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

H.M.P. Siljanen[1,2], L. Manoharan[2], A.S. Hilts[2], A. Bagnoud[2], R.J.E. Alves[2,a], C.M. Jones[3], F. 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

a. Present address: Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract

Microbes are the key players in the global cycling of nitrogen (N) and carbon (C) processes, availability and fluxes, and losses through emission of nitrous oxide (N_2O) and methane (CH_4), two very potent greenhouse gases. Thus, characterization of microbial functional guilds involved in these processes is high on the political agenda. Yet, typical sequence based characterization often represents 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 the 15 selected marker genes, resulting in 264,000 unique probes in total. These probes were used to selectively enrich the target genes in shotgun metagenomic libraries. In validation experiments with a mock community of cultured microorganisms, the target gene profiles were similar to those of the original community when sequenced with targeted metagenomics. Further, 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 35 times or 18% higher diversity estimates than when using shotgun metagenomics or targeted PCR amplification, respectively. Thus, captured 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 across ecosystems.

Introduction

The global nitrogen (N) and carbon (C) cycles are essential components of the Earth's biosphere. All major N transformation processes, i.e., nitrogen fixation, nitrification, denitrification and anammox, are performed exclusively by specific functional guilds of bacteria and archaea that are critical for ecosystem functioning. Their activities are thus crucial in regulating N availability for primary producers and microorganisms across ecosystems, as well as emissions of the potent greenhouse gas nitrous oxide (N_2O) and other volatile N compounds through nitrification and denitrification. Methane (CH_4) is 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 very specific groups of microorganisms. Importantly, CH_4 is the second most potential greenhouse gas after CO_2, and, together with N_2O,contributes at least 25% of the total global warming caused by greenhouse gases (The global carbon project 2020, Myhre et al., 2013).

Shotgun metagenomics have been useful to provide an unbiased picture of the natural microbiome. However, in complex and diverse microbial communities, such as those found in soils and sediments, even deeply sequenced metagenomes only generate reliable information for a small fraction of the diversity of microbial functional guilds involved in inorganic N and C cycling, mainly due to their low relative abundances. Therefore, 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 C cycling pathways (Knief 2015; Brauer et al. 2021; Alves et al. 2018; Romdhane et al. 2021). However, the gene diversity involved in these biogeochemical processes captured by PCR-based approaches is limited and biased by the primers used, rendering quantitative assessments and comparisons between genes difficult. Further, 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. 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).

To improve the analysis of critical, but low abundant microbial guilds involved in N and CH_4 cycling within natural communities, a probe-based targeted metagenomics approach for key genes involved in these processes could be advantageous . For eukaryotes, targeted, high-throughput sequencing approaches known as probe capture, also called hybridisation capture or hybridisation-based target enrichment, have been used to study complex eukaryotic samples such as human exons, ancient human genomes, plant transcriptomes, and cancer marker single nucleotide polymorphisms (Skoglund et al. 2014, Ichida et al. 2019, Futema et al. 2012, Kamil et al. 2021, Bewicke-Copley et al. 2019.). Different from sequence capture, where multiple probes are designed to cover a whole exon, captured metagenomics relies on large databases of multiple target genes clustered at predefined similarity cut-offs, which are used to design a large number of generic probes that are able to capture an extended sequence space for each of these sequence clusters (Kushwaha et al., 2015, Manoharan et al., 2015).

Here, we developed and tested a novel probe-based metagenomics approach targeting 15 distinct functional genes involved in inorganic N cycling (i.e., N fixation, nitrification, denitrification, anammox, and DNRA), and genes involved in CH_4 production and consumption, based on newly developed gene and probe databases. To aid in the design of probe libraries to capture the diversity of N and CH_4 cycling genes, the MetCap bioinformatics pipeline was developed as a protocol to produce probes from highly complex datasets for targeted metagenomics (Kushwaha et al., 2015). This approach has been successfully used earlier to identify genes involved in microbial C metabolism in soil (Manoharan et al., 2015, Manoharan et al., 2017). We evaluated our approach using mock communities comprising DNA from organisms involved in inorganic N cycling, or CH_4 production or consumption with varying gene GC% content, which is known to affect probe hybridization efficiency. As proof-of-concept for use in complex communities, we used agricultural and wetland soil to compare traditional metagenomics and targeted metagenomic, and also compared the diversity of archaeal *amoA* genes (marker for ammonia oxidizers) and *nosZ* genes (marker for nitrous oxide reduction) obtained using our targeted probe capture approach with that obtained by shotgun metagenomics or PCR amplification in samples from two distinct soil types.

Results

Development of databases and probes for targeted metagenomics

We compiled a total dataset of ~500,000 genes from public databases, after excluding redundant sequences and incorrect annotations, which comprised genes encoding the nitrogenase iron subunit (*nifH*), bacterial

and archaeal ammonia monooxygenase subunit A (*amoA)*, nitrite oxidoreductase beta subunit (*nxrB)*, hydrazine oxidoreductase A (*hzoA)*, formate dependent nitrite reductase (*nrfA*), periplasmic nitrate reductase alpha subunit (*napA*), respiratory nitrate reductase alpha subunit (*narG*), copper-containing nitrite reductase (*nirK*), cytochrome cd_1 nitrite reductase (*nirS*), nitric oxide reductase B subunit (*norB*), nitrous oxide reductase (*nosZ*), particulate methane monooxygenase subunit A (*pmoA*), soluble methane monooxygenase component A alpha chain (*mmoX*), and methyl-coenzyme M reductase I subunit alpha (*mcrA*). This information supported the generation of ~264,000 unique gene-specific biotin-labeled probes that were used for targeting the desired sequences in DNA from samples using streptavidin coated magneto-beads (Kushwaha et al., 2015; Manoharan et al., 2015).

Relative gene detection and quantification using targeted metagenomics

The targeted metagenomic approach was evaluated using mock microbial communities comprising predefined mixtures of genomic DNA from 15 bacterial and three archaeal strains, together containing all 15 functional genes targeted by the probes and mixed in different proportions to produce six different median G+C contents ranging from 47% to 62% (Table S1). The overall relative abundance of target genes in the captured metagenomes across all GC% categories was strongly positively correlated with those in the original mock communities ($r = 0.78$ ($P < 0.0001$) (Fig. 1). Comparisons between the relative abundance of individual genes in the mock communities as compared to captured metagenomes revealed nearly identical values (t = 1.8518e-10, df = 174.83, $P > 0.05$), with the exception of the *nirK* gene (Fig. S1a,b). Pearson correlation coefficients across all GC% categories varied between 0.76 and 0.93, although correlations were stronger within higher GC% categories (57%-63%; Rpearson = 0.80-0.93) than within lower GC% categories (47%-53%; Rpearson = 0.76-0.79) (Fig. S2).

Precision of gene identification

Target genes in the mock-communities were analyzed using both in-house and KEGG orthology Hidden Markov Models (HMMs) with precision and recall values calculated for each set of models. To determine these values, we identified which reads constituted a "true positive" or a "true negative" by mapping them to their corresponding proteomes. Protein sequences encoded by each genome were annotated using KOFAM (a customized HMM database of KEGG Orthologs). In turn, reads mapped to a given protein were assigned the function according to the above-mentioned KOFAM annotation. Matching assignments of reads from the HMMs (both in-house and from KEGG's orthology database) were counted as true positives. Precision for the in-house HMMs ranged from 74.1% (*nifH*) to 100.0% (*mcrA*) (Fig. S4). The average precision of the models was 93.4%, with a median of 99.8% (excluding *amoA* and *pmoA*). The recall ranged from 42.8% (*nxrB*) to 100.0% (*mcrA*). The average recall was 81.9%, with a median recall of

82.8%. Because KEGG only provides a model that combined the homologous genes *pmoA* and *amoA*, it was not possible to accurately calculate the precision and recall of the in-house models for these genes. An estimate of the precision could still be determined for the two genes though, where a "true positive" meant that the hit identified by the in-house models mapped to the HMM for (*a*/*p*)*moA* genes provided by KEGG (K10944). In this case, the precision of *amoA* gene was estimated to be 96.9%, and similarly, *pmoA* gene was estimated to be 100.0%. It is important to note that these estimates do not exclude the possibility that the in-house models predicted that a read belonged to an *amoA* gene, when it actually belonged to a *pmoA* gene, or vice versa.

Results using KEGG Orthologs' HMMs instead of in-house HMMs showed similar results. Note, however, that the results for *amoA* and *pmoA* genes were combined due to the model for the associated KO being the same, as mentioned above . Because KEGG only offered a model combining *amoA* and *pmoA* genes, the custom HMM produced a number of gene matches for *pmoA* and *amoA* for which also archaeal-*amoA* (*TamoA*) were possible to distinguish. Precision ranged from 76.5% (*a*/*pmoA*) to 100.0% (*mcrA*), while recall varied from 34.3% (*nifH*) to 100.0% (*mcrA* and *nxrB*). The average and median precisions were 94.5% and 97.7%, while the average and median recalls were 93.39% and 99.03%, respectively. The precision of KOs was lower for *napA* genes than for other genes when compared with custom DNA HMMs. This could be accounted for by the presence of formate dehydrogenase, a known homolog of *napA*. Formate dehydrogenase (K00123) was indeed identified among the genomes used for the mock community, so it is possible that the custom HMM was misclassifying reads originating from this known ortholog to *napA* (Cerqueira et al., 2015).

Overall the in-house HMM searches performed well when compared to reads mapped to the genomes and assigned KEGG Orthologs based on the full matched protein sequences, which indicates that the HMMs are able to accurately classify the reads in the mock-community. The in-house HMM was a feasible tool for detecting distantly related genes or novel clades of the gene, and can therefore be a more sensitive tool for detecting assigned KEGG Orthologs. As the results obtained with the custom HMM models generally yielded higher precision than those based on KEGG Orthologs, they were used for the final analyses. Noteworthy is that reads identified as false positives may in fact not be incorrect assignments, but could simply represent novel diversity of the genes of interest, not captured by the sequence database for the KEGG Orthologs.

Comparison between targeted and traditional metagenomes, and PCR amplicon sequences from complex environmental samples

We investigated the efficiency of the targeted metagenomics approach in natural complex ecosystems by directly comparing this approach with traditional untargeted metagenomes generated with a nearly 200 times higher sequencing depth (~220,000 reads and ~43 million reads per sample, respectively) in soils from a wetland in Bellefontaine, France, and from an agricultural field in Hungary (Table S2). These soils represent distinct ecosystems with average physicochemical conditions, such as water and oxygen conditions, and N availability that favor different distributions of functional genes (Jones et al. 2014). Despite the large difference in sequencing depth, targeted metagenomics generated a much larger set of target functional genes (Fig. 2), and up to 60 times as many identified gene sequences (Fig. 3), compared to traditional metagenomics. The relative abundances of functional genes in captured metagenomes correlated significantly with that in traditional metagenomes from both the agricultural field ($R_{Pearson}$ = 0.96, d.f.=14, $P < 0.00001$) and the wetland ($R_{Pearson}$ = 0.70, d.f,=14, $P < 0.01$). Relative target gene abundances were also not different between the two approaches ($P > 0.90$) (Fig. 3). However, reduced relative abundance was detected with targeted metagenomics in the agricultural site for *nxrB*, *nrfA*, *narG* and *nosZ* genes and the wetland site for *nxrB*, although, the targeted approach generated a much higher number of reads for these genes in both sites ($P > 0.0001$) (Fig. 3.). The lower relative abundance genes at the agricultural site can be due to some of the other genes being detected more efficiently with the targeted approach. The different relative abundances of functional genes in the two ecosystems reflected the expected differences in their environmental conditions: (i) N_2 fixation genes were higher in the wetland, which are not fertilized; (ii) Genes involved in ammonia oxidation and in nitric and nitrous oxide reduction were higher in the agricultural soil, which are likely to be more oxic and are also subject to N fertilization resulting in higher N_2O production rates ; (iii) Genes involved in methanogenesis were more abundant in the wetland, where anoxic conditions and methane production typically occur.

In order to compare the diversity of specific functional genes detected by targeted metagenomics, traditional metagenomics, and PCR amplicon sequencing, we used the archaeal *amoA* gene as a test case, for which there is a comprehensive reference database for classification available (Alves et al., 2018). Additionally, we compared the detection of *nosZ* genes between the targeted and untargeted metagenomics approaches in the agricultural soil, also based on an existing reference database (Graf et al., 2022). Comparisons were performed on different levels between: (i) targeted and traditional metagenomes, and PCR amplicons of the archaeal *amoA* gene in the agricultural soil (Fig. 4); (ii) targeted metagenomic reads of archaeal *amoA* genes in an arctic peatland soil and an archaeal *amoA* gene amplicon dataset (Fig. S3), generated from the same soil in a previous study (Siljanen et al., 2019); and (iii) *nosZ* gene sequences identified by targeted or traditional metagenomics in the agricultural soil (Fig. S4).

Comparison between agricultural and wetland soils showed that targeted metagenomics detected substantially higher diversity of archaeal *amoA* genes than PCR amplicon sequencing, while, as expected, traditional metagenomics detected a lower diversity than either of those two methods (Fig. 4A). Specifically, targeted metagenomics detected 71-83 distinct taxonomic *amoA* gene bins, whereas PCR amplicon sequencing and traditional metagenomics detected 65-68 and only two bins, respectively. Rarefaction analysis also indicated that targeted metagenomics retrieved a higher alpha diversity of archaeal *amoA* genes than PCR amplicon sequencing (Fig. 4B,C). Likewise, comparisons based on the peatland soils showed that targeted metagenomics detected a greater diversity of archaeal *amoA* genes than PCR amplicon sequencing, by detecting gene variants affiliated with nine different Nitrososphaera-related clades, in comparison with only two when amplicon sequencing was used (Fig. S3).

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

Discussion

We developed and evaluated a new targeted metagenomics approach using probe capture to investigate the diversity and distributions of microbial functional groups involved in the cycling of inorganic N and CH_4. This included the development of an extensive set of probes targeting 15 well-established marker genes encoding key metabolic enzymes and a bioinformatic pipeline to process and annotate the sequence data. In validation experiments with mock communities comprising 18 distinct organisms, we showed that this method successfully reproduced the expected relative abundances of all target functional genes, with one exception (i.e. *nirK*). We observed some variation in probe specificity among mock communities with different genomic GC content, with high GC communities having generally higher target gene detection efficiency. Such variation was likely due to stronger probe hybridization to sequences with higher GC content, given the higher binding energy between guanine and cytosine than between adenine and thymine. Targeted metagenomics also accurately reproduced the relative abundances of target genes obtained with traditional untargeted metagenomics in two distinct natural microbial communities from soils, showing that the probe capture step does not introduce significant quantitative bias. Using archaeal *amo*A genes as a case-study, targeted metagenomics with a sequencing depth of 120-180 Mb per sample generated at least

10 times more target gene sequences than traditional metagenomics (maximum number of *amoA* genes among whole metagenomes was two unique taxonomic bins) and 5 times less target genes than PCR amplicon sequencing with 12-20 Gb and 2.5 Mb datasets per sample, respectively. Moreover, targeted metagenomics also detected much higher gene sequence diversity than the two other methods. For example, the new targeted approach detected 35 times higher alpha-diversity of archaeal *amoA* genes than traditional metagenomics in the agricultural soil, and 5 times higher diversity of *nosZ* genes, especially that of genes affiliated with the less characterized Clade II. In turn, the new targeted method detected 18%, and 4.5 times higher diversity of archaeal *amoA* genes than PCR amplicon sequencing in agricultural, and arctic peatland soils, respectively.

It should be noted that the efficiency of this approach largely depends on the coverage and quality of the database used to generate the probes. However, the database is not constrained to full-length gene sequences from a limited diversity of full genomes and long metagenomic scaffolds available, and the probe/target diversity can be greatly extended through inclusion of gene fragments generated by PCR or from short metagenomic contigs. Despite its high target precision, this method may nevertheless capture non-target sequences, especially those of enzymes that are genetically closely related but have different functions, as we observed for *napA* genes and those encoding formate dehydrogenases. Such cases emphasize the importance of thorough data annotation and filtering procedures.

Despite the technological advances and declines in cost of high-throughput sequencing, traditional metagenomics remains impractical and prohibitively expensive to capture the diversity of specific, low abundant functional groups in complex environments. For example, the cost of generating these datasets was 100$/sample for targeted metagenomics, >1000$/sample for traditional metagenomics, and 80$/sample per gene for PCR amplicon sequencing. In turn, PCR-based assays are limited by the biases associated with any set of primers targeting a large diversity of sequences and assay efficiencies, which also preclude detection of highly divergent new sequence diversity. The probe capture approach circumvents these limitations by using multiple short probes targeting different regions of each gene, which largely increases the likelihood of capturing sequence variants that elude PCR amplification. Thus, this approach has not only the potential to capture rare and novel gene diversity in complex environments, but also to identify cryptic microorganisms in low-biomass samples or involved in suggested metabolisms, such as in the phyllosphere (Putkinen et al., 2021) and coral holobiont (Glance et al., 2021). Moreover, targeted metagenomics also overcomes issues associated with running and comparing multiple independent PCR assays when investigating several distinct targets. In that sense, our approach effectively represents a PCR-independent, multiplex approach to characterize simultaneously and in-depth the distributions of a broad range of functional genes.

The targeted metagenomics approach developed here provides an efficient and cost-effective new 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.

Materials and methods

Target database (TDB) collection

A Hidden Markov Model (HMM) model was built based on reference sequence alignments from curated databases already available (Alves et al., 2018; Knief et al., 2015; Jones et al., 2014; Graf et al., 2016; fungene; WGS), or generated here from gene sequence data publicly available from their protein-coding nucleotide sequences, to cover all known diversity in each protein family. Gene sequences from environmental surveys and cultured organisms were collected from Genbank based on annotation and gene structure (hmmer profile search). Structure-based searches were performed over the whole Genbank nt-database with tblastx and nhmmer using the profile HMMs (Wheeler & Eddy, 2013) on a local supercomputer cluster (Centre of scientific computing CSC, Espoo, Finland) (July 2017). Specificity of hmm searches were confirmed by comparing the results to the latest curated sequence databases of selected genes (Alves et al., 2018; Graf et al., 2022, primary data was obtained directly from the respective authors). The final output of these searches were collected and called target gene databases (TDBs). Since some of the genes are homologous, such as *amoA* and *pmoA* genes, and *nxrB* and nitrate reductase genes, the respective TDBs were curated with tblastx/diamond searches against *refseq* database (Altschul et al., 1990) to remove some of the duplicates or move them under correct gene annotation based on curation. This sequence selection process generated ~500,000 unique sequences across all genes.

Generation of probes for target genes

Targeted metagenomics strategy here was developed based on the NimbleGen SeqCap EZ protocol (Roche NimbleGen, Inc.). All the nucleotide sequences from the step above (TDBs) were then processed through the SeqCap pipeline (Kushwaha, et al., 2015) with default parameters for designing up to six unique 50 mer probes for each sequence cluster (with 80% identity as the sequence clustering threshold). This produced probes on clusters produced from ~500,000 unique sequences . The target databases (TDBs) for each gene were clustered at an identity cut-off of 80% using the SeqCap pipeline (Kushwaha, et al., 2015), and up to

six 50-mer probes were designed for each cluster. Information of these probes were then used to produce the NimbleGen SeqCap probe set for targeted metagenomics.

Extraction of DNA from cultures and environmental samples

DNA from mock communities encoding the genes of interest was generated by mixing DNA from different organisms in variable proportions, in order to simulate variation in overall GC% content. Mock communities included: *Methylosinus trichosporium* Ob3b, *Methylocella tundrae*, *Methylomicrobium buryatense*, *Nitrosospira multiformis*, *Nitrososphaera viennensis*, *Nitrospira defluvii*, *Pseudomonas aeruginosa*, *Escherichia coli*, *Shigella sonnei*, *Cupriavidus metallidurans*, *Cupriavidus necator*, *Dyadobacter fermentans*, *Pseudomonas stutzeri*, *Rhodobacter sphaeroides*, *Salinibacter ruber*, *Sulfurimonas denitrificans*, *Methanoregula boonei*, *Methanolacinia petrolearia*. Strain information, including genome size, number of target genes per organism, and relative abundance in mock communities are shown in Supplementary Table S1. The mock community samples used for validation of the targeted metagenomic approach, were pooled, resulting in different GC content categories due ending up to different median GC% median genome content of the genomes of each organism in the pool of targeted mock-community samples. This is , as shown as the relative abundance of the microorganism multiplying with the median genome GC content to have a weighted GC content of the pool of samples finally ending to: 47, 50, 53, 57, 60, and 63 GC%, Supplementary Table S1). Pooling of organisms was done based on the expected number of functional genes in extracted DNA, concentration of DNA, genome size, and finally it was confirmed with 16S rRNA qPCR in each DNA sample. Relative gene abundances in mock communities were calculated and compared to the sequences generated using targeted metagenomics.

Samples from different environments were collected for testing the effectiveness of the probe set to enrich functional genes in different sample contexts of various GC% produced with mock-community of cultured organism DNA. For the comparison environmental samples were collected from agricultural soil in Hungary, and from a wetland in Bellefontaine, France (Table S2). Extraction of DNA from environmental samples was performed as previously described by Siljanen et al., (2019), with phenol:chloroform:isoamyl alcohol extraction in CTAB buffer, followed by ethanol precipitation. DNA of cultured organisms was extracted from 1% SDS treated cells in CTAB buffer, followed by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation, as above. The quality of DNA extracts was assessed using a NanoDrop ND-1000 (Thermo), and DNA concentration was measured with a Qubit HS dsDNA kit (Thermo).

Library preparation, pooling, target enrichment of libraries and sequencing with Illumina Miseq

For each sample library, unique sequencing indexes and sequencing adapters were produced in the Centre for Genomic Research (CGR), University of Liverpool, Liverpool, UK. These libraries were then pooled

together uniformly, quality controlled with Agilent Bioanalyzer and hybridized to the NimbleGen SeqCap capturing probe set according to the manufacturer's instructions at 47°C for 72 hours in CGR as previously described (Manoharan et al., 2015). Sequencing was done with the Illumina MiSeq PE250 chemistry. Sequencing outcome was 30,000-200,000 PE250 per environmental samples and about 2,000,000 PE250 for cultured samples.

Shotgun sequencing with Illumina HiSeq and *amoA* gene amplicon sequencing

Comparison of different sequencing strategies were done by comparing targeted metagenomics, shotgun metagenomics and amplicon sequencing of archaeal *amoA* genes. Shotgun metagenomic sequencing was performed on environmental DNA from agricultural soil and wetland soil. Library preparation and sequencing in one Illumina HiSeq PE150 lane were performed at the Vienna Sequencing facility, Austria and Microsynth AG, Switzerland. Archaeal *amoA* gene amplicons were generated with PCR conditions described earlier by Siljanen et al., (2019), and sequencing was performed with Illumina MiSeq PE250, in Munich, Germany.

Mock-community genome protein prediction

Validation of HMM model predictions was performed on the mock communities using independent models. Predicted proteins for the genomes used in the mock community were retrieved from NCBI. For one genome (*Nitrospira defluvii*; NC_014355.1), proteins were predicted from the nucleotide sequence using Prodigal (v2.6.3) (Hyatt et al., 2010). All HMM models (i.e., the models associated with each KO) for prokaryotic organisms were downloaded from KEGG and concatenated. The concatenated set of models was run against each genome using HMMER's hmmsearch (v3.3)(Eddy, 1998). Each protein of each genome was assigned the top hitting KO. Where available, cut-offs provided with KOFAMscan (Aramaki et al., 2020) were used to filter results.

Read preparation and mapping

All six possible reading frames of the nucleotide sequence reads for the mock communities were translated using transeq from the EMBOSS package (v6.6.0.0). These were mapped to the genomes using DIAMOND blastp (v2.0.6.144). A minimum percentage identity of 60% was used for the mappings. No minimum coverage was used to account for reads that did not fully overlap with the proteins from the genomes. For each direction, up to four matches were retained, to account for cases where neighboring proteins were spanned by a single read. Multiple matches in the same direction could overlap by up to 15 amino acids to account for possible protein fusions or sequencing errors. Each of the, up to four, matches were used to assign KOs, based on the predicted KOs of the proteins (see above; Genome protein prediction) to which

they mapped. If, in a given direction, a read was assigned the same KO multiple times, this was collapsed into a single hit, retaining the hit with the best e-value.

Evaluating the methods for protein prediction of KO and in house-HMM searches

Protein predictions for reads were made using HMMs downloaded from KEGG in a similar manner to that described above. For a given read, using all six of the translated frames, up to four protein assignments could be made. Frames were classified according to the best models. If a model for a KO associated with a given function of interest was assigned, then the result was filtered based on a bit score or filtration filtering. Reads assigned KOs for the function of interest were counted as positively identified hits.

In-house HMMs were built using manually curated sequences for the proteins of interest. Protein predictions for each read were performed using nhmmer from the HMMER suite (v.3.3). An inclusion threshold of 0.0001 was used when assigning protein predictions for the reads.

Counts for the true positives, false positives, and false negatives were made for all models using all methods, and precision and recall calculated. Precision was calculated as the ratio of true positives (TP) to total positive detections (i.e., all true and false positives; FP).

Recall was the ratio of true positives, to all theoretically obtainable true positives (i.e., all true positives and false negatives; FN).

Read mapping (described above) was used to determine the total possible true positives of all reads for all proteins, which was in turn used to calculate false negatives, by subtracting the total true positives from the total possible positives.

Comparison of diversity in environmental samples produced with different sequencing strategies

Quality trimming of sequencing reads were done with *seqtk* -program and in dada2 pipeline by truncating amplicon reads to 200bp and discarding all reads with expected error greater than 2. Archaeal *amoA* reads from amplicon and targeted and non-targeted metagenomics, were classified with pipeline utilizing commands of dada2, USEARCH8, UCHIME, UCLUST and BLASTN (Altschul et al. 1990; Edgar, 2010; Edgar et al., 2011; Caporaso et al., 2010; Callahan et al., 2016) to assign amplicon/metagenome sequence variants (A/MSVs) as earlier described (Siljanen et al., 2019) and earlier described reference database of archaeal *amoA* genes were used (Alves et al., 2018). (script available at: https://github.com/alex-bagnoud/arctic-nitrifiers). Relative abundance of each functional gene was quantified either using BLASTN against TDB or by using the target gene hmmer profiles to search sequences of each functional gene with

maximum E-value cut-off (E < 0.001). The gene functions of each sequence were confirmed against the *refseq* database with tblastx in diamond sequence search accelerator (Buchfink et al., 2015) . Annotation and identity information was further processed with awk in unix and in Rversion 3.5.3 to produce lists of community structures of each gene. The sequencing depths for the *amoA* gene amplicon library was ~10,000 PE250 reads per sample for shotgun metagenomics 40-70 M PE150 reads per sample, and for targeted metagenomics 1.4-5.8k of PE250 reads per sample.

Phylogenetic placement of nosZ reads from shotgun and probe capture libraries.

A reference phylogeny and alignment for *nosZ* was generated from full length *nosZ* amino acid sequences obtained from genomes downloaded from the NCBI genome database (accessed October 2019). A profile HMM for *nosZ* was generated from a previously published *nosZ* dataset (Graf et al, 2022), and then used to search translated open reading frames in downloaded genomes (e-value = 0.0001). Hits were then dereplicated at 99% amino acid similarity and aligned by amino acid using *hmmalign* within the HMMer suite (v.3.1.2). Sequences that were homologous to *nosZ* were identified using FastTree (v. 2.1.11, Price et al., 2010) to generate an initial phylogeny with the resulting hits, which along with the alignment was imported into the ARB database software (Ludwig et al., 2004). The dataset was then manually curated to remove non-target homologues as well as sequences from poor quality genomes based on checkM screening (v1.0.18, Parks et al., 2015), as well as correct alignment errors. The final amino acid alignment was then used for phylogenetic analysis using IQ-TREE (v1.6.12). The Le-Gascuel (LG) substitution model with 10 rate categories was selected after automatic model selection and branch support was determined using the ultrafast bootstrap approximation, as implemented in IQ-TREE.

Phylogenetic placement of *nosZ* reads obtained using either shotgun metagenome sequencing or probe capture methods was performed by translating reads and aligning to the reference *nosZ* alignment using *hmmalign*. Placement of reads in the reference phylogeny was then performed using the next generation evolutionary placement algorithm (EPA-NG v0.3.8; Barbera et al., 2018), with the same model parameters as used for constructing the reference phylogeny. Placement positions with the highest likelihood weight ratio were plotted on the reference phylogeny in R using the 'ggtree' package (v3.4.2, Yu et al., 2016).
The metagenomic, targeted metagenomic and *amoA* amplicon sequencing data were archived in NCBI SRA under the Bioproject numbers PRJNA898102 and PRJNA488558.

Acknowledgements

H.S. was financially supported by the Atmosphere and Climate Competence Center (ACCC) Academy of Finland project number 337550, 342362. 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). 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.

References:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. 1990. Basic local alignment search tool. Journal of molecular biology. 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. Nature Communications 9: 1517. https://doi.org/10.1038/s41467-018-03861-1.

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. Bioinformatics. 36:2251-2252. doi: 10.1093/bioinformatics/btz859.

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. Systematic Biology 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. Nature methods. 12: 59.

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

Bräuer SL, Basiliko N, Siljanen HMP, Zinder SH. 2021. Methanogenic archaea in peatlands, *FEMS Microbiology Letters*, 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. Nature 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. Nature Methods 7, 335–336.

Cerqueira NMFSA, Gonzalez PJ, Fernandes PA, Moura JJG, Ramos MA. 2015. Periplasmic Nitrate Reductase and Formate Dehydrogenase : Similar Molecular Architectures with Very Different Enzymatic Activities. Accounts of Chemical Research 48, 2875-2884
DOI: 10.1021/acs.accounts.5b00333

Danilova, O., Suzina, N., Van De Kamp, J. *et al.* 2016. A new cell morphotype among methane oxidizers: a spiral-shaped obligately microaerophilic methanotroph from northern low-oxygen environments. *ISME J* 10: 2734–2743 https://doi.org/10.1038/ismej.2016.48

Futema M, Plagnol V, Whittall RA, et al. 2012. Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. Journal of Medical Genetics 49:644-649.

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 Microbiology Ecology, 98:1-11

Glaze TD, Erler DV, Siljanen HMP. Microbially facilitated nitrogen cycling in tropical corals. ISME J. 2021 Jul 5. doi: 10.1038/s41396-021-01038-1. Epub ahead of print. PMID: 34226659.

Eddy SR. 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.

Hyatt D, Chen GL, LoCascio RF. *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.

Jones, C., Spor, A., Brennan, F. et al. 2014. Recently identified microbial guild mediates soil N_2O sink capacity. Nature Clim Change 4: 801–805. https://doi.org/10.1038/nclimate2301

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

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.

Ludwig, W., Strunk, O., Westram, R., et al. 2004. ARB: a software environment for sequence data. Nucleic Acids Research 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 Research 22: 451-460.

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

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.

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

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

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490.

Putkinen, A., Siljanen, H.M.P., Laihonen, A., Paasisalo, I., Porkka, K., Tiirola, M., Haikarainen, I., Tenhovirta, S. and Pihlatie, M. 2021. New insight to the role of microbes in the methane exchange in trees: evidence from metagenomic sequencing. New Phytol 231: 524-536. https://doi.org/10.1111/nph.17365

Skoglund P, et al. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci USA 111, 2229–2234 .

Siljanen HMP, Alves RJE, Ronkainen JG, Lamprecht RE, Bhattarai HR, Bagnoud A., Marushchak ME, Martikainen PJ. Schleper C, Biasi C. 2019. Archaeal nitrification is a key driver of high nitrous oxide emissions from arctic peatlands. Soil Biol Biochem. 137: 107539

Wheeler, T. J., and 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 Ecology and Evolution 8: 28-36.

Figures:

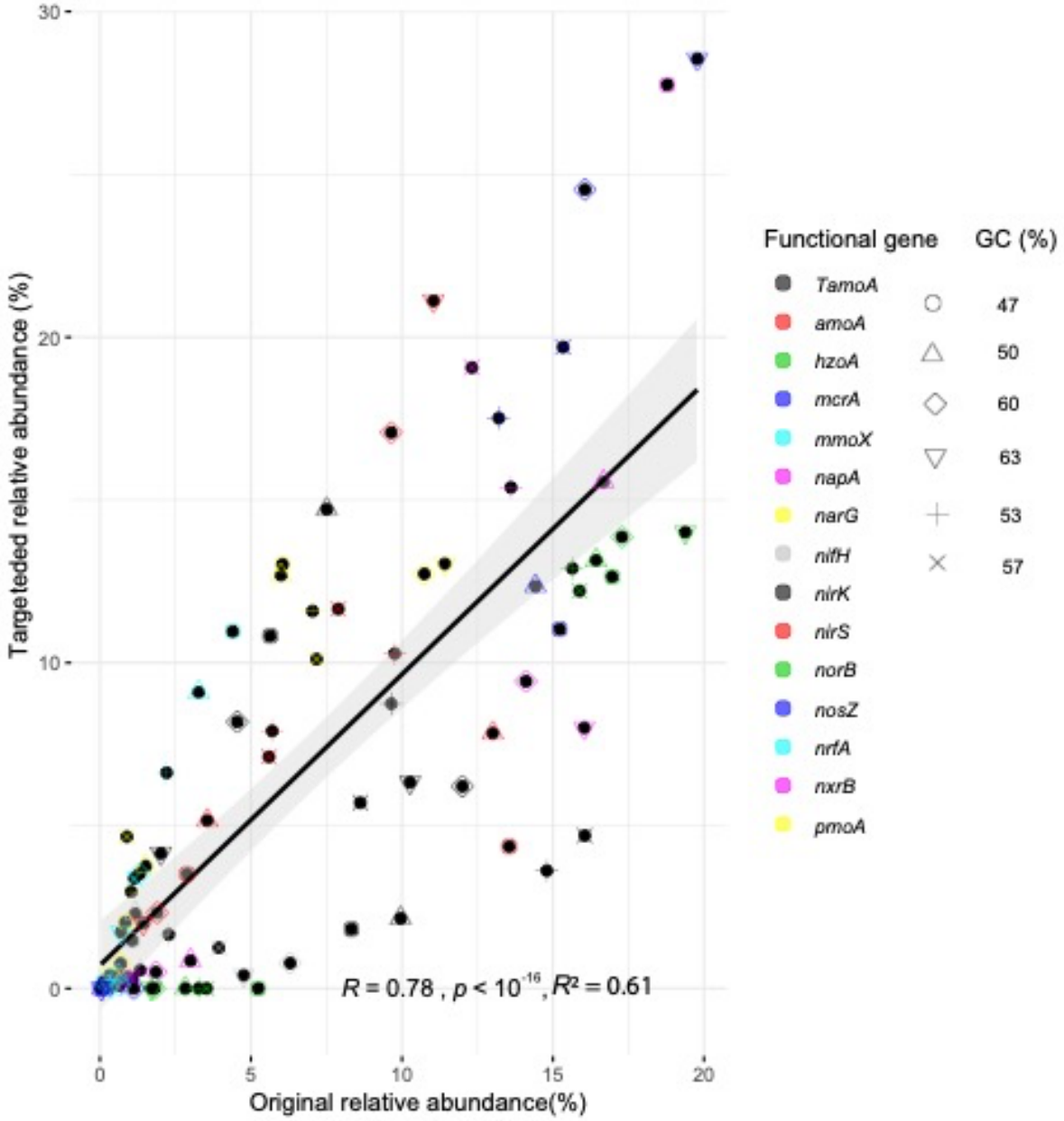


Fig. 1. Comparison between the original composition of the mock community and the relative abundance for each functional gene detected using probe captured metagenomics for all GC% combinations combined. The relative abundance in the mock community was calculated based on the amount of each functional

gene and organism as well as genome size. Color of the symbols depict gene and shape the category of GC content (%).

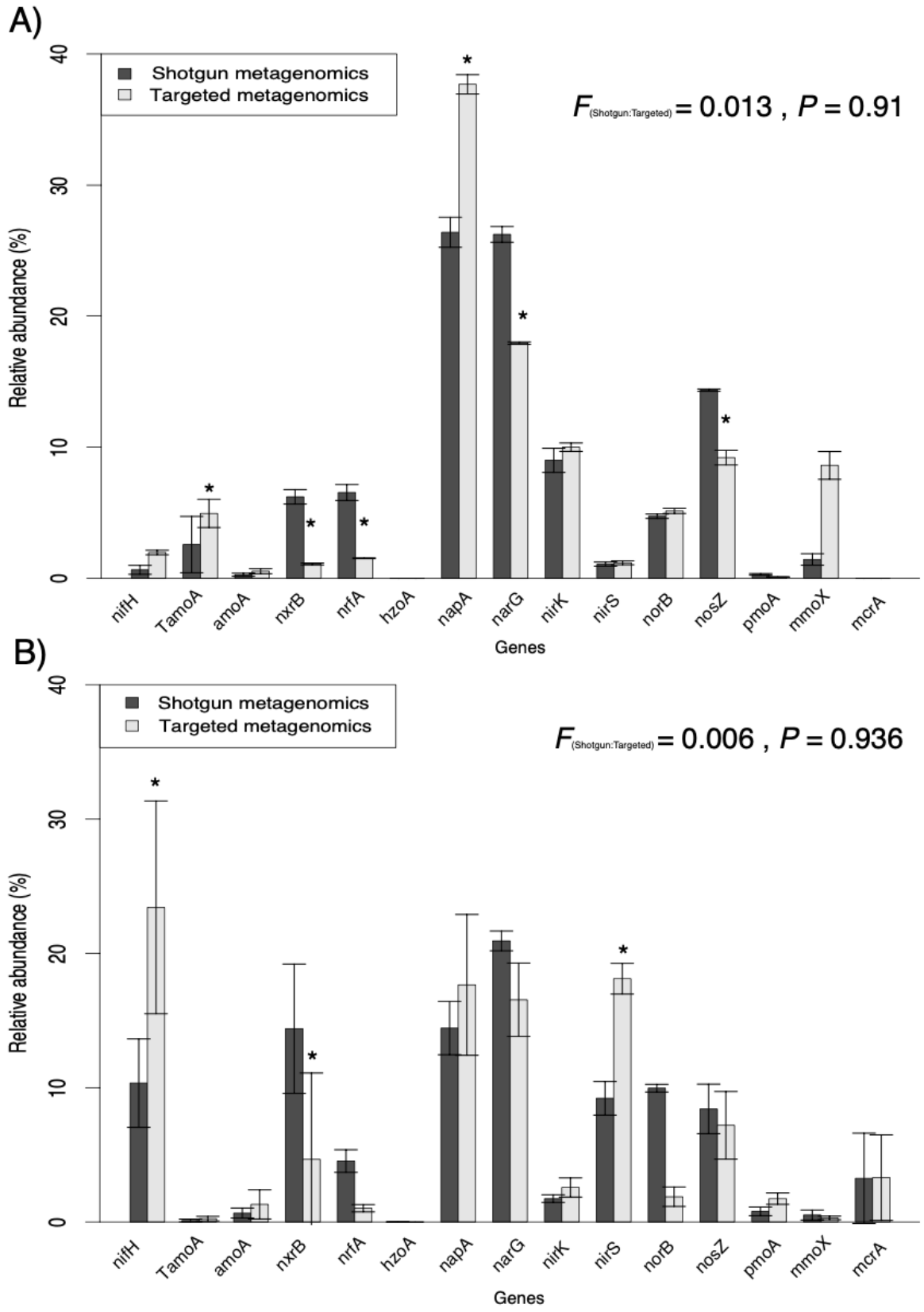

A)
Shotgun metagenomics
Targeted metagenomics
$F_{(Shotgun:Targeted)} = 0.013$, $P = 0.91$
Relative abundance (%)
0
10
20
30
40
nifH
TamoA
amoA
nxrB
nrfA
hzoA
napA
narG
nirK
nirS
norB
nosZ
pmoA
mmoX
mcrA
Genes
B)
Shotgun metagenomics
Targeted metagenomics
$F_{(Shotgun:Targeted)} = 0.006$, $P = 0.936$
Relative abundance (%)
0
10
20
30
40
nifH
TamoA
amoA
nxrB
nrfA
hzoA
napA
narG
nirK
nirS
norB
nosZ
pmoA
mmoX
mcrA
Genes

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).

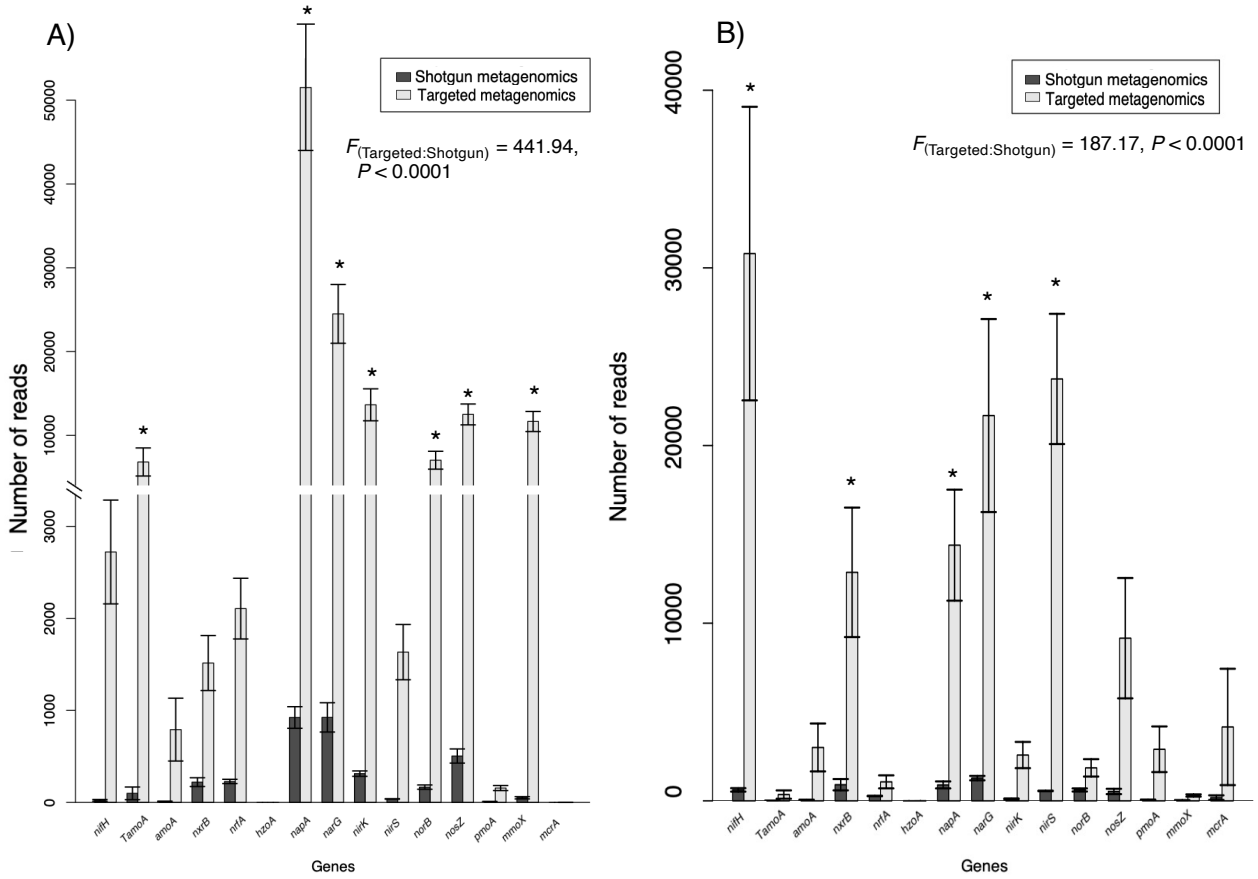


Fig. 3. Total reads of functional genes involved in inorganic nitrogen and methane cycling obtained by targeted metagenomics and shotgun metagenomics from A) agricultural soil, B) wetland soil B). Number of reads was retrieved with HMMER searches of each functional gene. Statistically significant differences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

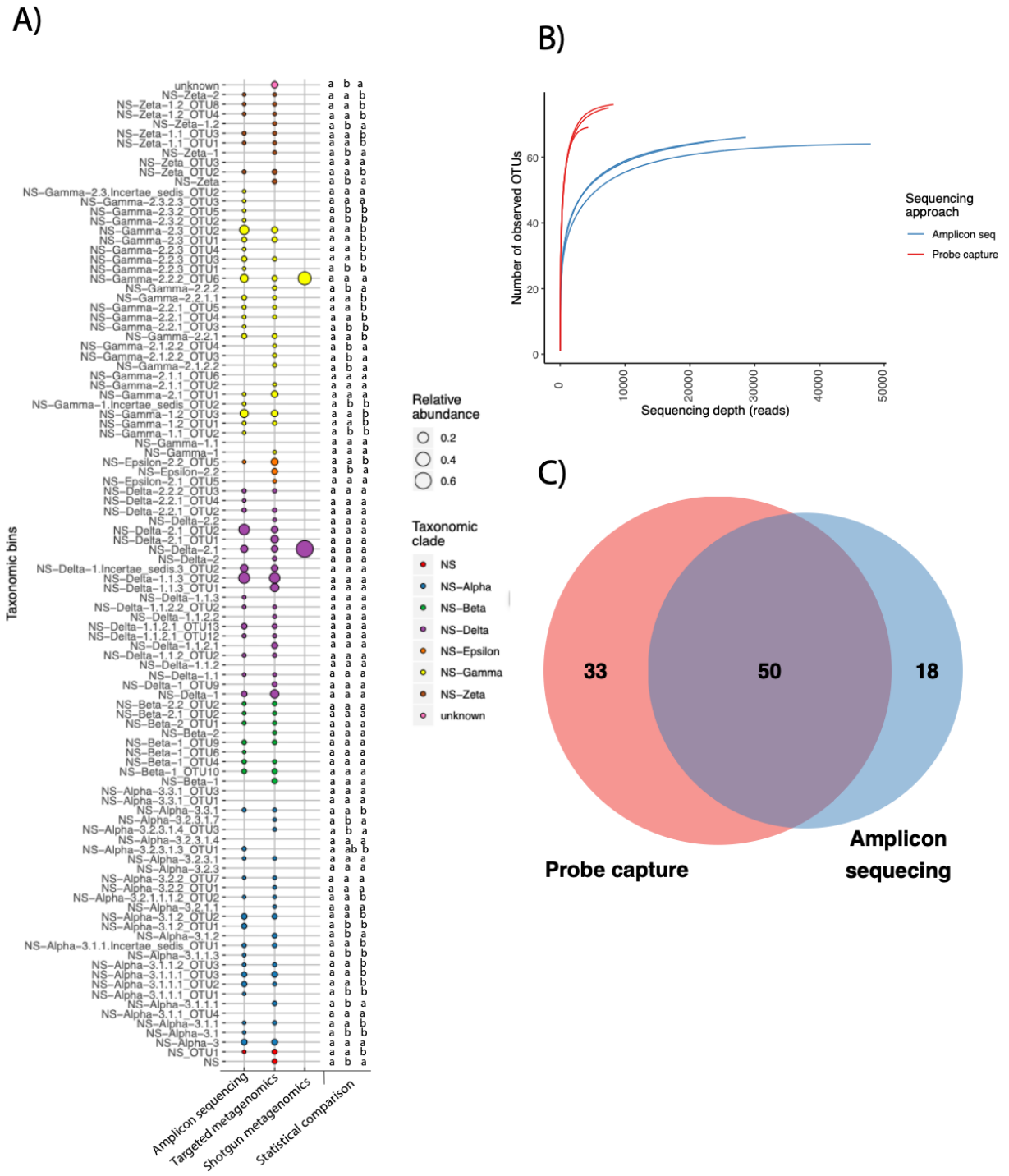
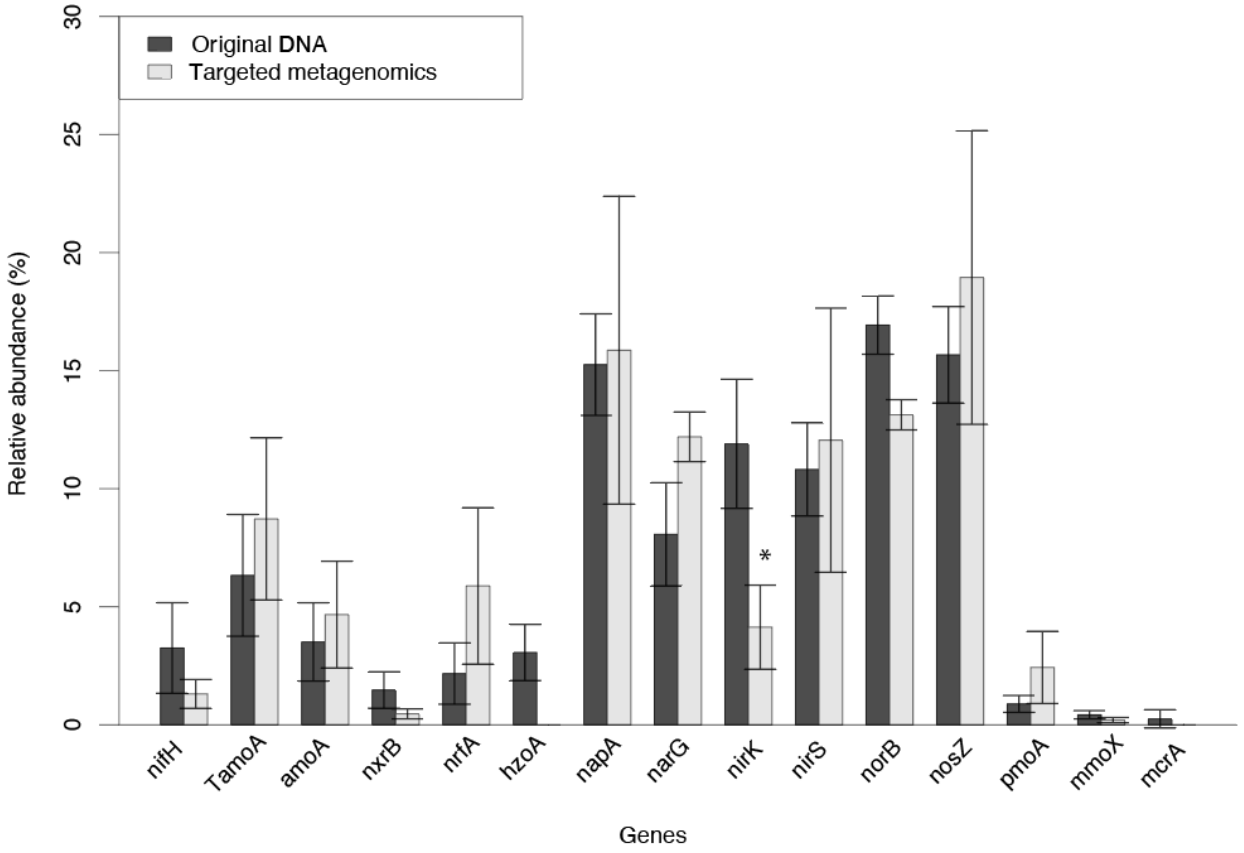


Fig. 4. A) Relative abundance of archaeal *amoA* ASVs sequenced with *amoA* amplicon sequencing (*amoA* gene PCR amplicons sequenced with Illumina MiSeq, about 18000 reads per sample), targeted metagenomics (genomic DNA sequenced with Illumina MiSeq, after target enrichment, about 12000 *amoA* reads out of 220000 captured reads per sample), and shotgun metagenomics (genomic DNA sequenced with Illumina HiSeq, about 44 million reads per sample, of which only two were archaeal *amoA*). In all

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

Supplementary materials:

A)



B)

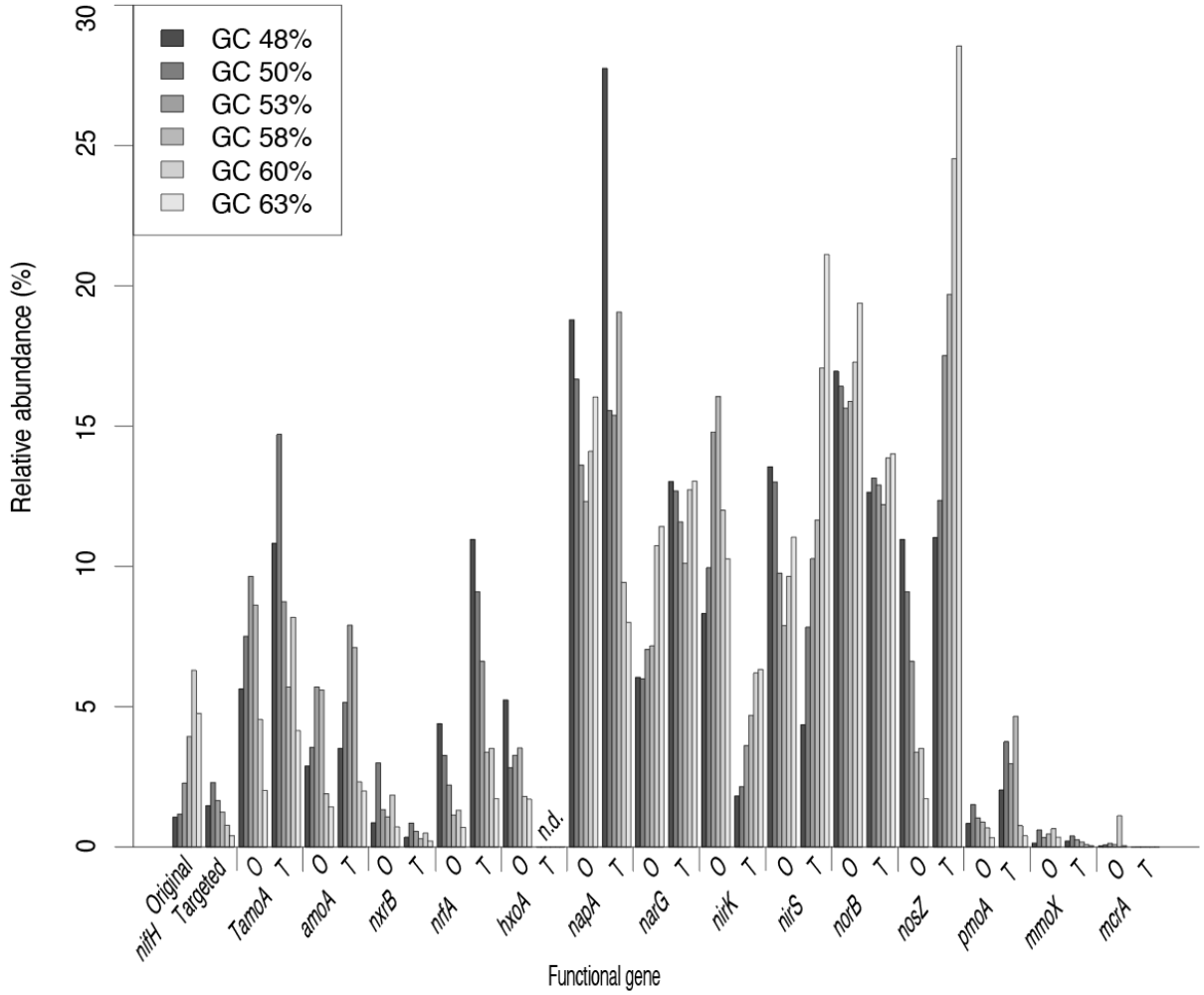


Fig. S1. Relative abundance of functional genes originally mixed into the samples and in captured metagenomes retrieved with the custom HMMs searches. A) Relative abundance of each functional gene in the original DNA mixture and that obtained by captured metagenomics (mean ± s.d., n=6) . Statistically significant differences between original and targeted metagenomic relative abundance according to pairwise comparisons with ANOVA are shown with asterisk ($P < 0.05$). B) Bars with different shades of grey show relative gene abundance values for each GC% category in the original community (O) and those obtained by targeted captured metagenomics (T).

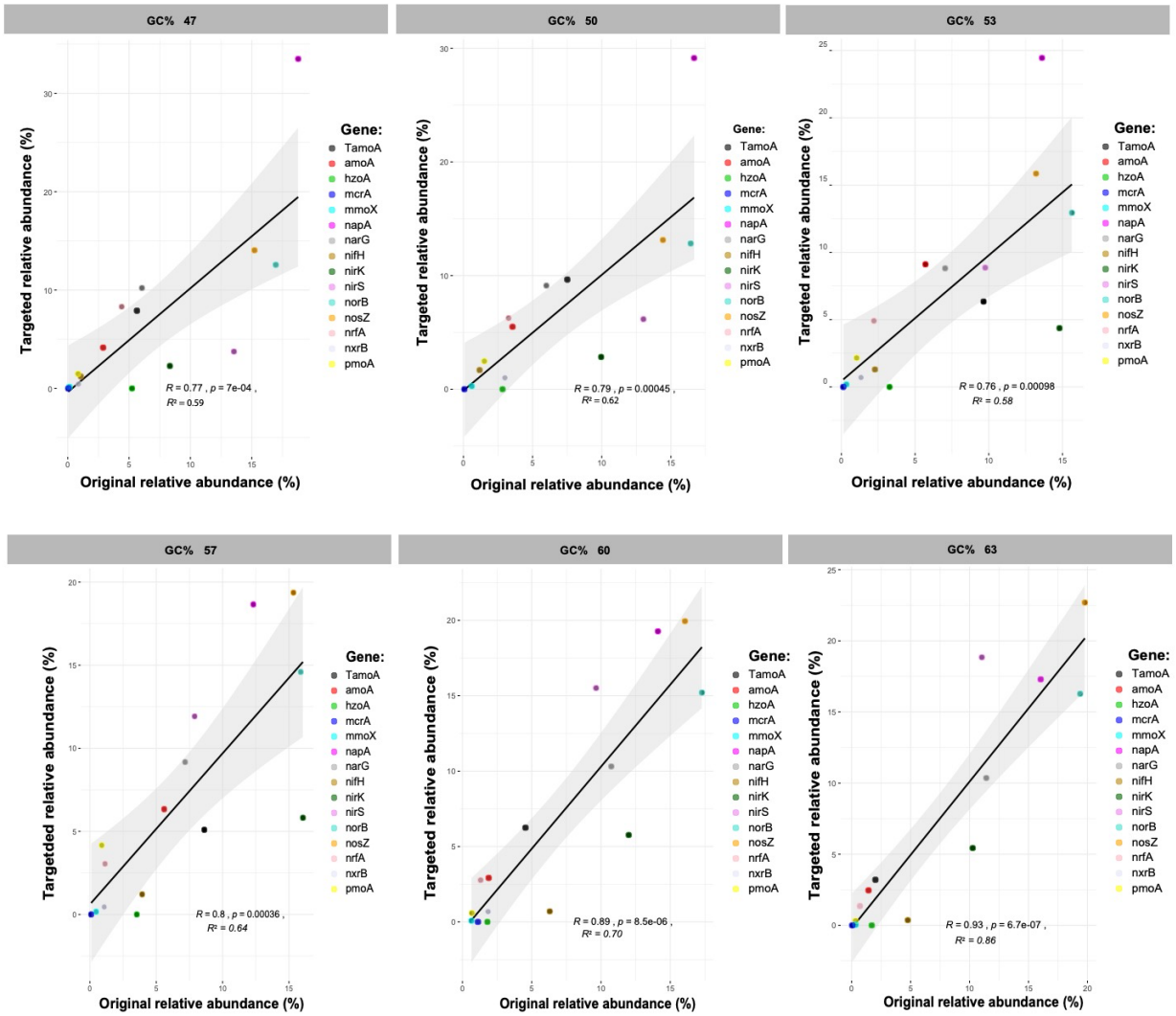


Fig. S2. Comparison of mock community between original (The relative abundance which was calculated with the amount of each functional gene and organism taken into the mock community and genome size.) and targeted metagenomics relative abundance produced for each different functional genes studied with probe hybridized targeted metagenomics for each GC% content separately.

Table S1. The mock-community relative abundance composition, and weighted GC% content of each microorganisms and on each GC% category samples. DNA quality was measured by for UV absorbance based on the with ratio of 260nm/280 nm ratio.

Organism:	**Genome accession (RefSeq)**	**Genome size (bp)**	**N-cycle or CH4-cycle genes**	**Median GC% of genome**	**260nm/280nm**
Nitrosospira multiformis	*NC_007614.1*	3234310	3*amoA, 1*nirK	53.3	2.03
Nitrososphaera viennensis	CP007536.1	2527938	1*TamoA, 1*nirK, putative nosZ?	52.07	2.07
Nitrospira defluvii	NC_014355	4317083	2*nxrB	59	1.27
Ca. Kuenenia stuttgartiensis/hzoA fragment in plasmid vector	(amplified hzoA fragment in vector)	4305	1*hzoA	46.02	ND
Pseudomonas aeruginosa PA96	CP007224.1	6444091	1*narG, 1*napA,1* nirS, 1*nor, 1*nosZ-I	66.2	2.11
Escherichia coli	AE014075.1	5231428	1*nrfA, 1*narG, 1*napA,	50.6	1.92
Shigella sonnei strain FC1706	CP014099.2	4878943	1*nrfA, 1*napA,	50.7	2.15
Cupriavidus metallidurans (Ralstonia eutropha) CH34/CCUG 13724	NC_007973.1	3928089	nosZ-I, 1*nirS, 1*nor, 1*narG, napA	63.58	2.15
Cupriavidus necator (Alcaligenes eutrophus/Ralstonia eutropha) ATCC 17699	NC_008313.1; NC_008314.1	4343281	nosZ-I, 1*nirS, 1*nor, 1*narG, napA	66.3	2.16
Dyadobacter fermentans DSM 18053	NC_013037.1	6967790	1*nosZ-II	51.5	2.05
Pseudomonas stutzeri JM300/DSM 10701	NC_018177	4174118	1*nirS, 1*nirK, 1*nosZ-I, 2*nor, 1*napA, 1*narG	63.3	2.17
Rhodobacter sphaeroides DSM 158/ATCC 17023	NC_007493.2; NC_007494.2	4131542	nirK, 1*nor, nosZ-I, 1*nifH, 1*napA	68.77	2.16
Salinibacter ruber DSM 13855	NC_007677.1; NC_007678.1	3587328	1*nosZ-II	65.98	1.9
Sulfurimonas denitrificans DSM 1251	NC_007575.1	2201561	2*nosZ-II, 1*nirS, 1*nor, 1*napA	34.5	2.13
Methylosinus trichosporium Ob3p	CP023737.1	4508832	3xpmoA, nifH, mmoX,	65.82	1.42
Methylocella tundraea	GCF_900749825.1	4305430	1*mmoX, 1*nifH	63	1.11
Methylomicrobium buryatense 5B	FO082060.1	4668296	1xpmoA, 1*norB, 1*mmoX	48.07	2
Methanoregula boonei	NC_009712	2542943	1*mcrA, 3*nifH	54.5	1.43
Methanolacinia petrolearia	NC_014507	2843290	2*mcrA, 3*nifH	47.4	ND

Organism:	**GC 47% [Rel. Abun.]**	**GC 47% Weighted GC%**	**GC 50% Rel. Abun.**	**GC 50% Weighted GC%**	**GC 53% [Rel. Abun.]**	**GC 53% Weighted GC%**	**GC 57% Rel. Abun.**	**GC 57% Weighted GC%**	**GC 60% Rel. Abun.**	**GC 60% Weighted GC%**	**GC 63% Rel. Abun.**	**GC 63% Weighted GC%**
Nitrosospira multiformis	0.024	1.296	0.034	1.820	0.053	2.801	0.052	2.790	0.022	1.182	0.018	0.977
Nitrososphaera viennensis	0.143	7.429	0.216	11.272	0.267	13.881	0.242	12.597	0.160	8.312	0.077	4.035
Nitrospira defluvii	0.011	0.646	0.043	2.548	0.018	1.087	0.015	0.891	0.032	1.916	0.014	0.816
Ca. Kuenenia stuttgartiensis/hzoA fragment in plasmid vector	0.133	6.103	0.081	3.750	0.090	4.157	0.099	4.560	0.063	2.911	0.066	3.020
Pseudomonas aeruginosa PA96	0.004	0.282	0.008	0.517	0.014	0.934	0.013	0.832	0.026	1.716	0.028	1.830
Escherichia coli	0.096	4.838	0.084	4.273	0.048	2.443	0.023	1.142	0.041	2.051	0.023	1.148
Shigella sonnei strain FC1706	0.016	0.792	0.010	0.491	0.013	0.648	0.009	0.477	0.005	0.277	0.004	0.212
Cupriavidus metallidurans (Ralstonia eutropha) CH34/CCUG 13724	0.012	0.779	0.027	1.720	0.045	2.839	0.048	3.037	0.091	5.803	0.105	6.696
Cupriavidus necator (Alcaligenes eutrophus/Ralstonia eutropha) ATCC 17699	0.019	1.227	0.029	1.895	0.043	2.852	0.054	3.604	0.090	5.991	0.149	9.905
Dyadobacter fermentans DSM 18053	0.170	8.776	0.068	3.490	0.114	5.860	0.099	5.083	0.054	2.796	0.020	1.015
Pseudomonas stutzeri JM300/DSM 10701	0.022	1.418	0.025	1.576	0.045	2.818	0.064	4.054	0.129	8.157	0.134	8.463
Rhodobacter sphaeroides DSM 158/ATCC 17023	0.021	1.463	0.011	0.788	0.045	3.115	0.092	6.337	0.111	7.606	0.165	11.347
Salinibacter ruber DSM 13855	0.021	1.392	0.029	1.928	0.050	3.294	0.117	7.704	0.114	7.554	0.171	11.260
Sulfurimonas denitrificans DSM 1251	0.285	9.849	0.287	9.889	0.123	4.260	0.043	1.475	0.002	0.073	0.008	0.282
Methylosinus trichosporium Ob3p	0.001	0.066	0.005	0.343	0.003	0.188	0.003	0.206	0.006	0.387	0.003	0.206
Methylocella tundraea	0.002	0.157	0.012	0.777	0.007	0.412	0.010	0.627	0.017	1.078	0.010	0.646

Methylomicrobium buryatense 5B	0.018	0.880	0.028	1.346	0.020	0.967	0.016	0.752	0.006	0.308	0.003	0.161
Methanoregula boonei	0.001	0.028	0.001	0.059	0.002	0.103	0.001	0.050	0.019	1.039	0.001	0.057
Methanolacinia petrolearia	0.000	0.011	0.001	0.024	0.001	0.042	0.001	0.040	0.010	0.479	0.000	0.023
sum:	1.00	47.43	1.00	48.51	1.00	52.70	1.00	56.26	1.00	59.64	1.00	62.10

Table S2. Soil properties of used samples in the shotgun vs. targeted analysis and targeted vs. amoA amplicon analysis.

Site	Coordinates	Corg/Norg	pH H_2O	NH_4^+ [µg NH_4^+ -N·g -1 dry soil]	NO_2^- [µg NO_2^- -N·g -1 dry soil]	NO_3^- [µg NO_3^- -N·g -1 dry soil]	Fe II [µmol·g -1 dry soil]	Fe III [µmol·g -1 dry soil]	Ref.:
Hungary, Field # & %	47°31N 16°59E	17.9±5.3	6.7±0.16	0.02±0.004	0.23±0.1	3.4±1.9	1.5±0.07	43.3±4.8	This study
Belfontaine wetland#	46°34N 6°04E	14±0.6	6.9±0.1	0.04±0	0.04±0	0.002±0.001	78.5 ± 43.9	78.5 ± 43.9	Bagnoud et al., 2020
Seida, Bare peat &	67°03 N 62°57 E		3.9±0.09	5.4±0.6	bd	11.4 ±3.9	NA	NA	Siljanen et al., 2019

bd = below detection limit

NA = not analyzed

= shotgun vr. Targeted comparison, Fig. 2,3.

& = Targeted vs. *amoA* amplicon analysis, Fig. 4, S3.

% = targeted vs. *nosZ* amplicon analysis, Fig. S4

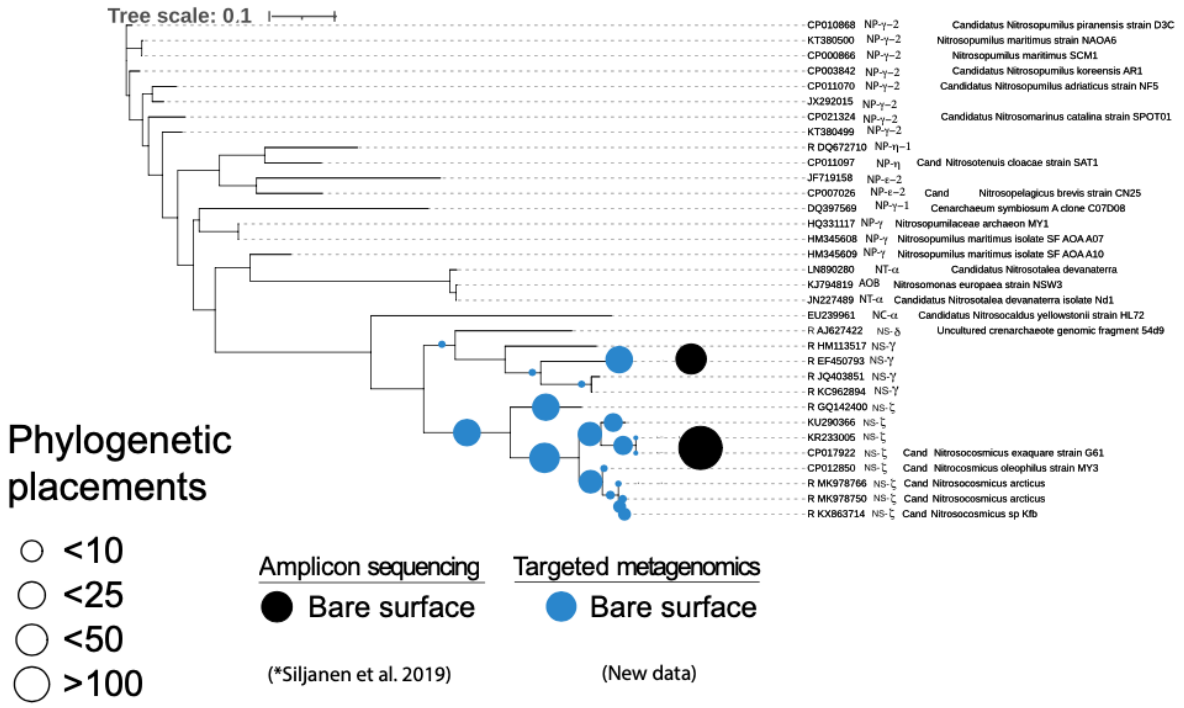


Fig. S3. Comparison of archaeal *amoA* diversity in Arctic permafrost peat with bare surfaces obtained with *amoA* amplicon sequencing and targeted metagenomics. The amplicon sequencing data was generated by Siljanen et al. (2019) and used here with permission from the authors.

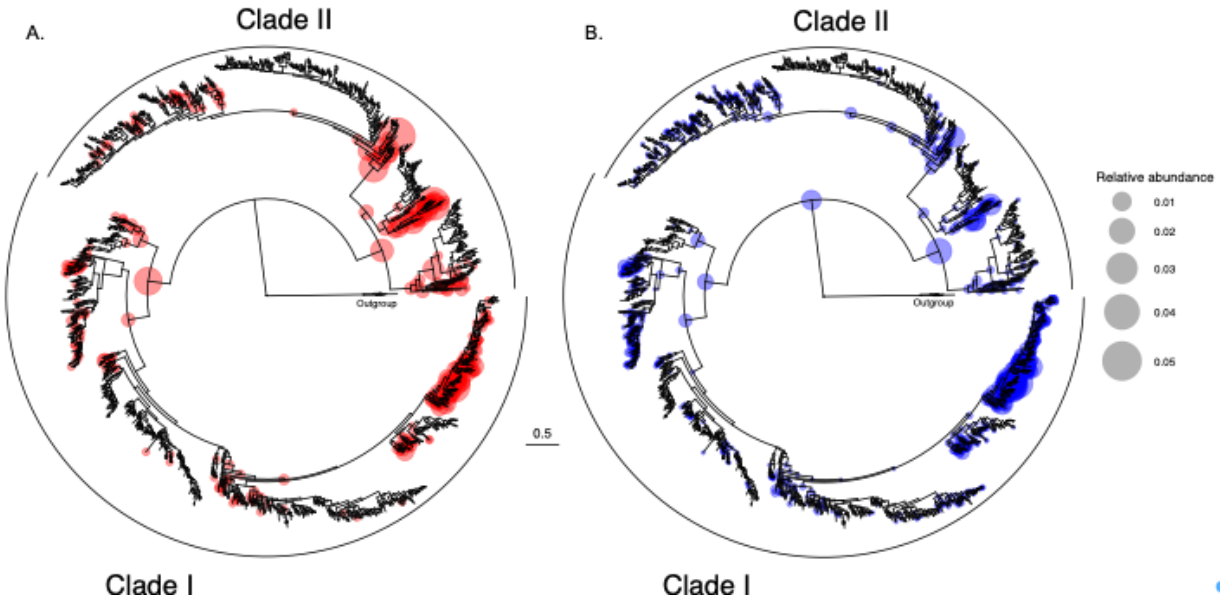
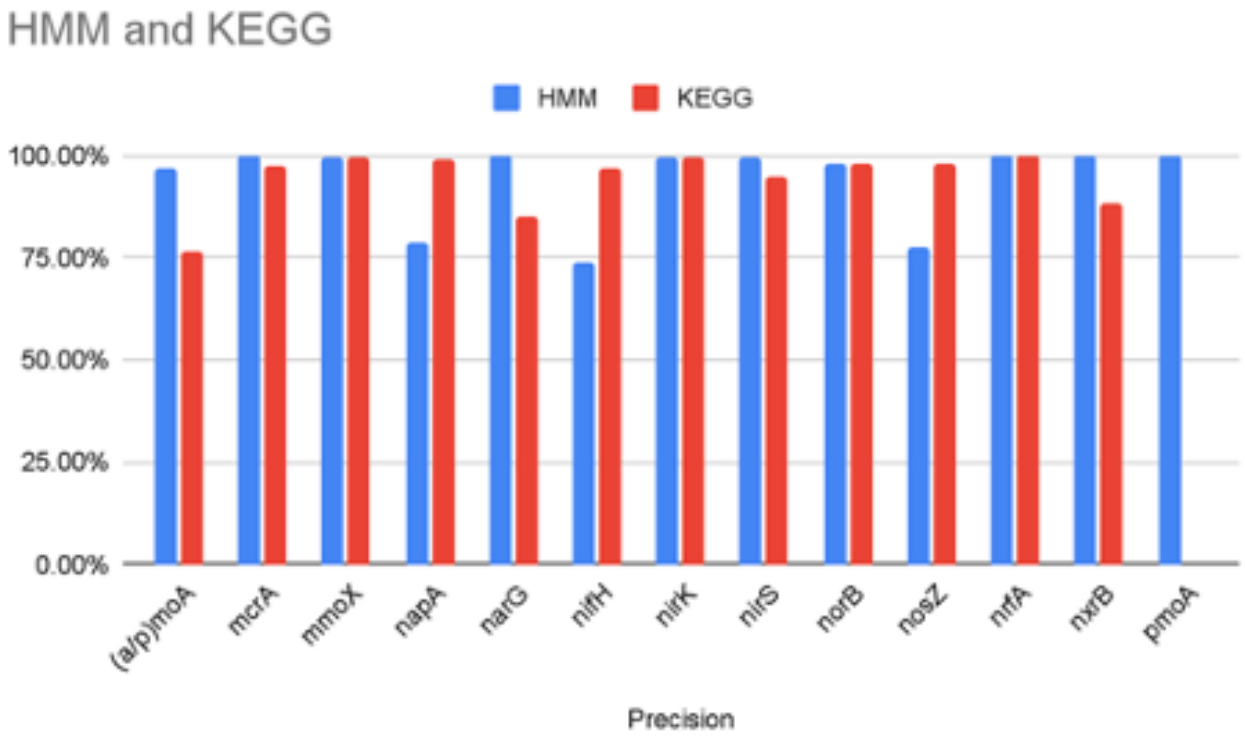


Fig. S4 Phylogenetic placement of *nosZ* reads obtained from agricultural soil using A) shotgun metagenomics and B) targeted metagenomics. The reference phylogeny (Graf et al., 2022) is based on

amino acid sequences analyzed using the LG+R10 substitution model in IQ-TREE, and node symbols indicate the location of placements in the reference tree. Symbol size corresponds to the relative abundance of reads placed at each node, and the scale bar at center indicates branch length in the reference tree. 'Outgroup' denotes distant homologues of *nosZ* with unknown function.



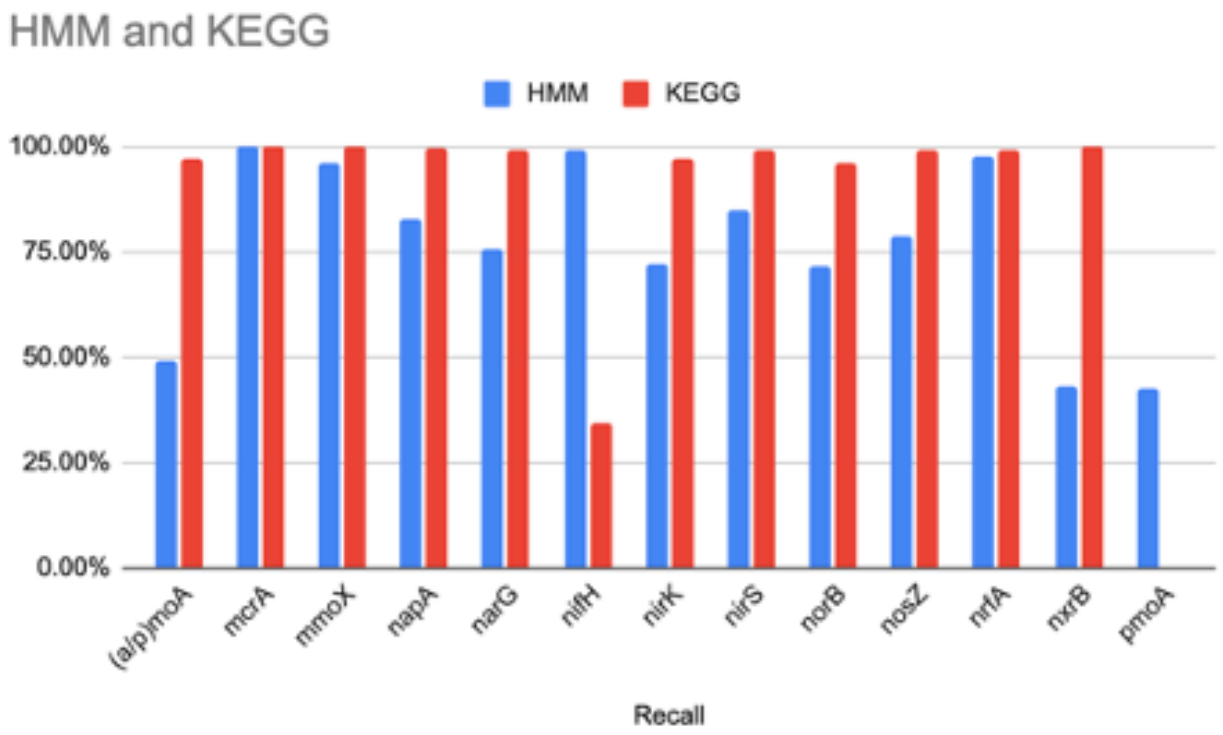


Fig. S4. Comaprison of KEGG annotation and custom HMMs precision and recall for the mock-community comparison with different GC% categories combined.