Supplementary Information

Extraction and preparation of nucleic acids. Nucleic acid extractions (RNA and DNA) were performed using a modification of the method described by Griffiths *et al*.[25] In brief, 0.5 volumes of both CTAB buffer and phenol chloroform isoamylalcohol were added to a lysing matrix E tube (Q-Biogene) containing 0.5 g of soil and cells were lysed in a homogenizer or FastPrep machine (Q-Biogene) at speed 5.5 followed by DNA precipitation with PEG 8000. Different homogenization times were used to optimize DNA extraction. For soils L-l, L-n, L-h, R-nt, R-p, B77T, B77V and E16 the time was set such that a maximum yield was obtained for DNA molecules with sizes >3 kb. (Longer homogenization times resulted occasionally in slightly higher yields but heavily sheared DNA, as measured by NanoDrop® and by scanning SYBRGreen I® -stained gels). For soils RUD, STO, GSF and KRO the procedure was optimized for the yield of archaeal and bacterial *amo*A genes individually. This utilised bead beating times of 30 s for quantification of archaeal *amo*A copies in all four soils and 120 s for RUD and GSF and 150 s for STO and KRO soils for quantification of bacterial *amo*A. Fig. S2 demonstrates the influence of bead beating times on *amo*A gene copy numbers for two different soils.

Quantification of DNA. A modified protocol of Zipper and colleagues[28] was used to quantify extracted DNA. Ten µl of diluted DNA samples was added to 40 µl of a 1,600-fold SYBRGreenI® dilution and incubated for 15 min at 22°C in the dark followed by detection of fluorescence intensity in an Opticon2 real-time Instrument (MJResearch). A standard curve was generated using known amounts of *Escherichia coli* DNA (Sigma Aldrich). Detection limits were 7000 – 70 ng ml^{-1} DNA. All measurements were performed in triplicate.

Real-time PCR. Archaeal *amo*A genes were quantified using a protocol described previously[8] with a dilution series of the linearized soil fosmid clone 54d9 as a standard with slight modifications in primer and probe concentrations (1 µM *amo*196F, 0.5 µM *amo*277R and 1 µM *amo*247) and the addition of 0.2 mg ml^{-1} bovine serum albumin (BSA) to reduce inhibitory effects of co-extracted polyphenolic compounds in soil. The bacterial *amo*A gene

was quantified with SybrGreenI as a fluorescent dye[29, 30] using primers *amo*A-1F*[14] and *amo*A-2R[13]. Each reaction was performed in a 20 µl volume containing 5 or 10 ng soil DNA, 0.2 mg ml^{-1} BSA, 0.5 µM of each primer and 10 µl of QuantiTect SYBR Green PCR Master Mix (Qiagen). Cycling conditions were as follows: 15 min at 95°C, 45 cycles of 1 min at 95°C, 1 min at 54°C, 1 min at 72°C, followed by a plate read, incubation at 76°C to avoid detection of unspecific products, such as primer dimers and a second plate read (which was used for quantification) followed by final extension reaction for 10 min at 72°C. Product specificity was confirmed by melting curve analysis and visualization in agarose gels showing specific product bands at the expected size of ca. 500 bp. Quantification standards consisted of a dilution series of a known amount of a linearized plasmid (pCR4-TOPO, Invitrogen) containing the *amo*A gene of *Nitrosomonas europaea* ATCC 19718*, Nitrosospira multiformis* ATCC25196*, Nitrosospira* NpAV or *Nitrosovibrio tenuis* NV-12, generated with primers A189F[31] and amoA-5R[32]. For AOA *amo*A quantification a standard curve with fosmid 54d9 was generated over seven orders of magnitude from 30 to 3 x 10^8 copies of template per assay; for quantification of AOB *amo*A the template copies varied between 10 and 4 x 10^6 per assay, respectively. High amplification efficiencies of 95-100% were obtained in AOA *amo*A quantification, with R^2 values between 0.990 and 0.997 and slopes between -3.31 and -3.47, and efficiencies of 84-113% were obtained for quantification of AOB *amo*A, with R^2 values between 0.987 and 0.999 and slopes between -3.04 and -3.78. The *amo*A copies determined with the different AOB standards varied maximally by a factor of 2. The mean values of all four standards were used for quantification. While reproducible standard curves were generated with the two qPCR protocols described here, the TaqMan®-based technique published for quantification of AOB *amo*A[16] did not give reproducible results in our hands. All qPCR experiments were conducted in a DNA Engine Opticon2 real time thermocycler (MJResearch) with Opticon Monitor software version 3.01 (BioRad), low profile thermo-strips and ultra clear cap strips (ABgene). Water was used instead of DNA as a non-template control and 5 ng *E. coli* DNA and fosmid 54d9 in AOB measurements or a mixture of the four AOB *amoA* standards used in AOA measurements served as negative controls. Several controls were performed to estimate the possible inhibition of PCR performance by co-

extracted polyphenolic compounds in soil extracts: (i) an *E. coli* plasmid (pUC) was added to DNA extracts of all soils and was quantified by qPCR with primers M13 and M13r according to Horz[30], (ii) different amounts of *amo*A standard were mixed with equal amounts of soil DNA extract and (iii) two different volumes of soil DNA extract from each soil sample used in this study were applied for quantification of *amo*A copies. In most cases, inhibition was not detected or was negligible. Modest inhibition (at most a factor of 2), was observed in a few DNA preparations that were strongly contaminated with humic acids with a dark brown appearance (see Supplementary Fig. S1b).

***Amo*A gene amplification, sequence and phylogenetic analysis.** *Amo*A sequences were amplified with the primers *amo*A19F and *amo*A643R in a PCR protocol as described in the titration PCR section and 25 to 30 cycles of amplification using DAp Goldstar DNA polymerase (Eurogentec). AOA *amo*A PCR products were purified (Qiaquick® PCR Purification Kit, Qiagen) and cloned into the pCR4-TOPO® vector for sequencing (Invitrogen) according to manufacturers instructions. Clones were checked for the right insert by PCR and sequencing was performed on an ABI Prism 3700 DNA Analyser using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). DNA sequences were aligned using ClustalW[33] implemented in BioEdit[34] before translation. Maximum likelihood distances were calculated using the JTT substitution model with site variation (invariable and eight variable gamma rates) in TREE-PUZZLE[35]. Bootstrap support using protein distance and protein maximum-parsimony methods was calculated using the PHYLIP package[36], and LogDet-paralinear DNA distance analysis[37] was performed with PAUP[38] using variable sites[39] (estimated from a maximum likelihood model) after 3rd codon positions were excluded to remove potentially saturated positions[40]. While the clustering of *amo*A protein sequences as shown in Supplementary Fig. S3 should reflect different "ecotypes" of ammonia oxidizers we can not fully exclude the possibility, that some of the sequences stem from yet unrecognized methane-oxidizing archaea, since closely related *amo*A/*pmo*A sequences are e.g. found in γ-proteobacteria.

Most probable number (MPN) PCR. A PCR-based assay using different primer sets specific for AOA and AOB was applied to determine the relative abundance of these groups in soil habitats, complementary to the real-time PCR techniques. A primer set specific for the 16S rRNA gene of betaproteobacterial AOB was used as previously described[31] with the forward primer being an equimolar mixture of CTO189fA, CTO189fB and CTO189fC, (without GC-clamp) and the reverse primer CTO654r. While these primers hybridize to most 16S rRNA genes of ammonia-oxidising betaproteobacteria, they may be biased against sequences of the *N. communis* cluster due to mismatches with primer CTO654r[41]. The primer pair Arch20F/958R[42] was used to amplify archaeal 16S rRNA genes. As these primers are not specific for crenarchaeota a clone library was made to analyse if only crenarchaeota were detected in the two top soils. All 43 sequences of RUD and 27 of L-h were affiliated with crenarchaeota and belonged to the soil group 1.1b lineage as described earlier for RUD soil[18]. AOB *amo*A genes were amplified with forward primers A189F[43] or *amo*A-1F*[14], respectively, and the reverse primer *amo*A-2R[13]. Crenarchaeal *amo*A gene specific primers were *amo*A19F (ATGGTCTGGCTWAGACG) and *amo*A643R[8]. DNA extracted from soil was added in amounts of 10 ng, 1 ng, 100 pg, 10 pg, 1 pg, 100 fg and 10 fg to individual reaction mixtures. Each experiment was performed in triplicate with an independently prepared DNA dilution series. PCR reactions were carried out in a 25 µl volume on a TGradient cycler (Biometra). After an initial denaturation step at 95°C for 10 min, 40 cycles of 95°C denaturation (1 min), 55°C annealing (1 min), 72°C elongation (40 s) were followed by a final elongation step of 10 min at 72°C. Five µl of the amplicons was visualized on agarose gels (1% TAE) stained with ethidium bromide, and the presence of a DNA band corresponding to the expected size was treated as a positive result. The comparable sensitivity of the different primer pairs was confirmed by titration-PCRs using dilution series of identical copy numbers of the soil clone 54d9 (for archaeal 16S rRNA and *amo*A genes) and of the *N. europaea amo*A and 16S rRNA genes, cloned into the TOPO-TA vector, as templates (10^6 to 10^1 gene copies) under the conditions described above. Results of one experimental series are presented in Fig. S4. PCR products of archaeal 16S rRNA genes with (primers 20F and 958R[42]) and *amo*A (19F, this study, and 643R[8]) were obtained repeatedly from 1 pg or 100 fg

of DNA template, while the minimal template that gave products for 16S RNA genes of AOB was 100 pg (primers CTO189f and CTO654r[31]) and 10 to 1 pg for *amo*A genes of AOB (primers A189F[43] or amoA-1F*[14] and amoA-2R[13]), again indicating dominance of AOA over AOB (Supplementary Table S3).

GDGT analysis. Soils were extracted twice with dichloromethane/methanol (1:1, v/v) using Accelerated Solvent Extraction (DIONEX ASE 200) at 65°C and 50 bars. Total lipid extracts were fractionated via solid phase extraction using activated Al_2O_3 (2,000 mg) as stationary phase and hexane/dichloromethane (9:1, v/v) as mobile phase into non-polar and polar lipid fractions. The polar lipid fraction was taken to dryness and then redissolved in hexane/2-propanol, centrifuged at 3,000 rpm for 10 minutes and the supernatant filtered through 0.45 µm PTFE filters.

For detection of tetraether lipids an HPLC-APCI (atmospheric pressure chemical ionisation)-MS method modified after Hopmans *et al.*[21] was used. Chromatographic separation was achieved on a Sphere Image 3 CN column (150 x 2 mm, 3 µm; CS-Chromatography, Düren, Germany). The mobile phase consisted of n-hexane A and isopropanol B (5 min 99% A, 1% B, linear gradient to 1.4% B within 22.5 min, in 1 min to 10% B, holding 5 min to clean the column and back to initial conditions in 1 min, held for 6 min for equilibration). The flow rate was set to 200 µl min^{-1} and injection was performed via an autosampler with a 5 µl loop. APCI source conditions were: corona current 5 µA giving a voltage of around 4 kV; vaporiser temperature 350°C; capillary temperature 200°C and voltage 7.5 V; nitrogen sheath gas at 60 psi (4 bar). Mass chromatograms were generated by selected ion monitoring (SIM) in the positive ion mode resulting in protonated $[M+H]^+$, $[M+H]^+$+1 and $[M+H]^+$+2 molecular ions. Multiplier voltage was set to 1,500 V, and the scan rate was 2 s. For compound identification full scan analysis over an *m/z* range of 900 to 1,400 a.m.u. was used. Quantitative determination of tetraether compounds was achieved by replicate analysis of an external synthetic diether archaeol standard (Avanti Polar Lipids Inc., AL, USA), assuming an identical response factor for the tetraether GDGT. Linearity and reproducibility were tested on three concentrations (0.1, 1, 10 µg ml^{-1}) analysed in triplicate and in parallel to the sample

sequence. Samples were run in duplicate with standard deviation of less than 8%. Integration involved the summed $[M+H]^+$, $[M+H]^+$+1 peak areas for all compounds and standards.

GDGT lipids occurred in significant abundance in the soils studied. Concentrations for total GDGT ranged between 0.14 and 76.42 µg g^{-1} soil and 0.04 to 3.24 µg g^{-1} soil were isoprenoidal GDGT attributed to archaea. Non-isoprenoidal GDGT have been proposed to derive from bacterial input[26]. Two of the ten soils analyzed show a different ratio of GDGT:crenarchaeol and also a much higher total organic carbon (TOC) than all other examined soils, indicating soils with very high organic matter in which a group of archaea that do not contain crenarchaeol might dominate. These two samples have been excluded from the plots shown in Fig. 2 and Fig. S5.

Construction of cDNA library, high-throughput sequencing and bioinformatic analysis. Total nucleic acid extraction and first strand cDNA synthesis were performed as described previously[8], with minor modifications. After DNase treatment, 5 or 10 µl of total RNA was subjected to reverse transcription at 37°C and 50°C, respectively, and the final 70°C incubation step was omitted. Second strand synthesis was performed using *E. coli* RNaseH, DNA ligase and DNA polymerase at 16°C for 2 h according to the manufacturer's instructions (Invitrogen). The size of the double-stranded (ds) cDNA (~1 µg) ranged from 100 to 1200 bp, as judged by agarose gel electrophoresis. Both preparations were combined for high-throughput sequencing on a Roche GS20 (Roche Applied Sciences/454 Life Sciences, Branford, CT).

Library construction was performed by shearing the cDNA into fragments which were blunt-ended and phosphorylated by enzymatic polishing using T4 DNA polymerase, T4 polynucleotide kinase and *E. coli* DNA polymerase. The polished DNA fragments were then subjected to adapter ligation followed by isolation of the single-stranded template DNA (sstDNA). The quality and quantity of the sstDNA library was assessed using the Agilent 2100 Bioanalyzer. Each sstDNA library fragment was captured onto a single DNA capture bead and clonally amplified within individual emulsion droplets. The emulsions were

disrupted using isopropanol, beads without an amplified sstDNA fragment were removed and beads with an amplified sstDNA fragment were recovered for sequencing. The recovered sstDNA beads were packed onto a 70x75mm PicoTiterPlate™ and loaded onto the GS 20 Sequencing System (Roche Applied Sciences/454 Life Sciences, Branford, CT) as previously described[22].

cDNA sequences (314,041) were aligned against the non-redundant nucleotide database[44] and ribosomal database[45] using megablast[46] at default cutoffs, i.e. expectation value of 10 and word size (length of best perfect match) of 28 bp. Keyword search using the classification names of the Taxonomic Outline of Prokaryote Genera[47] of Bergey's Manual was conducted on the BLAST output file. A sequence read was included in our read distribution statistics only when the best hit (highest bit score) could be assigned unanimously to a single taxon or genus.

Three approaches were used to pre-select reads potentially affiliated with crenarchaeota or AOB based on 16S rDNA, (1) The ratio of sequences affiliated with crenarchaeota and AOB were analysed by aligning all reads against a set of four complete crenarchaeal 16S rRNA genes (uncultured crenarchaeote 54d9 [gi:42557759], uncultured crenarchaeote 29i4 [gi:22797874], Candidatus *Nitrosopumilus maritimus* [gi:71668096], uncultured crenarchaeote 4B7 [gi:14548123]) and a set of six 16S rRNA genes covering the major phylogenetic groups of the AOB (betaproteobacteria: *Nitrosomonas europaea* [gi:30248031], *Nitrosomonas communis* [gi:1592803], *Nitrosomonas oligotropha* [gi:11545282], *Nitrosospira multiformis* [gi:82701135], *Nitrosospira* sp. B6 [gi:1236775]; gammaproteobacteria: *Nitrosococcus oceani* [gi:77163561]). A similarity cut-off of 97% was used. (2) The second approach involved alignment of group-specific probes to all reads in an "in silico" hybridization experiment, allowing up to one mismatch. A general archaeal probe (Arch958) and an updated probe specific for the crenarchaeal group 1.1b (Cren1.1b_537: CGACCACTCGGGGTGCTG, modified from probe Cren537[42, 48]) were used to identify archaeal reads. Six published probes targeting the betaproteobacterial AOB were applied, namely CTO189 and CTO654[31], Nsm156, Nso190 and Nso1225[49], and Nsp436[14]. Probes

CTO189 and Nso1225 target members of all betaproteobacterial AOB subgroups without mismatch[41]. The probe Nsm156 targets primarily *Nitrosomonas* species and CTO654, Nso190 Nsp436 primarily *Nitrosospira*[41] with perfect match. However most of the subgroups of betaproteobacterial AOB were covered by the probes in our search mode, i.e. allowing one mismatch. (3) In the third approach the reads were screened with the universal probe Uni1392[27] allowing one mismatch and archaea- and AOB-affiliated sequence reads were extracted from this dataset by key words search as described above. All sequences selected through these three approaches were subsequently verified manually in individual BLASTn searches against the non-redundant database (identified reads from approaches 2 and 3 are given in Supplementary Table S5).

We searched for ***amo*A cDNA** of AOA and AOB by two different approaches similar to those denoted above for 16S rDNA, (1) screening with *amo*A genes of Candidatus *Nitrosopumilus maritimus* [gi:71668108], uncultured crenarchaeote 54d9 [gi:42557759], *N. europaea* [gi:3252912], *N. oceani* [gi:2104719], *N. communis* [gi:11545288], *N. oligotropha* [gi:11545302], *N. multiformis* [gi:1085094] and *Nitrosospira* sp. B6 [gi:18996142] and (2) using the *amo*A-specific oligonucleotides and probes from the qPCR and MPN PCR. Identified sequence reads were verified as described above. The two verified reads from archaea are shown in Supplementary Table S5.