ISME Communications, 2025, 5(1), ycaf087
https://doi.org/10.1093/ismeco/ycaf087
Advance access publication: 23 May 2025
Original Article

# Predicting gene distribution in ammonia-oxidizing archaea using phylogenetic signals

Miguel A. Redondo[1,2,3,4,*], Christopher M. Jones[1], Pierre Legendre[2], Guillaume Guénard[2], Sara Hallin[1]

[1]Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Box 7026, 750 07 Uppsala, Sweden
[2]Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada
[3]National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Husargatan 3, 752 37 Uppsala, Sweden
[4]Department of Cell and Molecular Biology, Uppsala University, Husargatan3, 752 37 Uppsala, Sweden

*Corresponding author: National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Husargatan 3, 752 37 Uppsala, Sweden.
E-mail: miguel.angel.redondo@nbis.se.

## Abstract

Phylogenetic conservatism of microbial traits has paved the way for phylogeny-based predictions, allowing us to move from descriptive to predictive functional microbial ecology. Here, we applied phylogenetic eigenvector mapping to predict the presence of genes indicating potential functions of ammonia-oxidizing archaea (AOA), which are important players in nitrogen cycling. Using 160 nearly complete AOA genomes and metagenome assembled genomes from public databases, we predicted the distribution of 18 ecologically relevant genes across an updated *amoA* gene phylogeny, including a novel variant of an ammonia transporter found in this study. All selected genes displayed a significant phylogenetic signal and gene presence was predicted with an average of >88% accuracy, >85% sensitivity, and >80% specificity. The phylogenetic eigenvector approach performed equally well as ancestral state reconstruction of gene presence. We implemented the predictive models on an *amoA* sequencing dataset of AOA soil communities and showed key ecological predictions, e.g. that AOA communities in nitrogen-rich soils were predicted to have capacity for ureolytic metabolism while those adapted to low-pH soils were predicted to have the high-affinity ammonia transporter (*amt2*). Predicting gene presence can shed light on the potential functions that microorganisms perform in the environment, further contributing to a better mechanistic understanding of their community assembly.

**Keywords:** phylogenetic modeling; trait imputation; elastic net regularization; phylogenetic eigenvectors; ancestral state reconstruction

## Introduction

Genomic information is essential for indirect trait-based approaches in microbial ecology, which can provide mechanistic explanations to ecological dynamics and ecosystem functions [1]. Microbial enzyme-encoding genes are often associated with the organismal functional traits and tend to be phylogenetically conserved [2, 3], which provides a foundation for phylogeny-based trait prediction [4–6]. We can therefore use massive amounts of environmental sequencing data to predict the probability of presence of certain microbial genes, thus inferring the potential functions that microorganisms perform in the environment. In this way, we can functionally characterize taxa whose genomes are not yet available or that represent a minor fraction in metagenomes. Phylogeny-based imputations of traits has generally used either ancestral state reconstruction by means of phylogenetic generalized least squares [6–8] or phylogenetic eigenvector maps [9–11]. In contrast to ancestral state reconstruction, phylogenetic eigenvectors offer the additional advantage of accommodating different modes of evolution, as well as the possibility of using phylogenetic signals together with abiotic factors when predicting traits or species distributions [12]. While phylogenetic eigenvectors have been used in a wide range of studies for macroorganisms, they have yet to be applied to microbial communities.

Among microorganisms, ammonia-oxidizing archaea (AOA) are an optimal group to evaluate the power of phylogenetic eigenvectors for predicting gene distribution. First, they are key players in the nitrogen cycle and inhabit most ecosystems on earth [13, 14]. Given their ecological relevance, AOA genomes and metagenome-assembled genomes (MAGs) are increasingly available, thus expanding our knowledge of the potential functions of archaeal genes [15–17]. Second, there is a coherence between the organismal phylogeny and that of the *amoA* gene encoding the ammonia monooxygenase subunit A, which has long been used as a marker gene for AOA in environmental studies [13, 18–21]. This has resulted in the availability of a global *amoA* phylogeny [21] that reflects the distribution of the organism across different earth environments. This adds to previous studies pointing at the niche specialization of certain AOA clades across varying levels of pH and other environmental properties [22–24]. Whether or not gene content can be predicted using the *amoA* phylogeny, thus providing a basis for a more mechanistic understanding of AOA community assembly, remains uncertain.

The aim of this work was to predict gene presence/absence in AOA using the *amoA* phylogenetic signal. To reach this goal,

**Received:** 10 October 2024. **Revised:** 7 March 2025. **Accepted:** 21 May 2025

we updated a recent *amoA* gene reference phylogeny of AOA [21] by adding 160 highly complete AOA genomes and MAGs available in public databases. The updated phylogeny was then used together with phylogenetic eigenvector mapping [11] to predict the presence of a set of genes (Table 1) belonging to four functional categories selected from a comparative genomics study [16]. We validated the predictions using hold-out validations and compared them to estimations based on ancestral state reconstruction. Finally, we implemented the predictive approach on soil AOA communities obtained from a field study [25] and linked the predicted AOA gene presence to soil properties by means of simultaneous analysis of environmental characteristics, species distributions, and species traits [26]. This study demonstrates that phylogenetic eigenvector maps are useful for highly accurate predictions of gene distributions in AOA and can inform about their potential functions in the environment and the mechanisms underpinning community assembly.

## Materials and methods

### Update of the archaeal *amoA* reference phylogeny

To update the *amoA* phylogeny developed by Alves *et al.* [21], we first downloaded all genomes from isolates and MAGs from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and the Joint Genome Institute (JGI; https://genome.jgi.doe.gov/portal/), up to October 2021 using the search terms "Thaumarchaeota" and "Crenarchaeota," since the search was done prior to the current taxonomy of the phylum harboring AOA (Nitrososphaerota). Sequences of 1527 archaeal genomes and MAGs were initially screened for the presence of *amoA* genes using HMMER (http://hmmer.org) and the translated alignment from Alves *et al.* [21] as seed alignment. Significant hits (*e*-value $<10^{-6}$) were then aligned by amino acid to the original seed alignment using HMMER. An initial phylogeny of the significant hits, together with the 1190 sequences from Alves *et al.* [21] was generated from the nucleotide alignment using FastTree 2.1 [27] to ensure the correct identification of AOA *amoA* genes. It is important to note that we used a final dataset of 1190 sequences that Alves *et al.* [21] obtained as a result of excluding rogue sequences from their initial alignment of 1206 sequences. The alignment and tree were then inspected using the ARB software [28] to correct alignment errors and remove fragmented or poor quality (i.e. multiple "N"s) sequences, resulting in a total of 457 genomes and MAGs with *amoA* gene (Supplementary Table S1). We further discarded genomes/MAGs with <80% completeness or >5% contamination as determined by BUSCO [29] (archaea_odb10 database), as well as those with identical *amoA* sequences and the same presence/absence values of the selected genes (see below) to remove duplicate taxa. We further tested whether the *amoA* sequences were chimeras by using the reference-based chimera detection algorithm implemented in VSEARCH (v2.3.4) [30] software using the dataset of Alves *et al.* [21] as a reference and deleted 8 genomes/MAGs having chimeric *amoA* sequences. At the end, *amoA* sequences from 160 high-quality genomes and MAGs were added to the original alignment of 1190 sequences in Alves *et al.* [21], and the total alignment was used to build maximum likelihood phylogenies using the IQTree software [31], version 2.1.3. The tree search was carried out using three rounds of 20 independent tree searches, in which perturbation strength settings (*–pers* parameter in IQTree) of 0.1, 0.5, and 1.0 were used for each round, respectively. The resulting trees were checked individually, and the tree with both the highest likelihood and coherence with the Alves *et al.* [21] topology was selected as the final reference tree. Automatic model selection [32] resulted in GTR + F + R9 being selected as the best substitution model, and node support was determined by ultrafast bootstrap approximation and SH-aLRT tests using 1000 replicates for each support metric [33]. During the tree search one sequence from Alves *et al.*, "NC-Alpha-OTU2" was the same as two other *amoA* sequences in our high-quality genomes and was thus removed by IQ-tree. Therefore, the final phylogenetic tree contained 1189 *amoA* sequences from Alves *et al.* [21] and the 160 *amoA* sequences from the high-quality genomes and MAGs we added. The display of the phylogenetic trees was produced using iTOL [34] and the "ggtree" package in R [35].

### Selection of genes for predictions and screening of genomes

We selected 18 genes for phylogenetic modeling (Table 1) based on the comparative genomic study by Kerou *et al.* [16]. These genes were (i) distributed across different AOA lineages to avoid genes only present in an isolated clade within the phylogeny and (ii) involved in pathways related to different ecological functions and especially relevant to soil AOA. The selected genes belonged to four categories: nitrogen metabolism (N-metabolism in figures) (*amt* and *ureC* genes), carbon and amino acid metabolism (C/AA-metabolism in figures) (*metE*, *proDH*, *rocA* genes), chemotaxis and motility (*cheA*, *cheY*, *flaK*, *flaI*, *tadC* genes), and environmental adaptation (*ipct, nhaP*, *trk,* and *cspC* genes). We downloaded the orthologous gene protein alignment from each of these genes from the EggNOG 5.0 database (http://eggnog5.embl.de) using the arCOG ID reported by Kerou *et al.* [16]. When more than one arCOG was provided, we used each of them separately. With the alignments obtained from the EggNOG database, the 160 AOA genomes/MAGs were screened for the presence/absence of each gene using HMMER (*e*-value <0.01). To ensure that gene hits were accurate, we retrieved the protein sequences found on the genomes/MAGs, aligned them with the reference protein alignment from the EggNOG database, and constructed phylogenies using the fastTree2 software [27], version 2.1.11. The location of the hits relative to the reference sequences within the phylogenies was examined to make sure that they were in fact hits of the searched genes and not artifacts or other homologs with different functions.

In the specific case of the ammonium transporter gene (*amt*), and due to its ecological relevance, the gene phylogeny was used to further classify the hits as Amt2 (high-affinity) or Amt1 (low-affinity) type of transporter [36–38]. The hits falling together with *Nitrosocosmicus* taxa in the phylogeny were assigned the low-affinity transporter type (Amt1), as all sequenced *Nitrosocosmicus* taxa encode uniquely one low-affinity Amt [39–42]. It should be noted here that other studies have used the reverse nomenclature, in which Amt1 and Amt2 are defined as high- and low-affinity transporters, respectively [43–45]. To clarify these inconsistencies in the nomenclature, we screened the Amt hits of our study for the primers used by Nakagawa and Stahl [37] for both types of transporters and found that the *amt2* primers from Nakagawa and Stahl [37] matched the genes that were defined as *amt1* in Offre *et al.* [43] and vice-versa. Therefore, we refer here to the gene encoding the high-affinity transporter as *amt2* [36, 37], which corresponds to the *amt1* in Offre *et al.* [43]. The procedure of checking the *amt* hits using phylogenies allowed us to find a third type of ammonium transporter uniquely present in the *Nitrosocaldales* (*NC*) lineage (Supplementary Text S1; Supplementary Figs S1 and S2; and Supplementary Table S2). The gene names and arCOG references are provided in Table 1.

**Table 1.** Genes selected for phylogenetic modeling, statistics for phylogenetic signal test, and validation results using phylogenetic eigenvectors and ancestral state reconstruction.

| Gene | arCOG ID | Functional category (Kerou *et al.*, [16]) | Functional annotation/product (Kerou *et al.* [16]) | Phylogenetic signal | | | Validation of gene predictions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Phylogenetic eigenvectors | | | Ancestral state reconstruction | | |
| | | | | $D$[a] | $P(D \neq 1)$[b] | $P(D \neq 0)$[b] | Accuracy (%) | Sensitivity (%) | Specificity (%) | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| *amt2* | arCOG04397 | N-metabolism | Ammonia permease | 0.17 | **<.001** | .174 | 93.3 | 99.1 | 55.8 | 91.3 | 96.2 | 57.7 |
| *amt1* | arCOG04397 | N-metabolism | Ammonia permease | 0.06 | **<.001** | .328 | 88.4 | 92.7 | 84.4 | 89.6 | 93.6 | 86.0 |
| *amt-NC*[c] | arCOG04397 | N-metabolism | Ammonia permease | −0.38 | **<.001** | .899 | 98.0 | 81.2 | 98.9 | 97.6 | 69.6 | 99.4 |
| *ureC* | arCOG00698 | N-metabolism | Urea amidohydrolase (urease), subunit alpha | 0.36 | **<.001** | **.003** | 76.8 | 76.6 | 78.7 | 77.8 | 81.9 | 73.0 |
| *metE* | arCOG01876 | C/AA metabolism | Methionine synthase II (cobalamin-independent) | −0.36 | **<.001** | .986 | 96.3 | 88.0 | 98.3 | 97.6 | 88.0 | 99.9 |
| *metE2* | arCOG01877 | C/AA metabolism | Methionine synthase II (cobalamin-independent) | −0.29 | **<.001** | .957 | 95.0 | 85.3 | 97.4 | 96.5 | 84.7 | 99.3 |
| *proDH* | arCOG06322 | C/AA metabolism | Putative proline dehydrogenase | −0.08 | **<.001** | .713 | 88.7 | 89.1 | 88.6 | 91.3 | 90.2 | 92.7 |
| *rocA* | arCOG01252 | C/AA metabolism | 1-Pyrroline-5-carboxylate dehydrogenase | −0.14 | **<.001** | .814 | 94.2 | 97.3 | 82.4 | 94.0 | 97.6 | 80.6 |
| *cheA* | arCOG04403 | Chemotaxis/motility | Chemotactic sensor histidine kinase CheA | 0.11 | **<.001** | .225 | 84.5 | 76.1 | 89.2 | 88.4 | 78.2 | 93.3 |
| *cheY* | arCOG02391 | Chemotaxis/motility | Chemotaxis response regulator CheY | 0.05 | **<.001** | .364 | 88.5 | 94.3 | 70.8 | 88.9 | 95.4 | 69.4 |
| *cheY2* | arCOG02589 | Chemotaxis/motility | Chemotaxis response regulator CheY | 0.29 | **<.001** | **.009** | 79.1 | 88.3 | 66.3 | 82.7 | 91.7 | 70.3 |
| *tadC* | arCOG01808 | Chemotaxis/motility | Pilus assembly protein TadC | 0.14 | **<.001** | .131 | 82.1 | 84.8 | 80.5 | 82.3 | 84.9 | 80.4 |
| *flaK* | arCOG02298 | Chemotaxis/motility | Putative archaeal preflagellin peptidase FlaK | 0.09 | **<.001** | .218 | 85.5 | 87.1 | 84.8 | 84.6 | 88.8 | 81.5 |
| *flaI* | arCOG01817 | Chemotaxis/motility | ATPase involved in archaellum/pili biosynthesis | 0.09 | **<.001** | .247 | 86.1 | 91.5 | 81.2 | 85.3 | 89.8 | 81.0 |
| *ipct* | arCOG00673 | Environmental adaptation (osmotic regulation) | Bifunctional CTP:inositol-1-phosphate cytidylyltransferase/di-myo-inositol-1, 3-phosphate-1-phosphate synthase | 0.62 | **.004** | **.001** | 79.7 | 31.5 | 87.0 | 85.1 | 34.5 | 92.6 |
| *nhaP* | arCOG01962 | Environmental adaptation (osmotic regulation) | NhaP-type $Na^+/H^+$ and $K^+/H^+$ antiporter with a unique C-terminal domain/Cell volume regulation protein A | 0.15 | **<.001** | .197 | 89.8 | 95.9 | 50.5 | 90.5 | 96.4 | 52.8 |
| *trk* | arCOG04145 | Environmental adaptation (osmotic regulation) | Trk-type $K^+$ transport system, membrane component | −0.21 | **<.001** | .938 | 93.7 | 93.9 | 93.1 | 95.3 | 96.1 | 94.0 |
| *cspC* | arCOG02983 | Environmental adaptation (thermoadaptation) | Cold shock protein, CspA family | −0.21 | **<.001** | .957 | 92.0 | 96.5 | 88.6 | 92.8 | 95.2 | 90.6 |
| All genes (mean) | | | | | | | 88.4 | 86.1 | 82.0 | 89.5 | 86.3 | 83.0 |

[a] $D$ is the phylogenetic signal strength parameter, where values <0 indicates phylogenetic conservatism and values close to 1 indicate random distribution of gene value across the phylogeny. [b] $D \neq 1$ means departure from a random distribution, whereas $D \neq 0$ means departure from a Brownian motion. [c] *amt-NC* refers to a type of ammonia transporter gene that was found in this study to be present only in the *Nitrosocaldales* lineage (NC). Bold indicates $P$-value < 0.05.

## Test of phylogenetic signal for selected genes

We tested the phylogenetic signal of each gene using the phylo.d function from the caper package [46], running 1000 permutations. This method determines the strength and the statistical significance of the phylogenetic signal for binary traits, i.e. in this study binary trait refers to the presence/absence of a specific gene, compared to random and Brownian motion distributions of the trait [47]. The resultant parameter ($D$) equals 1 when the binary trait has a random distribution across the phylogeny, and 0 if under Brownian motion. $D$ values can be <0 if the trait distribution is more clustered than expected under Brownian motion, and >1 if it is more overdispersed than expected by random. The phylo.d function tests $D$ for significant departure from 1 when the trait is phylogenetically conserved, and a significant departure from 0 when the trait does not follow a Brownian evolutionary model.

## Prediction of gene presence/absence using phylogenetic eigenvectors and comparison to ancestral state reconstruction

The input for the predictions of gene presence was the final *amoA* phylogenetic tree and the matrix of presence/absence of the 18 genes across the 160 AOA genomes/MAGs. We then used phylogenetic eigenvectors obtained from the phylogenetic tree and compared the prediction results to those obtained by ancestral state reconstruction. In both cases, we predicted the probability of the presence of each gene in the taxa across the *amoA* phylogeny that had unknown genomes.

The phylogenetic eigenvector-based predictions were done in three steps: (1) decompose the phylogenetic tree into eigenvectors, (2) fit and regularize individual predictive models for each gene using the eigenvectors as the descriptors, and (3) estimate the presence/absence of the genes from the models of Step 2 on the taxa with unknown genomes given their locations in the phylogeny. To decompose the phylogenetic tree into eigenvectors (Step 1), we used R package MPSEM [48]. The input for the MPSEM package is the phylogenetic tree containing all tips, i.e. taxa with and without gene information. The MPSEM package calculates an influence matrix using only the tips of the tree for which there is information about presence/absence of genes by pruning from the tree the taxa without gene information. From the influence matrix, phylogenetic eigenvectors were obtained by singular value decomposition. When calculating the phylogenetic eigenvectors, we fixed argument $a = 0$ to assume a Brownian motion evolution of all genes. In Step 2, the phylogenetic eigenvectors were then used as fixed factors in a multiple logistic regression whose coefficients were regularized using elastic net regularization. For model regularization, we used the R package glmnet [49] with arguments family = "binomial" and $\alpha = 0.5$ to use same amounts of both $L_1$ (LASSO) and $L_2$ (ridge regression) shrinkage. The penalization hyper-parameter ($\lambda$) was tuned using leave-one-out cross validation within the training dataset and choosing the $\lambda$ value that provided the highest accuracy of the predictions. The outcome of Step 2 is a regularized model that predicts the probability of gene presence given the eigenvectors' values. We classified the probabilities of the predictive model into presence or absence by choosing a threshold that maximizes both true-positive (sensitivity) and true-negative rate (specificity) of the predictions. For this, we created ROC curves using the R package pROC [50] with the function roc and used the function coords to select a threshold for classification that would render the highest Youden's $J$ statistic, where $J$ = sensitivity + specificity − 1. Probabilities that had values equal to or greater than the selected threshold were classified as presence, whereas probabilities lower than the threshold were classified as absence. The final output of Step 2 is a model with tuned $\lambda$ and classification threshold parameter that predict the gene presence for a taxon given its phylogenetic eigenvector scores. We then obtained the eigenvector scores for taxa with unknown genomes and used them on the model of Step 2 to predict the gene presence. Function getGraphLocations from the MPSEM package places the taxa with unknown genomes in the initial influence matrix of Step 1 and function Locations2PEMscores obtains the phylogenetic eigenvector scores. It is important to note that the influence matrix and phylogenetic eigenvectors are not obtained using all tips of the tree, but only the ones with known gene information. For taxa with unknown information, we calculated the eigenvector score values, which are projections of new influence matrix coordinates on the eigenvectors obtained from the initial influence matrix (see Guénard *et al.* [11], for details on this procedure). Thus, the number of phylogenetic eigenvectors of the training dataset, and therefore the number of coefficients of the predictive model, is independent of the number of taxa to be predicted. We provide the code that can be used for analysis of other datasets by providing the phylogenetic tree and its associated presence/absence table with and without missing values in the input data.

The ancestral state reconstruction was done with R package picante R [51]. We used the phyEstimateDisc function to predict the genes of the AOA taxa with unknown genomes. In this procedure, for each taxon with unknown gene presence data, the phylogenetic tree is rerooted on the most recent ancestor common to the unobserved taxon and the rest of the phylogeny. The gene presence or absence of the unobserved taxon is then estimated from the ancestral state reconstruction of the root of the rerooted phylogeny [7, 52]. The function phyEstimateDisc provides a trait state, i.e. presence or absence, for a given threshold (default = 0.5), as well as a value for the statistical support of the state.

## Validation of the predictive models

We validated the predictions of each gene using a 20% hold-out validation. This procedure consists of randomly removing 20% of the initial dataset before creating the predictive model and validating it on the removed samples. We repeated this procedure 500 times, always randomizing the taxa included in the held-out dataset. For each gene, we obtained the mean accuracy, sensitivity, and specificity of the prediction. We also varied the proportion of taxa to be held out in the validations to 30% and 40% and obtained accuracies of 87.1% and 86.4%, respectively (compared to >88% accuracy at 20% hold out). We could not increase the proportion of data to be held out because many genes were class unbalanced. Both sensitivity and specificity were obtained using R package pROC.

We validated the accuracy of the predictions at the genome/MAG level using leave-one-out cross validation, in which we deleted each genome/MAG from the dataset in turn and used the rest of the genomes to predict the gene content of the previously removed genome or MAG.

## Implementation of predictive modeling on natural communities: a case study

To illustrate how predicting AOA gene distribution could link community composition, potential functions, and environmental properties, we used data from a previous study that characterized AOA communities by *amoA* amplicon sequencing on 50 sampling points across an agricultural area in which soil properties were measured (see Enwall *et al.* [53] and Jones and Hallin [25], for more information). The study site is a 44 ha farm divided into 14 fields, with sampling points taken at 51 locations throughout the fields

based on environmental gradients identified in a previous study [54]. We deleted one of the locations (S17) because it lacked data on soil properties. We predicted the presence/absence of the 18 genes on the AOA communities and linked the gene composition with the soil properties that were reported by Enwall *et al.* [53].

To predict the gene presence of the AOA members of the soil communities, we used the representative sequences of the operational taxonomical units (OTUs) (162 in total) from the study of Jones and Hallin [25]. In that study, OTUs were obtained by clustering the postprocessed reads at 97% nucleotide similarity using the UPARSE algorithm [55]. We then placed the *amoA* sequences of each OTU on the reference phylogeny using the EPA-ng [56] (accessed at https://github.com/pierrebarbera/epa-ng) and gappa software [57] (accessed at https://github.com/lczech/gappa). The output of these analyses is a phylogenetic tree containing all sequences of our reference phylogeny and the representative OTU sequences to be predicted as grafted leaves. We used this phylogenetic tree and grafted leaves to implement the phylogenetic eigenvector modeling described above. The output was a matrix with presence/absence of all 18 genes for each OTU in the study.

We studied the link between the predicted genes (Q matrix) and the soil properties (R matrix) mediated by community composition (L matrix) by performing a RLQ analysis followed by a univariate fourth-corner analysis [26, 58, 59]. RLQ is an ordination method that displays the covariation of traits, i.e. the predicted gene presence, and environmental properties providing site and species (hereinafter OTUs) scores, and a global test for significance. The fourth-corner correlations, on the other hand, provide tests of single associations between predicted gene presence and environmental properties. The two methods are complementary and can be performed sequentially. For the RLQ analysis and following Dray *et al.* [26], we calculated three separate ordinations using the ade4 package for R [60]: (1) a correspondence analysis for the community data, i.e. OTU table, using the function dudi.coa; (2) a combination of principal component analysis and multiple correspondence analyses using the environmental data matrix, i.e. soil properties, after standardization using the function dudi.hillsmith; and (3) a principal component analysis for the predicted genes presence/absence without standardization using the function dudi.pca. The three ordinations were analyzed together using the rlq function of the ade4 package for R [60]. The RLQ analysis was tested for significance using the randtest function of the ade4 package with argument modeltype = 6 for the permutation test. This model performs two sequential permutational tests, a first one testing the link between OTUs distribution and environmental conditions (Model 2), and a second testing the link between OTUs distribution and predicted gene presence (Model 4). When both tests are significant, the highest of the two *P*-value provides the statistical significance for the global tests of association between gene distribution and environment [61].

To test which specific gene was associated with each soil property, we performed a fourth-corner analysis using the fourth-corner function from the ade4 package, with modeltype = 6, 99 999 permutations, and "false discovery rate" as the multiple testing correction method for the *P*-value.

# Results

## Congruence between *amoA* phylogeny and content of genes for predictions

To perform the phylogenetic eigenvector-based predictions, we first updated the archaeal *amoA* gene reference phylogeny from Alves *et al.* [21], which now contains 1349 unique *amoA* sequences, including those from 160 highly complete AOA genomes of isolates and MAGs distributed across most *amoA* lineages (Supplementary Fig. S3). After screening the genomes and MAGs for the presence/absence of the genes selected for modeling (Table 1), we found that *Nitrososphaerales* (*NS*) taxa overall lacked genes associated with motility (*tadC*, with the exception of *NS-α*), osmotic regulation (*trk*), and thermoadaptation (*cpsC*). The *metE* gene, responsible for methionine synthesis in energy-limiting environments [16], was found in *Nitrosopumilales* (*NP*) clades associated to deep sea waters (*NP-α* and *-θ*), as well as in *NC* and *Nitrosotaleales* (*NT*) (Fig. 1A). Genes encoding the high- and low-affinity ammonium transporters (Amt2 and Amt1, respectively [36, 37], see details in Methods) were identified based on their positions within a phylogenetic tree of translated Amt sequences [43]. When doing this, we identified a gene encoding a novel variant of the ammonia transporter protein specifically associated with the *NC amoA* lineage, hereafter referred to as Amt-NC. All taxa having the *amt-NC* gene also had the *amt2* (Fig. 1A). We found two subgroups of the ammonium transporter gene *amt-NC*, hereinafter *amt-NC.1* and *amt-NC.2* (Supplementary Fig. S1), which corresponded to the groups into which Luo *et al.* [17] divided the *NC* lineage based on the concatenation of 122 archaeal genes (Supplementary Text S1). The amino acid composition of the Amt-NC was identical to the Amt types described by Offre *et al.* [43] at the ammonium-binding sites, i.e. they contained the same histidine lining the transporter pore [43] and had the same amino acids in several conserved loci (Supplementary Fig. S2). However, it differed in several loci across the regions described by Offre *et al.* [43] (Supplementary Fig. S2). Our findings of a potentially novel variant of the ammonia transporter are solely based on phylogenetic alignment and differences in amino acid composition. Future studies should validate its function and examine if there are functional differences between the proteins encoded by the *amt-NC* and *amt1* and *amt2* genes.

All selected genes were phylogenetically conserved (Table 1). The distance between the 160 isolate genomes and MAGs in a principal component ordination including all 18 genes reflected the lineage classification of the AOA taxa (Fig. 1B), supporting a coherence between the *amoA* phylogeny and the overall gene content of available genomes and MAGs.

## Phylogeny-based predictions of selected genes

We used the gene content and phylogenetic relatedness of the 160 genomes and MAGs to build predictive models of gene presence. For each of the 18 genes, we used elastic net regularized regressions, with the phylogenetic eigenvectors as predictors and the presence of each gene as response. The models with optimized $\lambda$ penalization and classification threshold parameters (Supplementary Table S3) were then used to predict gene presence across the *amoA* reference phylogeny using the phylogenetic eigenvector scores of unobserved taxa as input. The predicted presence of most genes varied across and within *amoA* lineages (Supplementary Table S4). For example, the gene encoding the high-affinity ammonia transporter (*amt2*) was predicted to be present in nearly all lineages except *NS-ζ*. In contrast, the low-affinity ammonia transporter (*amt1*) and urease (*ureC*) genes were predicted to be within most *NS* clades yet were more unevenly distributed or absent across *NT* and *NP* clades (Fig. 2; Supplementary Tables S4 and S5). Regarding genes involved in carbon and amino acid metabolism, the gene responsible for B-12 independent methionine synthesis (*metE*) was predicted to be present in all *NC* and in most *NT* taxa. Across *NP* clades, *metE*

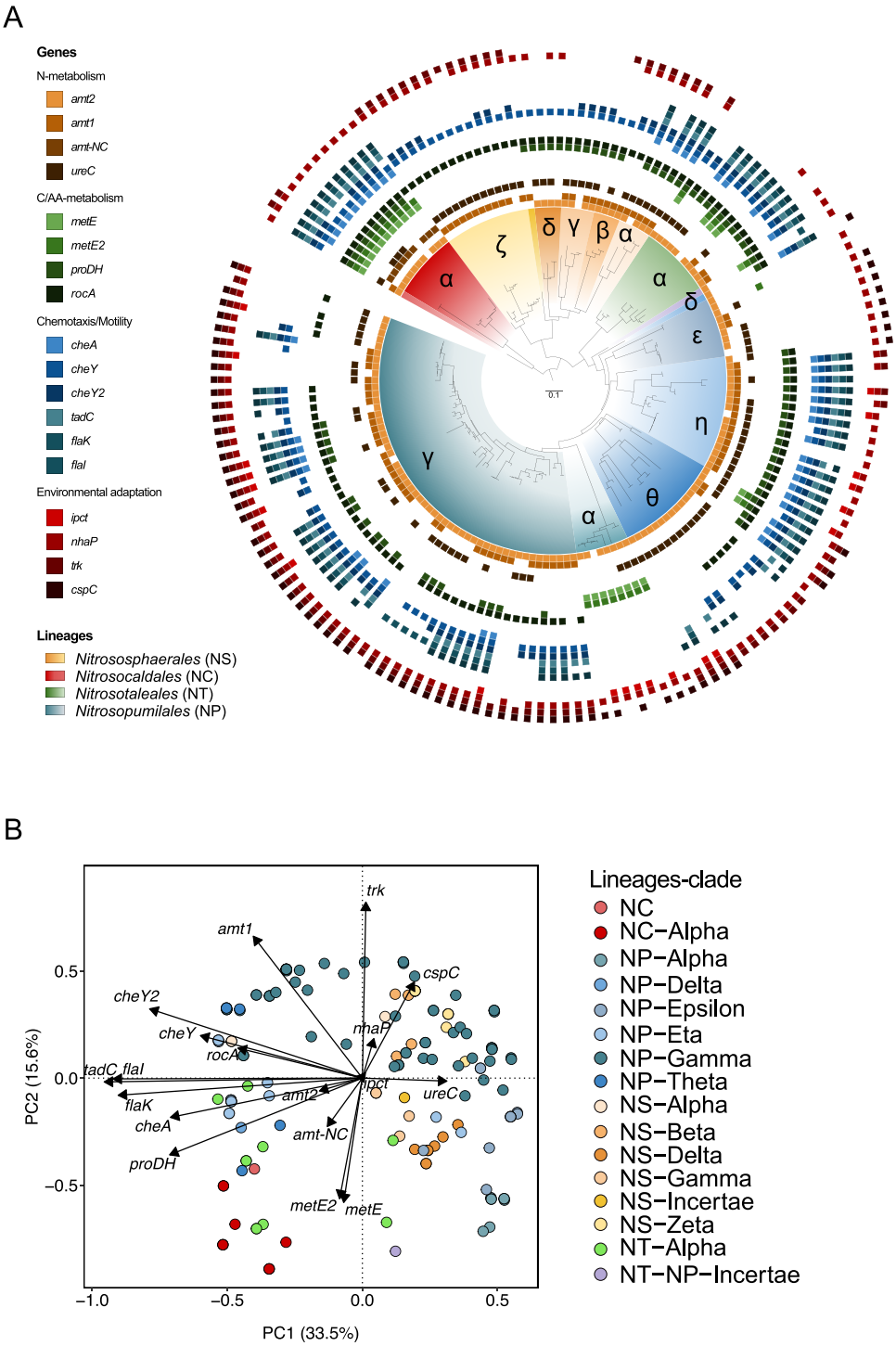


**Figure 1.** Distribution of the 18 selected genes in AOA genomes. (A) The presence/absence of 18 ecologically relevant genes (Table 1), across a phylogeny of 160 AOA isolate genomes and MAGs. Outer rings depict presence (filled squares) and absence (no squares) of specific genes. Clades within each AOA lineage are denoted by Greek letters. The phylogeny of 160 genomes is the result of pruning taxa that lack genomic information from the updated reference phylogeny. For bootstrap node support of the lineages of the *amoA* updated phylogeny, see Supplementary Fig. S3. (B) Principal component analysis of the 160 genomes and MAGs based on a presence–absence matrix for the 18 selected genes.

was predicted to be present in *NP-α* and *-β*, as well as in most *NP-θ* and some *NP-η*, *-ε*, and *-γ* (Fig. 2; Supplementary Tables S4 and S5). Genes involved in chemotaxis and motility showed the highest within-clade variation, where presence of genes related to archaellum formation (*flaK* and *flaI*) varied between taxa of the *NP-γ* clade (Fig. 2; Supplementary Tables S4 and S5), while nearly all lineages except *NP-ε* and *-α* were predicted to have the *cheY* gene, encoding a response regulator associated with chemotaxis. Regarding environmental adaptation, genes associated with osmotic regulation (*nhaP*, *trk*) and thermoadaptation (*cpsC*) were also predicted to be present more often in *NP* clades and less often in *NS* (Fig. 2; Supplementary Tables S4 and S5). The presence of

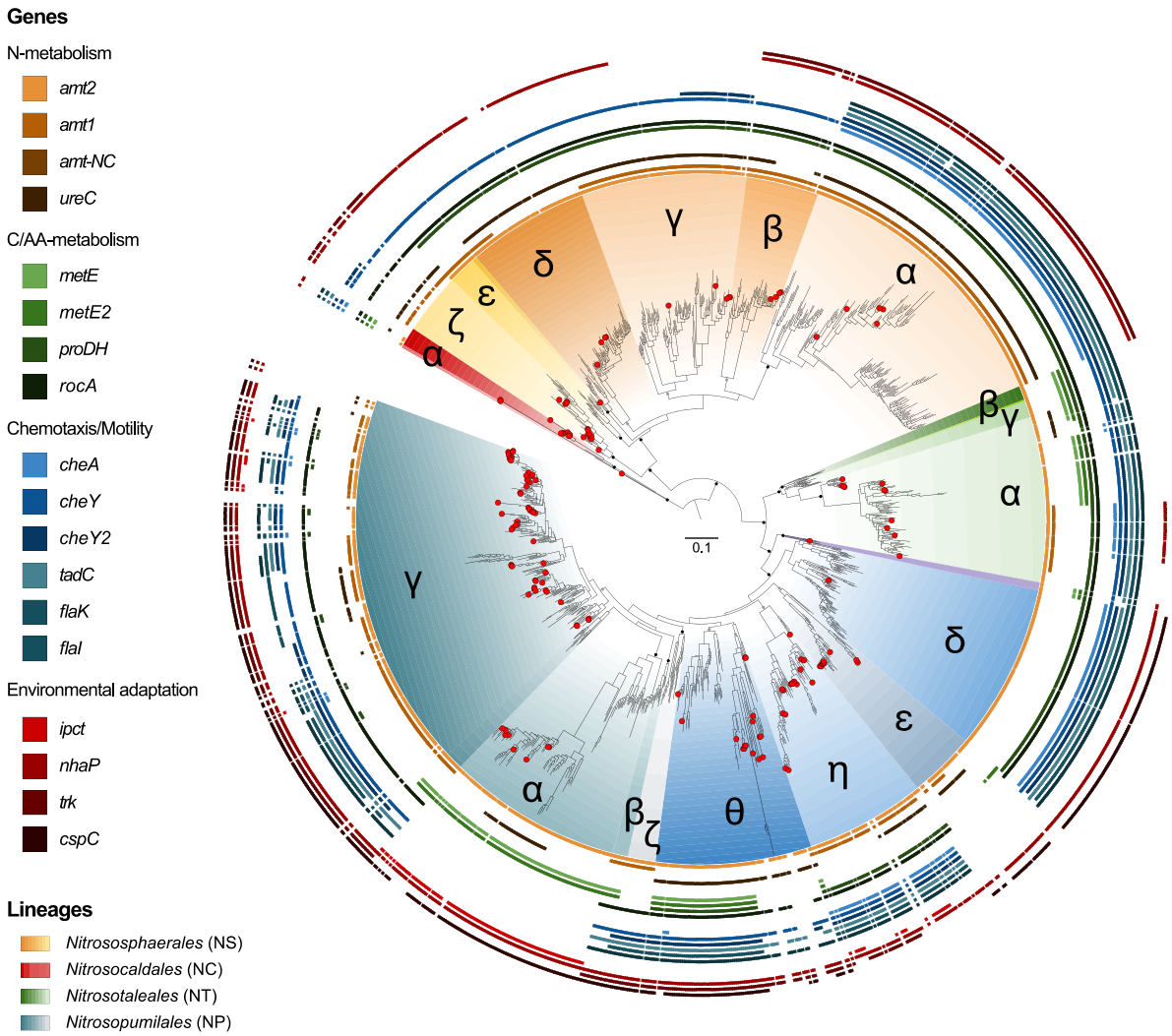


**Figure 2.** Predictions of gene presence/absence across the archaeal *amoA* reference phylogeny using phylogenetic eigenvectors. Outer rings depict predicted presence (filled squares) and absence (no squares) of specific genes (see Table 1 for definition). Circular markers at the tips represent the isolate genomes and MAGs with >80% completeness and <5% contamination that were used to train the predictive model. Clades within each AOA lineages are denoted by Greek letters. Nodes with >80 SH-aLRT and >70 ultrafast bootstrap are indicated by solid points, whereas those with >80 SH-aLRT support only are indicated by lighter-shaded points. For better visualization, we only display bootstrap values for lineages and clades. See Supplementary Fig. S3 for all bootstrap values across deeper nodes.

*ipct* gene was mostly restricted to taxa of the *NP-α* clade. A similar pattern of gene presence was obtained using ancestral state reconstruction (Supplementary Fig. S4; Supplementary Tables S6 and S7); a co-inertia analyses, i.e. a test of collinearity between two matrices, performed on phylogenetic eigenvector- and ancestral state reconstruction-based predictions had a $R^2$ of 0.73.

The phylogenetic eigenvector-based predictions of gene presence for all genes resulted in an average of 88.4% accuracy, 86.1% sensitivity, and 82% specificity based on a 20% hold-out-validation (Table 1). Ancestral state reconstruction of gene presence resulted in similar levels of accuracy, sensitivity, and specificity of predictions (89.5%, 86.3%, and 83%, respectively; Table 1). For both methods, the predictive accuracy increased linearly with the strength of the phylogenetic signal ($R^2 = 0.71$ and 0.68 for phylogenetic eigenvectors and ancestral state reconstructions, respectively, Supplementary Fig. S5).

### *Link between predicted genes and soil properties: a case study*

To exemplify how predictions of gene presence/absence can contribute to indirect trait-based studies, we placed the *amoA* sequences of the members of 50 AOA communities from arable land in the updated *amoA* reference phylogeny, predicted the gene presence among these *amoA*-based OTUs (Fig. 3A), and tested the link between predicted genes and soil properties by performing RLQ [58] and fourth-corner [59] analyses. The first RLQ axis showed that OTUs predicted to have genes encoding the high-affinity ammonia transporter (*amt2*), chemotaxis response (*cheY2*), and proline dehydrogenase (*proDH*) genes yet lacking the *nhaP* gene were more associated with most of the sites with lower pH (pH = 5.7–6.0 in sites S19, 21, 23, and 28; Figs. 3B and C). The second RLQ axis highlighted two specific genes, i.e. the low-affinity ammonia transporter (*amt1*) and the *ureC*, which were most closely associated to sites with higher levels of total nitrogen and carbon, and more moderate soil pH (pH = 6.1–6.2 in sites S29, 31, 34, 37; Figs. 3B and C). When testing the univariate associations of each gene and soil property using the fourth-corner approach, significant correlations were only found before correcting the *P*-values for multiple testing (Supplementary Fig. S6).

## Discussion

Predicting gene presence can inform about key potential functions that microorganisms perform in the environment without

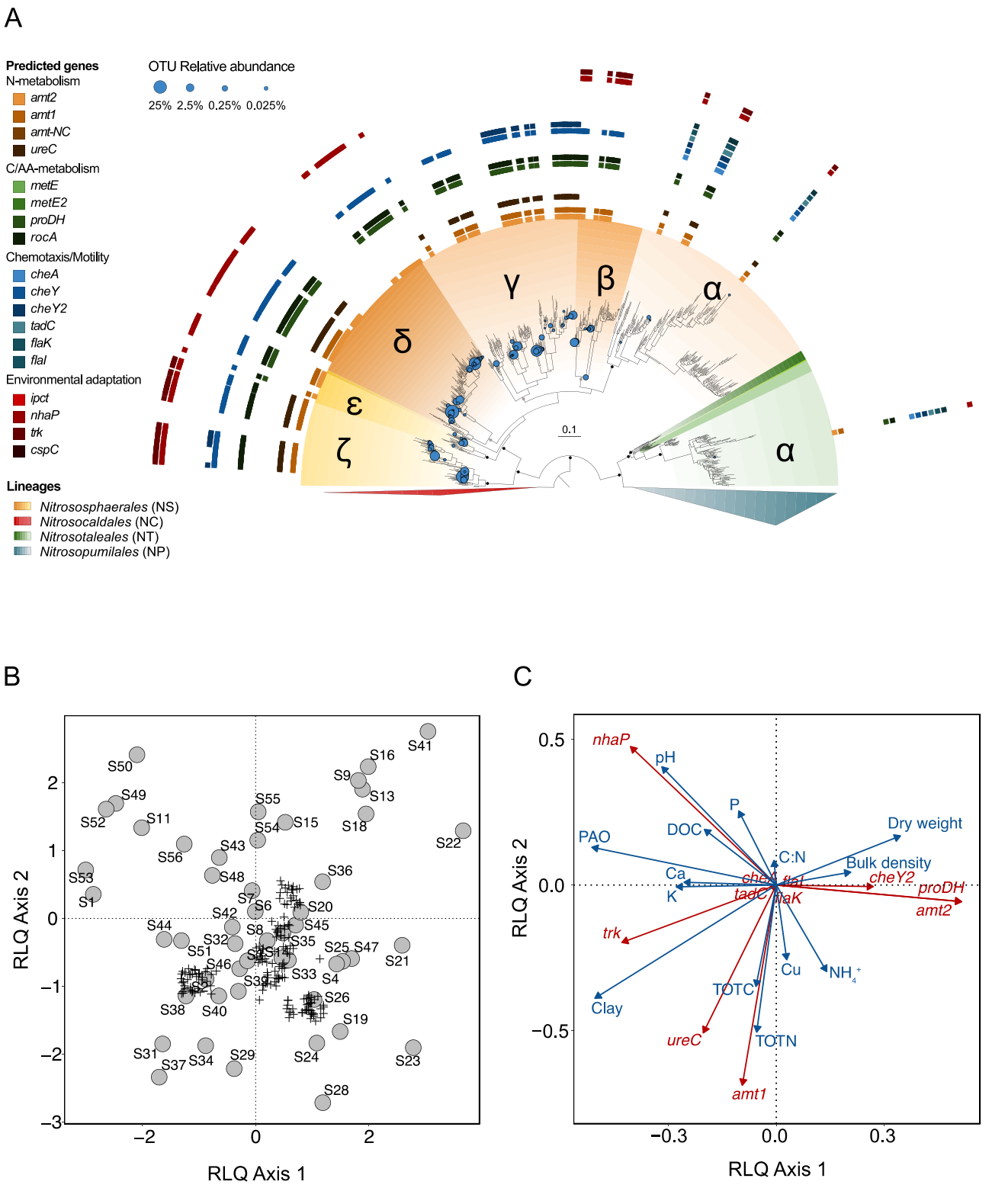


**Figure 3.** Association of predicted genes of AOA communities with soil properties across 44 ha of arable land. (A) Our updated *amo*A reference phylogeny with the placements of the 162 operational taxonomical units (OTUs) as grafted leaves. For visualization purposes, the relative abundances of the OTUs are displayed as circles on the nodes where they were placed. The circle size is proportional (fifth root) to their relative abundance. Outer rings depict predicted presence (filled squares) and absence (no squares) of specific genes for each OTU (see Table 1 for definition). None of the OTUs belonged to *NP* nor *NC*, and therefore both lineages are collapsed in the tree. Clades within each AOA lineage are denoted by Greek letters. Nodes with >80 SH-aLRT and >70 ultrafast bootstrap are indicated by solid points, whereas those with >80 SH-aLRT support only are indicated by lighter-shaded points. For better visualization, we only display bootstrap values for lineages and clades. See Supplementary Fig. S3 for all bootstrap values across deeper nodes. (B) Biplot of the RLQ analysis displaying scores of 50 sites (S1–S50) and 162 operational taxonomical units (OTUs) in circles and plus (+) signs, respectively. OTU symbols were jittered for visualization purposes. (C) Biplots of the RLQ analysis displaying association between predicted genes (red) and observed soil properties (blue). PAO refers to potential ammonia oxidation rates. The genes whose predicted values were all 0 or 1 were not included in the RLQ analysis. The global *P*-value associated with the RLQ analysis was 0.02.

having access to their full genomes. In this study, we predicted the presence of a set of genes of AOA with >88% accuracy using the *amoA* phylogenetic signals implemented in phylogenetic eigenvectors. The overall distribution of our isolate genomes and MAGs across the phylogeny supports the reliability of these predictions, particularly for AOA belonging to clades with good genome representation (see Supplementary Table S8). While genome and MAG incompleteness may limit inferences of potential microbial functions [62], we did not find any association between predictive accuracy and genome completeness (Supplementary Fig. S7A). We only observed a weak trend suggesting that the presence of genes in highly complete genomes was predicted with lower sensitivity and higher specificity than in less complete genomes (Supplementary Fig. S7B and C). Thus, including more incomplete genomes in the training dataset could increase false-negative predictions.

The high prediction accuracy can be explained by the strength of the phylogenetic signal of the selected genes. All the screened genes were conserved phylogenetically, in line with previous studies [16, 19], and the strength of the phylogenetic signal was positively correlated with the accuracy of the predictions [5, 6]. For example, the *amt-NC* gene had one of the strongest phylogenetic signals and displayed the highest prediction accuracy, while the *ureC* gene had one of the weakest signals and displayed the lowest accuracy. The differences in phylogenetic signal between genes are likely to be the evolutionary result of niche specialization. The *amt-NC* variant of the *amt* gene, described here for the first time, encodes an ammonia transporter that is present only in *Nitrosocaldus* AOA inhabiting thermal waters (Supplementary Text S1), and is therefore localized in a single clade within the *amoA* phylogeny. By contrast, the *ureC* gene tends to be phylogenetically dispersed. In agreement, soil AOA thrive in conditions in which ammonia is supplied slowly through mineralization of organic matter [63], and urease genes are abundant in Nitrososphaerota communities in oligotrophic marine environments [64]. The overall high accuracy of the predictions may also be the result of the availability of AOA genomes and MAGs across the *amoA* phylogeny. Accordingly, predictions on genomes and MAGs belonging to the *NS-ζ* and *NC-α*, *NP-ε*, and *NP-α* clades had the highest accuracies, and these AOA lineages were well-represented by genomes with close phylogenetic relatedness to the within-clade available relatives [5, 65] (Supplementary Table S8). Phylogenetic modeling of potential microbial traits should be therefore restricted to groups of organisms with good genomic representation in the databases [4, 66].

Predictions of gene distribution can provide a mechanistic understanding of community assembly in the environment. By implementing the gene presence predictions in multivariate models of AOA communities in soils with varying physical and chemical properties, we show that AOA communities adapted to low pH soils are more likely to have the high-affinity ammonia transporter (*amt2*), chemotaxis gene *cheY2*, and proline dehydrogenase gene *proDH*. This makes sense given the low availability of ammonia at low pH and is in line with previous studies showing down-regulation of bacterial chemotaxis genes at high pH and involvement of proline metabolism in abiotic stress response [67, 68]. By contrast, high pH sites were associated with OTUs predicted to have the *nhap* gene encoding a protein related to $Na^+/H^+$ antiporters, which are more important for homeostasis under neutral or alkaline conditions [69]. Resource rich sites with high amounts of total nitrogen and soil organic carbon, located in areas with higher influx of nitrogen via biological $N_2$ fixation [22], were related to the increase of relative abundance of AOA taxa with predicted potential for ureolytic metabolism (*ureC*). Accordingly, nitrogen addition may increase the abundance of *ureC* genes [70]. Higher potential ammonia oxidation activity was linked to taxa with the low-affinity ammonium transporter. This could be because the high ammonium concentrations used in the activity assay favored the ammonia oxidizers with low ammonia affinity or those that prefer inorganic to organic N sources, including ammonia-oxidizing bacteria [71]. Overall, our predictions on soil AOA communities show that the combination of gene presence predictions with RLQ and fourth-corner analyses can shed light on mechanisms of community assembly. This approach can be expanded to microbial groups other than AOA to decipher ecological dynamics, given a well curated reference database and phylogeny to which either amplicon, metagenomic, or metatranscriptomic data can be mapped.

Phylogenetic eigenvector mapping can complement the broadly used ancestral state reconstruction when performing phylogeny-based modeling. We show that both methods have similar values of accuracy, sensitivity, and specificity of the predictions, and accuracy values between these two methods across all genes were highly correlated ($R^2 = 0.97$). Based on co-inertia testing, both predictions across the phylogeny were associated with a high *RV* coefficient value, i.e. a correlation metric between two multivariate sets of variables [10], of 0.73. When setting the threshold for classifying probabilities at 0.5 in our phylogenetic eigenvector models, which is the default for ancestral state reconstruction in the R package picante [51], the *RV* increased to 0.78. The use of either phylogenetic eigenvector maps or ancestral state reconstruction depends on the role of phylogeny in the analyses [72, 73]. When predicting presence of gene using only phylogenetic signals, both methods give similar results, with ancestral state reconstruction not requiring the extra step of selecting or regularizing a model [6]. Regularizing a model may particularly be challenging when dealing with uneven class distribution. For example, when predicting the presence of the amt-NC gene across the phylogeny, the optimized $\lambda$ hyper-parameter of the elastic net was 0.31, and all coefficient estimates shrank to 0, i.e. all taxa in the phylogeny would have the same probability of gene presence. Although modifying the $\lambda$ value would render the same class predictions between phylogenetic eigenvectors and ancestral state reconstruction, the latter method does not rely on hyper-parameter tuning and therefore more useful on cases with very few presences or absences of specific genes. On the other hand, phylogenetic eigenvectors can be used when modeling phylogenetic signal together with other factors, e.g. abiotic variables or gene co-occurrences, which is useful when estimating traits on partially incomplete databases [74, 75]. Although in this study we use phylogenetic eigenvector maps to predict the probability of gene presence and not functions themselves, our approach could be used to predict high-level functional trait activities when information about them is present in the input data. In addition, the procedure used by the MPSEM R [48] package to calculate phylogenetic eigenvectors can use phylogenetic networks as input. Thus, phylogenetic eigenvectors can potentially be used for microorganisms for which horizontal gene transfer occurs, as well as for organisms that undergo reticulate speciation or hybridization [76]. Finally, the sets of latent descriptors produced by phylogenetic eigenvector mapping can be used with other modeling approaches, such as support vector machines, gradient boost machines, or artificial neural networks. Although some studies report a better performance of ancestral state reconstruction over phylogenetic eigenvectors [77], perhaps because of model selection issues, the present study shows that both can provide similar accuracies, while

phylogenetic eigenvector maps constitute a more versatile tool for phylogenetic modeling.

To conclude, we show that we can move toward gene distribution predictions in microbial ecology. Whereas the absence of many microbial genomes can limit the implementation of trait-based studies in microbial ecology, many microbial genes are conserved phylogenetically, and their presence can be predicted using such tools as phylogenetic eigenvectors and ancestral state reconstruction. Predictive modeling of potential microbial functions can provide useful information to understand how evolution shapes the genetic content of microorganisms, how that determines their distribution in the environment, and how that ultimately may impact ecosystem functions.

## Acknowledgments

We are grateful to Gabriel Dansereau of Université de Montréal for insightful discussions during the analysis of the data.

## Author contributions

M.A.R., C.M.J., and S.H. conceptualized the study and acquired the funding. M.A.R. and C.M.J. compiled and screened the genomes and metagenomes and updated the reference phylogeny. M.A.R., G.G., and P.L. performed the analyses. M.A.R. was responsible for the visualization of results. All authors contributed to the manuscript writing.

## Supplementary material

Supplementary material is available at *ISME Communications* online.

## Conflicts of interest

The authors declare that they have no competing interest.

## Funding

This research was supported by The Swedish Research Council (grant 2020-00434 to M.A.R.).

## Data availability

The datasets with gene content, phylogenetic trees, and code to replicate the predictive analyses are available on Github (https://github.com/RedondoMA/AOA_gene_predictions). The NCBI and JGI accession ID of the 457 AOA genomes that were used in this study are provided in Supplementary Table S1. The alignment to build the reference phylogeny tree is available as separated fasta file in supplementary information. The phylogenetic eigenvector- and ancestral state reconstruction-based predictions for the taxa of the reference phylogeny are provided in Supplementary Tables S5 and S7. The complete sequence dataset of the case study is available in the NCBI Short Read Archive under BioProject Accession no. PRJNA436119.

## References

1. Lajoie G, Kembel SW. Making the most of trait-based approaches for microbial ecology. *Trends Microbiol* 2019;**27**:814–23. https://doi.org/10.1016/j.tim.2019.06.003
2. Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J* 2013;**7**:830–8. https://doi.org/10.1038/ismej.2012.160
3. Morrissey EM, Mau RL, Hayer M. *et al.* Evolutionary history constrains microbial traits across environmental variation. *Nat Ecol Evol* 2019;**3**:1064. https://doi.org/10.1038/s41559-019-0918-y
4. Langille MGI, Zaneveld J, Caporaso JG. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;**31**:814–21. https://doi.org/10.1038/nbt.2676
5. Goberna M, Verdú M. Predicting microbial traits with phylogenies. *ISME J* 2016;**10**:959–67. https://doi.org/10.1038/ismej.2015.171
6. Walkup J, Dang C, Mau RL. *et al.* The predictive power of phylogeny on growth rates in soil bacterial communities. *ISME Commun* 2023;**3**:1–8. https://doi.org/10.1038/s43705-023-00281-1
7. Garland T, Ives AR. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am Nat* 2000;**155**:346–64. https://doi.org/10.1086/303327
8. Kembel SW, Cowan PD, Helmus MR. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 2010;**26**:1463–4. https://doi.org/10.1093/bioinformatics/btq166
9. Diniz-Filho JAF, de Sant'Ana CER, Bini LM. An eigenvector method for estimating phylogenetic inertia. *Evol Int J Org Evol* 1998;**52**:1247–62. https://doi.org/10.1111/j.1558-5646.1998.tb02006.x
10. Legendre P, Legendre L. *Numerical Ecology*. Amsterdam: Elsevier, 2012.
11. Guénard G, Legendre P, Peres-Neto P. Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods Ecol Evol* 2013;**4**:1120–31. https://doi.org/10.1111/2041-210X.12111
12. Guénard G, Lanthier G, Harvey-Lavoie S. *et al.* Modelling habitat distributions for multiple species using phylogenetics. *Ecography* 2017;**40**:1088–97. https://doi.org/10.1111/ecog.02423
13. Stahl DA, de la Torre JR. Physiology and diversity of ammonia-oxidizing archaea. *Ann Rev Microbiol* 2012;**66**:83–101. https://doi.org/10.1146/annurev-micro-092611-150128
14. Lehtovirta-Morley LE. Ammonia oxidation: ecology, physiology, biochemistry and why they must all come together. *FEMS Microbiol Lett* 2018;**365**:fny058. https://doi.org/10.1093/femsle/fny058
15. Kerou M, Offre P, Valledor L. *et al.* Proteomics and comparative genomics of Nitrososphaera viennensis reveal the core genome and adaptations of archaeal ammonia oxidizers. *Proc Natl Acad Sci USA* 2016;**113**:E7937–46. https://doi.org/10.1073/pnas.1601212113
16. Kerou M, Ponce-Toledo RI, Zhao R. *et al.* Genomes of Thaumarchaeota from deep sea sediments reveal specific adaptations of three independently evolved lineages. *ISME J* 2021;**15**:2792–808. https://doi.org/10.1038/s41396-021-00962-6
17. Luo Z-H, Narsing Rao MP, Chen H. *et al.* Genomic insights of "Candidatus Nitrosocaldaceae" based on nine new metagenome-assembled genomes, including "Candidatus Nitrosothermus" Gen Nov. and two new species of "Candidatus Nitrosocaldus". *Front Microbiol* 2021;**11**:608832. https://doi.org/10.3389/fmicb.2020.608832
18. Leininger S, Urich T, Schloter M. *et al.* Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 2006;**442**:806–9. https://doi.org/10.1038/nature04983
19. Oton EV, Quince C, Nicol GW. *et al.* Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. *ISME J* 2016;**10**:85–96. https://doi.org/10.1038/ismej.2015.101

20. Wang H, Bagnoud A, Ponce-Toledo RI. *et al.* Linking 16S rRNA gene classification to amoA gene taxonomy reveals environmental distribution of ammonia-oxidizing archaeal clades in peatland soils. *mSystems* 2021;**6**:e00546-21. https://doi.org/10.1128/mSystems.00546-21
21. Alves RJE, Minh BQ, Urich T. *et al.* Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes. *Nat Commun* 2018;**9**:1–17. https://doi.org/10.1038/s41467-018-03861-1
22. Wessén E, Söderström M, Stenberg M. *et al.* Spatial distribution of ammonia-oxidizing bacteria and archaea across a 44-hectare farm related to ecosystem functioning. *ISME J* 2011;**5**:1213–25. https://doi.org/10.1038/ismej.2010.206
23. Gubry-Rangin C, Hai B, Quince C. *et al.* Niche specialization of terrestrial archaeal ammonia oxidizers. *Proc Natl Acad Sci* 2011;**108**:21206–11. https://doi.org/10.1073/pnas.1109000108
24. Dai S, Liu Q, Zhao J. *et al.* Ecological niche differentiation of ammonia-oxidising archaea and bacteria in acidic soils due to land use change. *Soil Res* 2018;**56**:71–9. https://doi.org/10.1071/SR16356
25. Jones CM, Hallin S. Geospatial variation in co-occurrence networks of nitrifying microbial guilds. *Mol Ecol* 2019;**28**:293–306. https://doi.org/10.1111/mec.14893
26. Dray S, Choler P, Dolédec S. *et al.* Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. *Ecology* 2014;**95**:14–21. https://doi.org/10.1890/13-0196.1
27. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 2010;**5**:e9490. https://doi.org/10.1371/journal.pone.0009490
28. Ludwig W, Strunk O, Westram R. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res* 2004;**32**:1363–71. https://doi.org/10.1093/nar/gkh293
29. Simão FA, Waterhouse RM, Ioannidis P. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**:3210–2. https://doi.org/10.1093/bioinformatics/btv351
30. Rognes T, Flouri T, Nichols B. *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584. https://doi.org/10.7717/peerj.2584
31. Minh BQ, Schmidt HA, Chernomor O. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:1530–4. https://doi.org/10.1093/molbev/msaa015
32. Kalyaanamoorthy S, Minh BQ, Wong TKF. *et al.* ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;**14**:587–9. https://doi.org/10.1038/nmeth.4285
33. Hoang DT, Chernomor O, von Haeseler A. *et al.* UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 2018;**35**:518–22. https://doi.org/10.1093/molbev/msx281
34. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;**49**:W293–6. https://doi.org/10.1093/nar/gkab301
35. Yu G, Smith DK, Zhu H. *et al.* Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;**8**:28–36. https://doi.org/10.1111/2041-210X.12628
36. Wright CL, Lehtovirta-Morley LE. Nitrification and beyond: metabolic versatility of ammonia oxidising archaea. *ISME J* 2023;**17**:1358–68. https://doi.org/10.1038/s41396-023-01467-0
37. Nakagawa T, Stahl DA. Transcriptional response of the archaeal ammonia oxidizer *Nitrosopumilus maritimus* to low and environmentally relevant ammonia concentrations. *Appl Environ Microbiol* 2013;**79**:6911–6. https://doi.org/10.1128/AEM.02028-13
38. Qin W, Zheng Y, Zhao F. *et al.* Alternative strategies of nutrient acquisition and energy conservation map to the biogeography of marine ammonia-oxidizing archaea. *ISME J* 2020;**14**:2595–609. https://doi.org/10.1038/s41396-020-0710-7
39. Alves RJE, Kerou M, Zappe A. *et al.* Ammonia oxidation by the Arctic terrestrial Thaumarchaeote Candidatus Nitrosocosmicus arcticus is stimulated by increasing temperatures. *Front Microbiol* 2019;**10**:1571. https://doi.org/10.3389/fmicb.2019.01571
40. Lehtovirta-Morley LE, Ross J, Hink L. *et al.* Isolation of 'Candidatus Nitrosocosmicus franklandus', a novel ureolytic soil archaeal ammonia oxidiser with tolerance to high ammonia concentration. *FEMS Microbiol Ecol* 2016;**92**:fiw057. https://doi.org/10.1093/femsec/fiw057
41. Sauder LA, Albertsen M, Engel K. *et al.* Cultivation and characterization of Candidatus Nitrosocosmicus exaquare, an ammonia-oxidizing archaeon from a municipal wastewater treatment system. *ISME J* 2017;**11**:1142–57. https://doi.org/10.1038/ismej.2016.192
42. Jung M-Y, Kim J-G, Sinninghe Damsté JS. *et al.* A hydrophobic ammonia-oxidizing archaeon of the Nitrosocosmicus clade isolated from coal tar-contaminated sediment. *Environ Microbiol Rep* 2016;**8**:983–92. https://doi.org/10.1111/1758-2229.12477
43. Offre P, Kerou M, Spang A. *et al.* Variability of the transporter gene complement in ammonia-oxidizing archaea. *Trends Microbiol* 2014;**22**:665–75. https://doi.org/10.1016/j.tim.2014.07.007
44. Liu L, Liu M, Jiang Y. *et al.* Production and excretion of polyamines to tolerate high ammonia, a case study on soil ammonia-oxidizing archaeon "Candidatus Nitrosocosmicus agrestis". *mSystems* 2021;**6**:e01003–20. https://doi.org/10.1128/msystems.01003-20
45. Zou D, Wan R, Han L. *et al.* Genomic characteristics of a novel species of ammonia-oxidizing archaea from the Jiulong River estuary. *Appl Environ Microbiol* 2020;**86**:e00736–20. https://doi.org/10.1128/AEM.00736-20
46. Orme D, Freckleton R, Thomas G. *et al. Caper: Comparative Analyses of Phylogenetics and Evolution in R*. R package version 1.0.1; 2018.
47. Fritz SA, Purvis A. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv Biol* 2010;**24**:1042–51. https://doi.org/10.1111/j.1523-1739.2010.01455.x
48. Guenard G, Legendre P. *MPSEM: Modeling Phylogenetic Signals Using Eigenvector Maps*. R package version 0.4-1; 2022.
49. Friedman J, Hastie T, Tibshirani R. *et al. Glmnet: Lasso and Elastic—Net Regularized Generalized Linear Models*. R package version 4.1-7; 2023.
50. Robin X, Turck N, Hainard A. *et al. pROC: Display and Analyze ROC Curves*. R package version 1.18.0; 2021.
51. Kembel SW, Ackerly DD, Blomberg SP. *et al. Picante: Integrating Phylogenies and Ecology*. R package version 1.8.2; 2020.
52. Kembel SW, Wu M, Eisen JA. *et al.* Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* 2012;**8**:e1002743. https://doi.org/10.1371/journal.pcbi.1002743
53. Enwall K, Throbäck IN, Stenberg M. *et al.* Soil resources influence spatial patterns of denitrifying communities at scales compatible with land management. *Appl Environ Microbiol* 2010;**76**:2243–50. https://doi.org/10.1128/AEM.02197-09
54. Söderström M, Lindén B. *Using Precision Agriculture Data for Planning Field Experiments e Experiences from a Research Farm in Sweden. In: Proceedings of the 12th International Conference on Mechanization of Field Experiments. IAMFE, St. Petersburg, Russia* 2004; 161–168.

55. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 2013;**10**:996–8. https://doi.org/10.1038/nmeth.2604
56. Barbera P, Kozlov AM, Czech L. *et al.* EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst Biol* 2019;**68**:365–9. https://doi.org/10.1093/sysbio/syy054
57. Czech L, Barbera P, Stamatakis A. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* 2020;**36**:3263–5. https://doi.org/10.1093/bioinformatics/btaa070
58. Dolédec S, Chessel D, ter Braak CJF. *et al.* Matching species traits to environmental variables: a new three-table ordination method. *Environ Ecol Stat* 1996;**3**:143–66. https://doi.org/10.1007/BF02427859
59. Legendre P, Galzin R, Harmelin-Vivien ML. Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology* 1997;**78**:547–62. https://doi.org/10.1890/0012-9658(1997)078[0547:RBTHST]2.0.CO;2
60. Dray S, Dufour A-B, Thioulouse J. *ade4: Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences*. R package version 1.7-22; 2023.
61. ter Braak CJF, Cormont A, Dray S. Improved testing of species traits–environment relationships in the fourth-corner problem. *Ecology* 2012;**93**:1525–6. https://doi.org/10.1890/12-0126.1
62. Eisenhofer R, Odriozola I, Alberdi A. Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun* 2023;**3**:1–5. https://doi.org/10.1038/s43705-023-00221-z
63. Hink L, Gubry-Rangin C, Nicol GW. *et al.* The consequences of niche and physiological differentiation of archaeal and bacterial ammonia oxidisers for nitrous oxide emissions. *ISME J* 2018;**12**:1084–93. https://doi.org/10.1038/s41396-017-0025-5
64. Alonso-Sáez L, Waller AS, Mende DR. *et al.* Role for urea in nitrification by polar marine Archaea. *Proc Natl Acad Sci* 2012;**109**:17989–94. https://doi.org/10.1073/pnas.1201914109
65. Guénard G, von der Ohe PC, de Zwart D. *et al.* Using phylogenetic information to predict species tolerances to toxic chemicals. *Ecol Appl* 2011;**21**:3178–90.
66. Djemiel C, Maron P-A, Terrat S. *et al.* Inferring microbiota functions from taxonomic genes: a review. *GigaScience* 2022;**11**:giab090. https://doi.org/10.1093/gigascience/giab090
67. Maurer LM, Yohannes E, Bondurant SS. *et al.* pH regulates genes for flagellar motility, catabolism, and oxidative stress in *Escherichia coli* K-12. *J Bacteriol* 2005;**187**:304–19. https://doi.org/10.1128/JB.187.1.304-319.2005
68. Liang X, Zhang L, Natarajan SK. *et al.* Proline mechanisms of stress survival. *Antioxid Redox Signal* 2013;**19**:998–1011. https://doi.org/10.1089/ars.2012.5074
69. Krulwich TA, Sachs G, Padan E. Molecular aspects of bacterial pH sensing and homeostasis. *Nat Rev Microbiol* 2011;**9**:330–43. https://doi.org/10.1038/nrmicro2549
70. Abdo AI, Xu Y, Shi D. *et al.* Nitrogen transformation genes and ammonia emission from soil under biochar and urease inhibitor application. *Soil Tillage Res* 2022;**223**:105491. https://doi.org/10.1016/j.still.2022.105491
71. Hazard C, Prosser JI, Nicol GW. Use and abuse of potential rates in soil microbiology. *Soil Biol Biochem* 2021;**157**:108242. https://doi.org/10.1016/j.soilbio.2021.108242
72. Swenson NG. Phylogenetic imputation of plant functional trait databases. *Ecography* 2014;**37**:105–10. https://doi.org/10.1111/j.1600-0587.2013.00528.x
73. Zaneveld JRR, Thurber RLV. Hidden state prediction: a modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. *Front Microbiol* 2014;**5**:431. https://doi.org/10.3389/fmicb.2014.00431
74. Debastiani VJ, Bastazini VAG, Pillar VD. Using phylogenetic information to impute missing functional trait values in ecological databases. *Eco Inform* 2021;**63**:101315. https://doi.org/10.1016/j.ecoinf.2021.101315
75. Guénard G. *A Phylogenetic Modelling Tutorial Using Phylogenetic Eigenvector Maps (PEM) as Implemented in R Package MPSEM (0.6-1)*, 2025. Available at: https://cran.r-project.org/web/packages/MPSEM/vignettes/REM_with_MPSEM.html (13 June 2025, date last accessed).
76. Stull GW, Pham KK, Soltis PS. *et al.* Deep reticulation: the long legacy of hybridization in vascular plant evolution. *Plant J* 2023;**114**:743-66. https://doi.org/10.1111/tpj.16142
77. Swenson NG, Weiser MD, Mao L. *et al.* Phylogeny and the prediction of tree functional diversity across novel continental settings. *Glob Ecol Biogeogr* 2017;**26**:553–62. https://doi.org/10.1111/geb.12559





*ISME Communications*, 2025, **5(1)**, ycaf087
https://doi.org/10.1093/ismeco/ycaf087
Original Article