBI genome database (accessed October 2019). A profile
312 HMM for *nosZ* was generated from a previously published *nosZ* dataset, as described above (Graf et al,
313 2022). Reads were converted to amino acid sequences and then aligned using *hmmalign* within the HMMer
314 suite (v.3.1.2), and the homologous sections were identified using FastTree (v. 2.1.11, Price et al., 2010).
315 The reference protein alignment was used for phylogenetic reconstruction using IQ-TREE (v1.6.12) with
316 best model selection and 1000 ultrafast bootstraps. The Le-Gascuel (LG) substitution model with 10 rate
317 categories was selected after automatic model selection. For the phylogenetic placement of *nosZ* reads
318 obtained using either shotgun metagenome sequencing (n=3) or targeted metagenomics (n=3) of the
319 agricultural site, the relevant reads were pooled together and clustered with CD-HIT at 90% sequence
320 identity before translating and aligning to the reference *nosZ* alignment using *hmmalign*. The translated
321 reads were then placed in the reference phylogeny using the next generation evolutionary placement
322 algorithm (EPA-NG v0.3.8; Barbera et al., 2019), with the same model parameters as used for constructing
323 the reference phylogeny. The placement positions with the highest likelihood weight ratio were plotted on
324 the reference phylogeny in R using the 'ggtree' package (v3.4.2, Yu et al., 2016).
325

326 **Data availability**

327 The shotgun metagenomic, targeted metagenomic and *amoA* amplicon sequencing data have been deposited
328 in NCBI SRA under the Bioproject numbers PRJNA898102 and PRJNA488558.
329

330 **Results**

331

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

333 The targeted metagenomic approach was evaluated using mock microbial communities comprising
334 predefined mixtures of genomic DNA from 14 bacterial and three archaeal strains, together containing all
335 14 functional genes targeted by the probes and mixed in different proportions to produce six different
336 median G+C contents ranging from 47% to 62% (Table S1). The overall relative abundance of target genes
337 in the captured metagenomes across all GC% categories was strongly positively correlated with their
338 abundance in the original mock communities (r = 0.78 (P < 0.0001) (Fig. 1). Comparisons between the
339 relative abundance of individual genes in the mock communities as compared to captured metagenomes
340 revealed nearly identical values (t = 1.8518e-10, df = 174.83, P > 0.05), with the exception of the *nirK* gene
341 and the *hzoA* gene (Fig. S1a,b). Pearson correlation coefficients across all GC% categories varied between
342 0.76 and 0.93, although correlations were stronger within higher GC% categories (57%-63%; Rpearson =

343 0.80-0.93) than within lower GC% categories (47%-53%; Rpearson = 0.76-0.79) (Fig. S2). The probe
344 capture produced 61-72% of targeted sequences from the total sequenced library (Table S4).
345
346 **Precision of gene identification in mock communities**
347 Precision for the in-house HMMs ranged from 74.1% (*nifH*) to 100.0% (*mcrA*) (Fig. S6). The average
348 precision of the models was 93.4%, with a median of 99.8% (excluding *amoA* and *pmoA*). The recall ranged
349 from 42.8% (*nxrB*) to 100.0% (*mcrA*). The average recall was 81.9%, with a median recall of 82.8%.
350 Because KEGG only provides a single model covering the homologous genes *pmoA* and *amoA*, it was not
351 possible to accurately calculate the precision and recall for these genes individually. A precision estimate
352 could still be determined for the two genes, where a "true positive" is defined as a hit for a read identified
353 by the in-house models as *amoA* or *pmoA* mapped to a CDS identified by the HMM for (*a*/*p*)*moA* genes
354 provided by KEGG (K10944). In this case, the precision values for the *amoA* and *pmoA* genes were
355 estimated to be 96.9%, and 100.0%. respectively It is important to note that these estimates do not exclude
356 the possibility that the in-house models misclassified reads belonging to these homologous protein families.
357
358 Precision and recall values were similar using KOFAMs instead of in-house HMMs. Note, however, that
359 the results for *amoA* and *pmoA* genes were combined due to a single model for the associated KO for both
360 genes, as mentioned above. Precision ranged from 76.5% (*a*/*pmoA*) to 100.0% (*mcrA*), while recall varied
361 from 34.3% (*nifH*) to 100.0% (*mcrA* and *nxrB*). The average and median precision values were 94.5% and
362 97.7%, while the average and median recall values were 93.39% and 99.03%, respectively. The precision
363 of KOs was lower for *napA* genes than for other genes when compared with custom DNA HMMs. This
364 could be accounted for by the presence of formate dehydrogenase, a known homolog of *napA*. Formate
365 dehydrogenase (K00123) was indeed identified among the genomes used for the mock community, so it is
366 possible that the custom HMM was misclassifying reads originating from this known ortholog to *napA*
367 (Cerqueira et al., 2015).
368
369 Of note is the relative performance of the KEGG models. The true gene assignments were based on the
370 KEGG models, so while these analyses demonstrate the ability of the models to predict putative function
371 even from small DNA fragments (*i.e.,* metagenomic reads), a relatively high precision and recall would be
372 expected. In other words, when the models used to predict the reads are the same as those used to assign
373 the true function based on the full aligned protein length, it is expected that the performance should be
374 relatively high compared to other using other models for predictions. In this sense, the comparison of the
375 in-house models to the "true positive" KEGG assignments is probably more reflective of a real-world
376 analysis.

377 Overall, the in-house HMM searches performed well when comparing reads mapped to the genomes and
378 assigned KEGG Orthologs based on the fully matched protein sequences, indicating that the HMMs can
379 accurately classify the reads in the mock community, and are sensitive enough to detect distantly related
380 genes or novel variants of a gene. As the results obtained with the custom HMM models generally yielded
381 higher precision than those based on KEGG Orthologs, they were used for the final analyses. It is
382 noteworthy that reads identified as false positives may in fact not be incorrect assignments but could simply
383 represent novel variants of the genes of interest, not captured by KEGG Orthologs.
384
385 **Higher number of reads for all target functional genes but similar relative abundance obtained with**
386 **targeted compared to shotgun metagenomics from complex environmental samples**
387 We investigated the efficiency of the targeted metagenomics approach in natural complex ecosystems by
388 directly comparing this approach with shotgun untargeted metagenomics generated with a nearly 100 times
389 higher amount of sequencing (~66 Mb and ~6.4 Gb per sample, respectively) from wetland soils in
390 Bellefontaine, France, and from an agricultural field in Hungary (Table S3). These soils represent distinct
391 ecosystems with average physicochemical conditions, such as water content and oxygen concentration, and
392 N availability that favors different distributions of functional genes (Jones et al. 2014). Despite the large
393 difference in sequencing depth, targeted metagenomics generated a much larger set of target functional
394 genes (Fig. 2), and up to 60 times as many identified gene sequences (Fig. 3), compared to shotgun
395 metagenomics. The relative abundances of functional genes in captured metagenomes correlated
396 significantly with that in shotgun metagenomes from both the agricultural field ($R^{\text{Pearson}} = 0.96$, d.f.=14, P
397 < 0.00001) and the wetland ($R^{\text{Pearson}} = 0.70$, d.f,=14, $P < 0.01$). The relative target gene abundances were
398 also not different between the two approaches ($P > 0.90$) (Fig. 2). However, reduced relative abundance
399 was detected with targeted metagenomics compared to shotgun metagenomics in the agricultural site for
400 *nxrB*, *nrfA*, *narG* and *nosZ* genes and the wetland site for *nxrB*, but a higher relative abundance with targeted
401 metagenomics was obtained for *TamoA*, *napA* and *mmoX* genes in the agricultural site and for *nifH* and
402 *nirS* genes in the wetland site (Fig. 2.). The targeted approach generated a much higher number of reads for
403 all genes in both sites ($P > 0.0001$) (Fig. 3. Table S4).
404
405 The different relative abundances of functional genes in the two ecosystems reflected known differences in
406 their environmental conditions: (i) N_2 fixation genes were expectedly higher in the unfertilized wetland;
407 (ii) Genes involved in ammonia oxidation and nitric and nitrous oxide reduction were higher in the
408 agricultural soil, which is likely to be more oxic and also subject to N fertilization resulting in higher N_2O
409 production rates ; (iii) Genes involved in methanogenesis were more abundant in the wetland, where anoxic
410 conditions and methane production typically occur.

411

412 **Comparing diversity recovered for specific target genes from amplicon, targeted and shotgun**
413 **metagenomics from complex environmental samples**

414 In order to compare the diversity of specific functional genes detected by targeted metagenomics, shotgun
415 metagenomics and PCR amplicon sequencing we focused on the archaeal *amoA* gene as a test case, as it is
416 the second most sequenced gene in NCBI indicative of its ecological importance, and for which there is a
417 comprehensive reference database for classification available (Alves et al., 2018). Moreover, archaeal
418 ammonia oxidizers (AOA) typically outnumber their bacterial counterparts in most environments
419 (Leininger et al. 2006), and even though they can be low abundant, their diversity is striking, and underlines
420 their distinct metabolic capabilities (Alves et al. 2018; Huang et al. 2021). Additionally, we compared the
421 detection of *nosZ* genes between the targeted and shotgun metagenomics approaches in the agricultural soil,
422 also based on an existing reference database (Graf et al., 2022). The following datasets were compared: (i)
423 targeted and shotgun metagenomes, and PCR amplicons of the archaeal *amoA* gene in the agricultural soil
424 (Fig. 4); (ii) *nosZ* gene sequences identified by targeted or shotgun metagenomics in the agricultural soil
425 (Fig. S4).

426

427 Comparison from applying all three methods in agricultural soil samples showed that targeted
428 metagenomics detected a substantially higher diversity of archaeal *amoA* genes than PCR amplicon
429 sequencing, while, as expected, shotgun metagenomics detected a lower diversity than either of the other
430 two methods (Fig. 4A). Specifically, targeted metagenomics detected 71-83 distinct taxonomic *amoA* gene
431 MSV bins, whereas PCR amplicon sequencing and shotgun metagenomics detected 65-68 ASVs and only
432 two MSV bins, respectively. Rarefaction analysis also indicated that targeted metagenomics retrieved a
433 higher alpha diversity of archaeal *amoA* genes than PCR amplicon sequencing (Fig. 4B,C).

434

435 Similarly, targeted metagenomes yielded a higher number of taxonomic bins for *nosZ* genes from the
436 agricultural soil (~6,800 bins) than shotgun metagenomics (~1,300 bins) at a 90% identity cutoff, showing
437 that the first approach detected a greater gene diversity with a much lower sequencing depth (Fig. S4). Gene
438 sequence reads from both methods mapped to similar lineages in a comprehensive reference phylogeny of
439 *nosZ* genes (Graf et al., 2022). However, a greater proportion of reads from the targeted metagenomics
440 approach were assigned to deeper nodes of the phylogeny, particularly within the less studied clade II,
441 suggesting that this method also captured novel *nosZ* variants not represented in the reference phylogeny.

442

443 **Discussion**

444 In this study, we report the development and evaluation of a new probe set for targeted metagenomics for
445 14 key genes involved in the microbial cycling of inorganic N and CH_4. This included the compilation of
446 databases containing all variants of the target genes from public databases, development of an extensive set
447 of probes targeting 14 well-established marker genes and a bioinformatics pipeline to process and annotate
448 the sequence data. This approach has been successfully used earlier to identify genes involved in microbial
449 C metabolism in soil (Manoharan et al., 2015, Manoharan et al., 2017).
450
451 In validation experiments with mock communities comprising 18 distinct organisms encoding the target
452 genes, we showed that this method successfully reproduced the expected relative abundances of all target
453 functional genes, with two exceptions, i.e. *nirK* and *hzoA*. While 10.3% of the generated probes are
454 detecting nitrifying *nirK* genes, these were detected at a much lower relative abundance than present in the
455 original mock community DNA. Instead, the relative abundance profile of the detected *nirK* match the
456 amount of the *nirK* present in heterotrophic organisms in the mock community (Fig. S1). From this we
457 conclude that hybridization of the *nirK* probes on the nitrifying organisms was unsuccessful with the current
458 probe database. However, the sequencing technique used can also influence the outcome due to known GC
459 biases with certain sequencing platforms, such as the false low coverage observed in genomic regions with
460 non-optimal GC% content (50-60%) in MiSeq sequencing (Browne et al. 2020) therefore biasing the *nirK*
461 relative abundance. We indeed observed some variation in probe specificity among mock communities with
462 different genomic GC% content, with high GC% communities having a generally higher target gene
463 detection efficiency. Such variation was likely due to stronger probe hybridization to sequences with higher
464 GC% content, given the higher binding energy between guanine and cytosine than between adenine and
465 thymine. Regarding *hzoA* the gene was detected in environmental samples, however, in the mock-
466 community a PCR-generated genome fragment from Ca. *Kuenenia stuttgartiensis* was used to supply this
467 gene in the mock community sample (other genes in the mock community gDNA of organisms was
468 provided), therefore it is possible that degradation of the fragment prevented its detection during the
469 hybridization reaction.
470
471 Targeted metagenomics also closely reproduced the relative abundances of target genes obtained with
472 shotgun metagenomics in two distinct natural microbial communities from soils, showing that the probe
473 capture step does not introduce significant relative quantitative bias ($P > 0.90$, Fig. 2.). Reassuringly, this
474 was also true for *nirK*, suggesting that perhaps certain compositional characteristics of this gene family
475 render read mapping problematic (Fig. S2). The GC% content of the nitrifier *nirK* genes is smaller (mean
476 GC% is about 53%) than that of the heterotrophic *nirK* genes (mean GC% >60) and as we established in
477 the mock community experiment, higher GC% gene content leads to higher detection efficiencies. This

478 effect is also visible for *nosZ* , where the *Dyanobacter fermentas nosZ* variant with a median GC% of 52.9
479 (Table S1) is detected more efficiently than the *Sulfurimonas denitrificans nosZ* variant with a GC% of
480 32.7%. In these cases of target genes with low GC% care needs to be taken during probe design to ensure
481 that probes are not placed on well conserved regions, where higher GC% probes (matching to non-target
482 homologous variants with higher GC%, as exemplified in the *nirK* case) might compete with lower GC%
483 probes. The lower relative abundance values obtained with targeted metagenomics for *nxrB*, *nrfA*, *narG*
484 and *nosZ* at the agricultural site (Fig. 2) can be due to the fact, that some of the other genes are being
485 detected more efficiently with the targeted approach, than with the shotgun metagenomics (Fig. 3.).
486 However, as we observed during the benchmarking run using mock communities, the relative abundance
487 with targeted metagenomics was strongly positively correlated with their actual abundance for all GC%
488 categories (Fig. S3). This suggests that we can obtain a more accurate representation of the relative
489 abundance values of the communities in question with targeted metagenomics, compared to a shotgun
490 approach.
491
492 The archaeal *amo*A gene (*TamoA*) and the *nosZ* gene were used as case studies to investigate the differences
493 in the detected diversity of single functional guilds that are known to be highly diverse (Alves et al. 2018,
494 Graf et al. 2022). Targeted metagenomics with a sequencing depth of 117-160 Mb per sample generated at
495 least 2200-4200 times more *TamoA* gene sequences than shotgun metagenomics with 7-17 Gb per sample
496 (where only two quality-controlled *amoA* sequence reads were detected (Fig 4)). Conversely, targeted
497 metagenomics generated five times less *TamoA* reads (Fig. 4B) than PCR amplicon sequencing, with a
498 sequencing depth of 15-31 Mb per sample. The rarefaction curve of *amoA* targeted metagenomics also
499 reaches a plateau much faster and with a lower number sequences than amplicon sequencing (Fig. 4B).
500 Nevertheless, the targeted metagenomics approach detected a much higher gene sequence diversity for both
501 *TamoA* (Fig 4A, 4B) and *nosZ* (Fig. S4) than the two other methods in the agricultural soil samples. For
502 *TamoA*, the targeted approach detected 41 times higher alpha-diversity than shotgun metagenomics, while
503 5 times higher diversity of *nosZ* genes was recovered, importantly covering genes affiliated with the less
504 characterized Clade II. In turn, the targeted approach detected 1.2 times higher diversity of archaeal *amoA*
505 genes than PCR amplicon sequencing in agricultural soils, respectively. At first glance, PCR amplicon
506 sequencing seems to have resulted in 18 more taxonomic bins than targeted metagenomics. However, these
507 sequences were generated by a 35-cycle PCR, known to produce chimeras. Examples of such errors have
508 been shown even for shorter amplification cycles (Kozich et al., 2013). Problems due to chimeric sequences
509 in public databases from amplicon studies are known and were encountered during the compilation of the
510 thaumarchaeal *amoA* reference database (Alves et al., 2018). While we used standard pipelines for detection
511 of chimeric sequences during data processing of amplicon results in this study, we are aware that we might

512 not have been able to detect them all, resulting putatively in the higher number of sequences (post-
513 processing) in the amplicon dataset. Conversely, since no amplification step is performed in targeted
514 metagenomics, the issue of chimeric sequences is avoided, revealing an important quality advantage for
515 this approach.
516
517 It should be noted that the efficiency of this approach largely depends on the coverage and quality of the
518 database used to generate the probes. However, the database is not constrained to full-length gene sequences
519 from a limited diversity of complete genomes and long metagenomic scaffolds available, and the
520 probe/target diversity can be greatly extended through inclusion of gene fragments generated by PCR or
521 from short metagenomic contigs. Despite its high target precision, this method may nevertheless capture
522 non-target sequences, especially for orthologous gene families that have evolved different functions, as is
523 the case for *napA* genes and formate dehydrogenases. Such cases emphasize the importance of thorough
524 data annotation and filtering procedures. With the extended database used to generate probes in this study,
525 we managed to detect a higher diversity for the case study of *TamoA* genes than the shotgun sequencing
526 approach, showcasing the advantage of "casting a wider net". Similar results were obtained in other targeted
527 metagenomics applications, such as for detection of diverse resistome-virulome elements (Noyes et al.
528 2017) or for improved taxonomic microbial community characterization via 16S rRNA enrichment
529 (Beaudry et al. 2021).
530
531 Despite the technological advances and decline in cost of high-throughput sequencing, shotgun
532 metagenomics remain impractical and prohibitively expensive to capture the diversity of specific, low
533 abundant functional groups in complex environments, such as soils and sediments, especially in
534 longitudinal studies. The cost of generating the probes is about 20-50€ per sample depending on the probe
535 manufacturer. If the goal is to have a focused, comprehensive view of the diversity of functional guilds
536 involved in inorganic nitrogen or methane cycling in a certain ecosystem, then targeted metagenomics can
537 circumvent the high cost and overabundance of data generated by shotgun metagenomics, as well as provide
538 more quantitatively more information on the diversity of the genes of interest . In turn, PCR-based assays
539 are limited by the biases associated with any set of primers targeting a large diversity of sequences, assay
540 efficiencies which preclude detection of highly divergent new sequence diversity as well as known issues
541 with generation of chimeric sequences during amplification. The targeted metagenomics approach
542 circumvents these limitations by using multiple short probes targeting different regions of each gene, which
543 largely increases the likelihood of capturing sequence variants that elude PCR amplification. Thus, this
544 approach has not only the potential to capture rare and novel gene diversity in complex environments, but
545 also to identify cryptic microorganisms in low-biomass samples or involved in suggested CH_4 metabolisms,

such as in the tree phyllosphere (Putkinen et al., 2021) and nitrogen cycling in coral holobiont (Glaze et al., 2021). Moreover, targeted metagenomics can also overcome issues associated with running and comparing multiple independent PCR assays when investigating several distinct targets. In that sense, this approach effectively represents a PCR-independent, multiplex approach to characterize simultaneously and in-depth the distributions of a broad range of functional genes, providing a holistic view of the status of the nitrogen and methane cycles in the studied ecosystems. This is especially advantageous when combined with functional studies, such as determination of N-transformation rates and *in situ* fluxes, as showcased by a study of N_2O emissions in thawing Yedoma permafrost sites over time (Maruschak et al. 2021). In this study, the application of targeted metagenomics with the N-cycling probe dataset presented here revealed that changes in the N-cycling microbial community composition were responsible for an increase in N_2O emissions in revegetated Yedoma soils, which had undergone thawing a decade prior.

Conclusively, the targeted metagenomics approach developed here provides an efficient and cost-effective strategy for studying microbial functional guilds that typically represent small fractions of natural microbiomes, and whose diversity is generally underestimated and highly underrepresented in metagenomic datasets. This approach also circumvents the limitations and biases associated with PCR-based methods and has higher potential to capture rare or novel functional gene diversity.

Acknowledgements

H.S. was financially supported by the Atmosphere and Climate Competence Center (ACCC) Academy of Finland project number 337550, and Nitrobiome project 342362, in addition H.S. was supported by Saastamoinen foundation, FEMS Society and Niemi foundation. L.M., A.B., and C.S. were funded by ERC Advanced Grant TACKLE (695192), and R.J.E.A. was funded by project P25369 of the Austrian Science Fund (FWF). A.S.H and F.L.S were funded by ERC Starting Grant EvolPhisiol (grant agreement 803768). S.H and C.M.J were supported by the Swedish University of Agricultural Sciences (senior career grant 2019-2024). We thank Lea Wittorf for providing DNA extracts from isolates for the mock libraries.

Authors contribution statement

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

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

References:

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

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

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

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

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

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

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

Beaudry, M.S. et al. (2021) Improved Microbial Community Characterization of 16S rRNA via Metagenome Hybridization Capture Enrichment. *Front. Microbiol.* 12. Available at: https://doi.org/10.3389/fmicb.2021.644662.

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., Stamatakis, A. 2019. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.* 68: 365-369.

Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., Korfi, K., Wang, J. 2019. Applications and analysis of targeted genomic sequencing in cancer studies. *Comput. Struct. Biotechnol. J.* 17:1348-1359. doi: 10.1016/j.csbj.2019.10.004.

Buchfink, B., Xie, C., Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. methods.* 12: 59.

Bodelier, P.L.E., Laanbroek, H.J. 2004. Nitrogen as a regulatory factor of methane oxidation in soils and sediments. *FEMS Microbiol. Ecol.* 47: 265–277, https://doi.org/10.1016/S0168-6496(03)00304-0

Bonilla-Rosso, G., Wittorf, L., Jones, C. *et al.* 2016. Design and evaluation of primers targeting genes encoding NO-forming nitrite reductases: implications for ecological inference of denitrifying communities. *Sci. Rep.* 6: 39208 https://doi.org/10.1038/srep39208

Borrel, G., Adam, P.S., McKay, L.J. *et al.* Wide diversity of methane and short-chain alkane metabolisms in uncultured archaea. *Nat. Microbiol.* **4**, 603–613 (2019). https://doi.org/10.1038/s41564-019-0363-3

Browne, P.D. Nielsen. T.K., Kot, W., Aggerholm, A., Thomas, M., Gilbert, P., Puetz, L., Rasmussen, M., Zervas, A., Hansen L.H. 2020. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience* 9: giaa008, https://doi.org/10.1093/gigascience/giaa008

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

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.

Graf, D.R., Jones, C.M., Hallin, S. 2014. Intergenomic comparisons highlight modularity of the denitrification pathway and underpin the importance of community structure for N_2O emissions. *PloS one* 9: e114118.

Graf, D.R.H., Jones, C.M., Zhao, M., Hallin, S. 2022. Assembly of root-associated N2O-reducing communities of annual crops is governed by selection for nosZ clade I over clade II. *FEMS Microbiol. Ecol.* 98:1-11

Gruber, N., Galloway, J. An Earth-system perspective of the global nitrogen cycle. 2008. *Nature* 451, 293–296. https://doi.org/10.1038/nature06592

Glaze, T.D., Erler, D.V., Siljanen, H.M.P. 2021. Microbially facilitated nitrogen cycling in tropical corals. *ISME J.* 16: 68–77 doi: 10.1038/s41396-021-01038-1.

Eddy, S.R.. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755–763, https://doi.org/10.1093/bioinformatics/14.9.755

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *BMC Bioinformatics* 26: 2460–2461

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. *BMC Bioinformatics* 27: 2194–2200.

Fish, J. A., Chai, B., Wang, Q., Sun, Y., Brown, C. T., Tiedje, J. M., Cole, J. R. 2013. FunGene: the Functional Gene Pipeline and Repository. *Front. Microbiol.* 4: 291.

720
721 Hallin S, Lindgren PE. 1999. PCR detection of genes encoding nitrite reductase in denitrifying bacteria.
722 *Appl Environ Microbiol.* 65:1652-7. doi: 10.1128/AEM.65.4.1652-1657.1999.
723
724 Hoang, D.T., Chernomor, O., Von Haeseler, A. et al. 2018. UFBoot2: improving the ultrafast bootstrap
725 approximation. *Mol Biol Evol* 35:518
726
727 Huang, L., Chakrabarti, S., Cooper, J., Perez,A., John, S.M., Daroub, S.H., Martens-Habbena, W. 2021.
728 Ammonia-oxidizing archaea are integral to nitrogen cycling in a highly fertile agricultural soil. *ISME*
729 *Comm.* 1:19 ; https://doi.org/10.1038/s43705-021-00020-4
730
731 Håvelsrud, O.E., Haverkamp, T.H., Kristensen, T. *et al.* A metagenomic study of methanotrophic
732 microorganisms in Coal Oil Point seep sediments. *BMC Microbiol* **11**, 221 (2011).
733 https://doi.org/10.1186/1471-2180-11-221
734
735 Hyatt, D., Chen, G.L., LoCascio, R.F. *et al.* 2010. Prodigal: prokaryotic gene recognition and translation
736 initiation site identification. *BMC Bioinformatics* 11: 119 https://doi.org/10.1186/1471-2105-11-119
737
738 Ichida, H., Morita, R., Shirakawa, Y., Hayashi, Y., Abe, T. 2019. Targeted exome sequencing of unselected
739 heavy-ion beam-irradiated populations reveals less-biased mutation characteristics in the rice genome.
740 *Plant J.* 98:301-314. Doi: 10.1111/tpj.14213.
741
742 Jones, C.M., Graft, D.R.H., Bru, D., Philippot, L., Hallin, S. 2013. The unaccounted yet abundant nitrous
743 oxide-reducing microbial community: a potential nitrous oxide sink. *ISME J* 7: 417-426.
744
745 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K. et al. 2017. ModelFinder: fast model selection for accurate
746 phylogenetic estimates. *Nat. Methods* 14:587–9.
747
748 Kamil, G., Yoon, J.Y., Yoo, S. et al. 2021. Clinical relevance of targeted exome sequencing in patients with
749 rare syndromic short stature. *Orphanet J Rare Dis* 16: 297 https://doi.org/10.1186/s13023-021-01937-8
750
751 Katoh, K., Rozewicki, J., Yamada, K.D. 2019. MAFFT online service: multiple sequence alignment,
752 interactive sequence choice and visualization. *Brief Bioinform* 20: 1160–66.
753

Knief, C. 2015. Diversity and Habitat Preferences of Cultivated and Uncultivated Aerobic Methanotrophic Bacteria Evaluated Based on *pmoA* as Molecular Marker. *Front Microbiol* 6:1346. Doi: 10.3389/fmicb.2015.01346

Kushwaha, S.K., Manoharan, L., Meerupati, T., Hedlund, K., & Ahrén, D. 2015. MetCap: a bioinformatics probe design pipeline for large-scale targeted metagenomics. *BMC bioinformatics* 16: 65.

Kuypers, M., Marchant, H. & Kartal, B. 2018. The microbial nitrogen-cycling network. *Nat Rev Microbiol* 16: 263–276. https://doi.org/10.1038/nrmicro.2018.9

Kolb S, Knief C, Stubner S, Conrad R. 2003. Quantitative detection of methanotrophs in soil by novel pmoA-targeted real-time PCR assays. *Appl Environ Microbiol.* 69:2423-9. Doi: 10.1128/AEM.69.5.2423-2429.2003.

Lan, X., K.W. Thoning, and E.J. Dlugokencky: Trends in globally-averaged CH4, N2O, and SF6 determined from NOAA Global Monitoring Laboratory measurements. Version 2023-09, https://doi.org/10.15138/P8XG-AA10

Letunic, I, Bork, P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 47:W256–9

Li, J., Liu, T., McIlroy, S.J. et al. Phylogenetic and metabolic diversity of microbial communities performing anaerobic ammonium and methane oxidations under different nitrogen loadings. *ISME Commun.* 3, 39 (2023). https://doi.org/10.1038/s43705-023-00246-4

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.

Nnadozie, C.F., Lin, J. and Govinden, R. (2015) 'Selective isolation of bacteria for metagenomic analysis: Impact of membrane characteristics on bacterial filterability'. *Biotechnol Prog.* 31: 853–866. https://doi.org/10.1002/btpr.2109.

Noyes, N.R. et al. (2017) Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome* 5. 142. https://doi.org/10.1186/s40168-017-0361-8.

R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Romdhane, S., Spor, A., Aubert, J. *et al.* 2021. Unraveling negative biotic interactions determining soil microbial community assembly and functioning. *ISME J.* https://doi.org/10.1038/s41396-021-01076-9

822 Palmer, K., Biasi, C., Horn, M.A. 2012. Contrasting denitrifier communities relate to contrasting N_2O
823 emissions patterns from acidic peat soils in arctic tundra. *ISME J.* 6:1058-1077.
824
825 Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of
826 microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25: 1043-1055.
827
828 Pan Y, Bodrossy L, Frenzel P, Hestnes AG, Krause S, Lüke C, Meima-Franke M, Siljanen H, Svenning
829 MM, Bodelier PL. Impacts of inter- and intralaboratory variations on the reproducibility of microbial
830 community analyses. *Appl Environ Microbiol.* 2010 Nov;76(22):7451-8. doi: 10.1128/AEM.01595-10
831
832 Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large
833 alignments. *PLoS ONE* 5: e9490.
834
835 Prosser, JI, Hink, L, Gubry-Rangin, C, Nicol, GW. Nitrous oxide production by ammonia oxidizers:
836 Physiological diversity, niche differentiation and potential mitigation strategies. Glob Change Biol. 2020;
837 26: 103– 118. https://doi.org/10.1111/gcb.14877
838
839 Putkinen, A., Siljanen, H.M.P., Laihonen, A., Paasisalo, I., Porkka, K., Tiirola, M., Haikarainen, I.,
840 Tenhovirta, S. and Pihlatie, M. 2021. New insight to the role of microbes in the methane exchange in trees:
841 evidence from metagenomic sequencing. *New Phytol* 231: 524-536. https://doi.org/10.1111/nph.17365
842
843 Skoglund, P., et al. 2014. Separating endogenous ancient DNA from modern day contamination in a
844 Siberian Neandertal. *Proc Natl Acad Sci USA* 111: 2229–2234 .

845 Siljanen, H.M.P., Alves, R.J.E., Ronkainen, J.G., Lamprecht, R.E., Bhattarai, H.R., Bagnoud, A.,
846 Marushchak, M.E., Martikainen, P.J. Schleper, C., Biasi, C. 2019. Archaeal nitrification is a key driver of
847 high nitrous oxide emissions from arctic peatlands. *Soil Biol Biochem* 137: 107539
848
849 Siljanen, H.M.P., Saari, A., Krause, S., Lensu, A., Abell, G.C.J., Bodrossy, L., Bodelier, P.L.E.,
850 Martikainen, P.J. 2011. Hydrology is reflected in the functioning and community composition of
851 methanotrophs in the littoral wetland of a boreal lake. *FEMS Microbiol Ecol*. 75:430–445,
852 https://doi.org/10.1111/j.1574-6941.2010.01015.x
853

854 Sinha, R. et al. 2017. Assessment of variation in microbial community amplicon sequencing by the
855 Microbiome Quality Control (MBQC) project consortium. *Nature Biot* 35: 1077–1086.
856 https://doi.org/10.1038/nbt.3981.
857
858 Stamatakis, A. 2014 RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large
859 phylogenies. *Bioinformatics* 30:1312–13.
860
861 Stein, L.Y., Klotz, M.G. 2016. The nitrogen cycle. *Curr Biol.* 26: R94-R98,
862 https://doi.org/10.1016/j.cub.2015.12.021.
863
864 Sui, H.Y. et al. (2020) Impact of DNA Extraction Method on Variation in Human and Built Environment
865 Microbial Community and Functional Profiles Assessed by Shotgun Metagenomics Sequencing. *Front*
866 *Microb.* 11: . Available at: https://doi.org/10.3389/fmicb.2020.00953
867
868 Trifinopoulos, J., Nguyen, L.T., von Haeseler, A., Minh, B.Q. 2016. W-IQ-TREE: a fast online
869 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44(W1):W232-5. doi:
870 10.1093/nar/gkw256.
871
872 Thomson R.A.J., Giannopoulos, G., Pretty, J., Baggs, E.M., David J. 2012. Biological sources and sinks
873 of nitrous oxide and strategies to mitigate emissions. *Phil. Trans. R. Soc. B* 2012 367, doi:
874 10.1098/rstb.2011.0415, published 26 March 2012
875
876 Throbäck, I.N., Enwall, K., Jarvis, Å. and Hallin, S. 2004. Reassessing PCR primers targeting nirS, nirK
877 and nosZ genes for community surveys of denitrifying bacteria with DGGE. *FEMS Microbiol Ecol* 49: 401-
878 417. https://doi.org/10.1016/j.femsec.2004.04.011
879
880 Orellana, L.H., Hatt, J.K., Iyer, R. *et al.* Comparing DNA, RNA and protein levels for measuring microbial
881 dynamics in soil microcosms amended with nitrogen fertilizer. *Sci Rep.* **9**, 17630 (2019).
882 https://doi.org/10.1038/s41598-019-53679-0
883
884 Ouyang Y., Norton J.M. 2020. Nitrite Oxidizer Activity and Community Are More Responsive Than Their
885 Abundance to Ammonium-Based Fertilizer in an Agricultural Soil. *Front Microbiol.* 11:
886 DOI=10.3389/fmicb.2020.01736
887

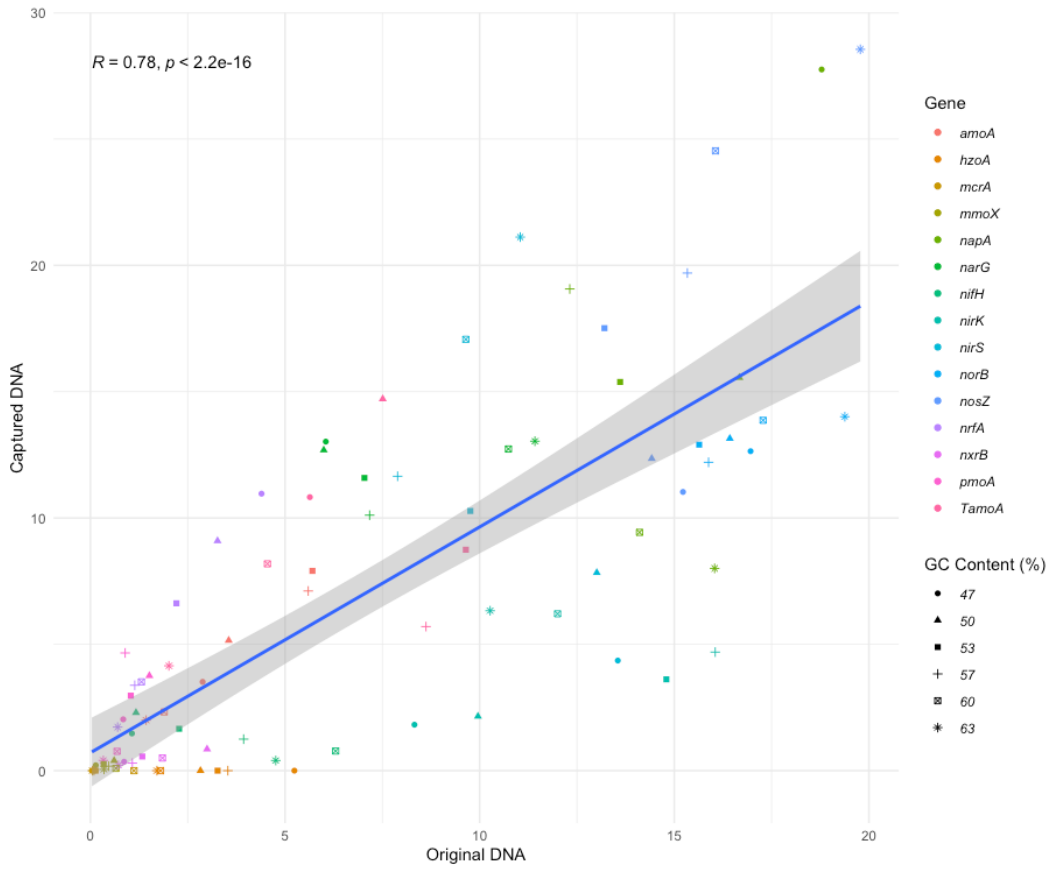
Wheeler, T.J., Eddy, S.R. 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29: 2487-2489.

Yu, G.C., Smith, D.K., Zhu, H.C., Guan, Y., Lam, T.T.Y. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecol Evol.* 8: 28-36.

Zaehle S. 2013.Terrestrial nitrogen–carbon cycle interactions at the global scale. Phil. Trans. R. Soc. B3682013012520130125 http://doi.org/10.1098/rstb.2013.0125

Zhao, J., Huang, L., Chakrabarti, S. et al. 2023. Nitrogen and phosphorous acquisition strategies drive coexistence patterns among archaeal lineages in soil. *ISME J.* 17, 1839–1850 https://doi.org/10.1038/s41396-023-01493-y

902 **Figures:**



903

904 Fig. 1. Comparison between the original composition of the mock community and the relative abundance
905 for each functional gene detected using probe captured metagenomics for all GC% combinations combined.
906 The relative abundance in the mock community was calculated based on the amount of each functional
907 gene and organism as well as genome size. Color of the symbols depict gene and shape the category of GC
908 content (%).

909

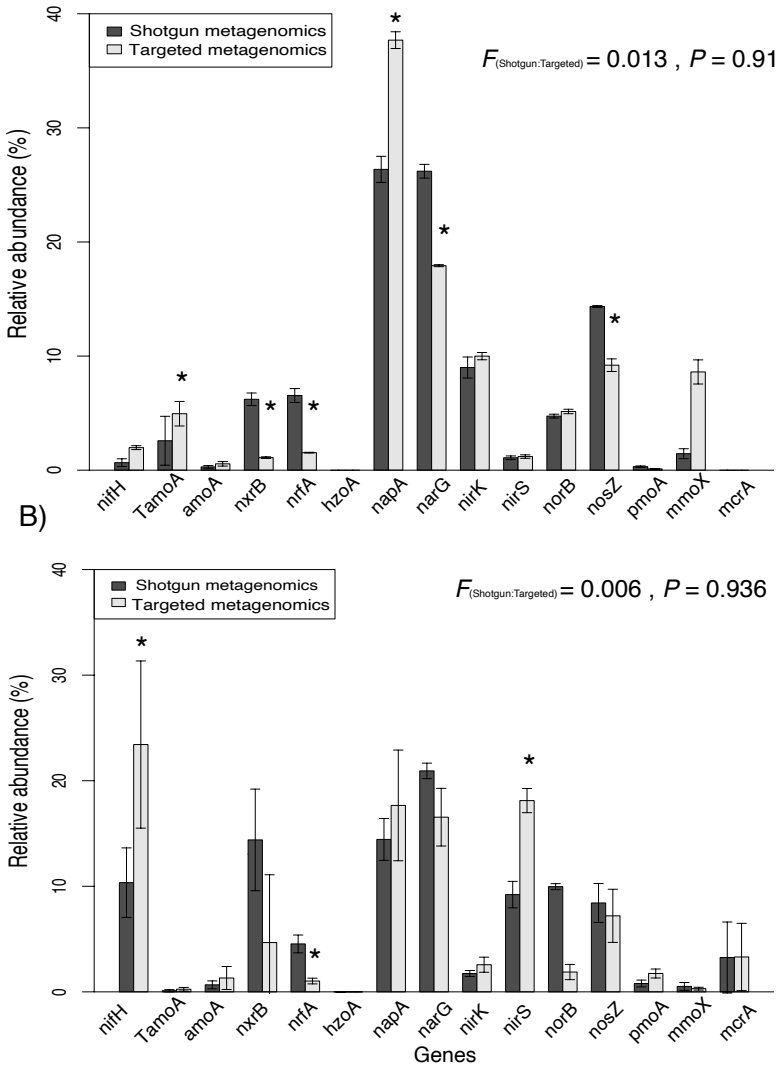


Fig. 2. Relative abundance of functional genes involved in inorganic nitrogen and methane cycling obtained by targeted metagenomics and shotgun metagenomics from A) agricultural soil; and B) wetland soil. Relative abundance was retrieved with HMMER searches of each functional gene. Statistically significant differences between shotgun and targeted relative abundances according to pairwise comparisons with ANOVA are shown with asterisk ($P < 0.05$, n = 3).

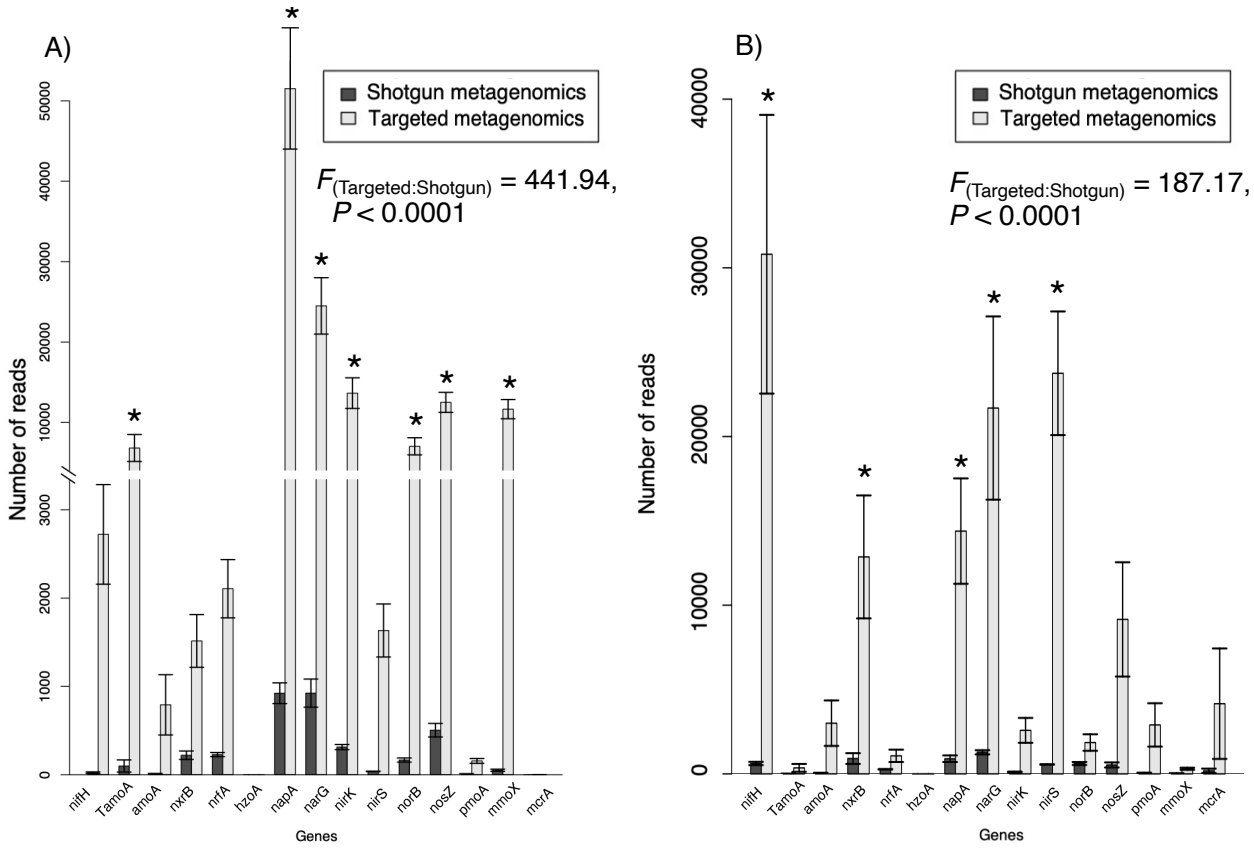


Fig. 3. Total reads of functional genes involved in inorganic nitrogen and methane cycling obtained by targeted metagenomics and shotgun metagenomics from A) agricultural soil, B) wetland soil B). Number of reads was retrieved with HMMER searches of each functional gene. Statistically significant differences between shotgun and targeted metagenomic numbers of reads according to pairwise comparisons with ANOVA are shown with asterisk ($P < 0.05$, n=3).

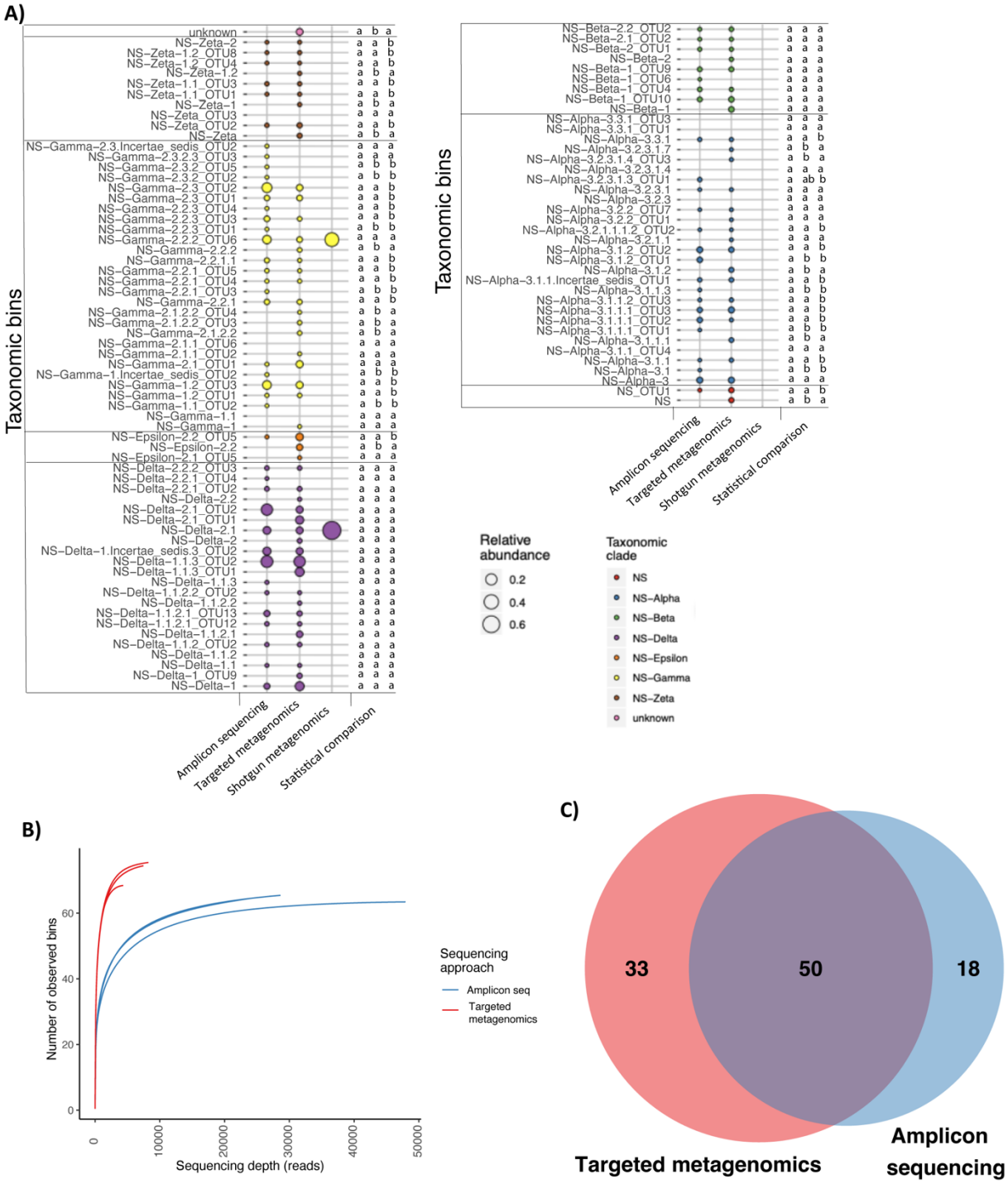


Fig. 4. A) Relative abundance of archaeal *amoA* ASVs sequenced with *amoA* amplicon sequencing (*amoA* gene PCR amplicons sequenced with Illumina MiSeq, about 22000-47000 *amoA* reads per sample) (n = 3), targeted metagenomics (genomic DNA sequenced with Illumina MiSeq, after target enrichment, about 4500-8500 *amoA* reads out of 220000-310000 targeted reads per sample)(n = 3), and shotgun metagenomics (genomic DNA sequenced with Illumina HiSeq, about 22-57 million reads per sample, of which only two

929 were archaeal *amoA*)(n = 3). In all cases, the analysis pipeline consisted of USEARCH comparison on
930 *amoA* reference database (Alves et al., 2018) and taxonomy were analyzed with DADA2/QIIME1 v1.9.1
931 pipeline as published earlier (Siljanen et al., 2019). Sample having largest diversity is shown. Bubble
932 symbol size is proportional to the relative abundance and color indicates Nitrosospheria (NS)-clade of
933 taxonomic bin. Statistically significant differences in relative abundance of taxonomic bins between
934 sequencing approaches based on Kruskall-Wallis rank comparison are indicated with different letters ($P <$
935 0.05, n=3) B) Rarefaction analysis of archaeal *amoA* obtained by amplicon sequencing and captured
936 metagenomics. C) Venn-diagram of archaeal *amoA* sequences obtained by amplicon sequencing and
937 targeted metagenomics.

938

939

940