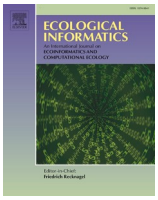
Ecological Informatics 83 (2024) 102817

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Modelling soil prokaryotic traits across environments with the trait sequence database *ampliconTraits* and the R package *MicEnvMod*

Jonathan Donhauser [a,*], Anna Doménech-Pascual [b], Xingguo Han [c], Karen Jordaan [d], Jean-Baptiste Ramond [e], Aline Frossard [c], Anna M. Romaní [b], Anders Priemé [a,f]

[a] Department of Biology, University of Copenhagen, Copenhagen, Denmark
[b] Research Group on Ecology of Inland Waters (GRECO), Institute of Aquatic Ecology, University of Girona, Girona, Spain
[c] Forest Soils and Biogeochemistry, Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Birmensdorf, Switzerland
[d] Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa
[e] Extreme Ecosystem Microbiomics & Ecogenomics (E^2ME) Lab., Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile
[f] Center for Volatile Interactions (VOLT), University of Copenhagen, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Trait sequence database
DNA sequencing
Microbial community
Cross validation
Weighted ensemble model

ABSTRACT

We present a comprehensive, customizable workflow for inferring prokaryotic phenotypic traits from marker gene sequences and modelling the relationships between these traits and environmental factors, thus overcoming the limited ecological interpretability of marker gene sequencing data. We created the trait sequence database *ampliconTraits*, constructed by cross-mapping species from a phenotypic trait database to the SILVA sequence database and formatted to enable seamless classification of environmental sequences using the SINAPS algorithm. The R package *MicEnvMod* enables modelling of trait – environment relationships, combining the strengths of different model types and integrating an approach to evaluate the models' predictive performance in a single framework. Traits could be accurately predicted even for sequences with low sequence identity (80 %) with the reference sequences, indicating that our approach is suitable to classify a wide range of environmental sequences. Validating our approach in a large trans-continental soil dataset, we showed that trait distributions were robust to classification settings such as the bootstrap cutoff for classification and the number of discrete intervals for continuous traits. Using functions from *MicEnvMod,* we revealed precipitation seasonality and land cover as the most important predictors of genome size. We found Pearson correlation coefficients between observed and predicted values up to 0.70 using repeated split sampling cross validation, corroborating the predictive ability of our models beyond the training data. Predicting genome size across the Iberian Peninsula, we found the largest genomes in the northern part. Potential limitations of our trait inference approach include dependence on the phylogenetic conservation of traits and limited database coverage of environmental prokaryotes. Overall, our approach enables robust inference of ecologically interpretable traits combined with environmental modelling allowing to harness traits as bioindicators of soil ecosystem functioning.

1. Introduction

Microorganisms play a pivotal role in soil ecosystem functioning and the number of studies addressing microbial communities and their links with soil processes has increased rapidly during the last years. However, it has proven difficult to identify interpretable microbial indicators of soil function. For instance, when using a taxonomy-based marker gene approach, limited ecological information is available for the mostly uncultivated microbial taxa in soil (Donhauser et al., 2020), hampering their use as bioindicators. Metagenome sequencing has been used as an alternative approach, allowing identification of functional genes associated with a certain environment or future climatic conditions. Metagenomic analyses typically involve millions of functional genes. Therefore, in many studies, only a small fraction of the most strongly changing genes is presented or genes are aggregated in broad categories according to databases that are not organized according to ecological functions. Thus, a key interest in microbial ecology is to infer traits from sequencing data that can be linked to soil processes such as carbon (C)

* Corresponding author at: Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark.
E-mail addresses: jonathan.donhauser@bio.ku.dk (J. Donhauser), a.domenech@udg.edu (A. Doménech-Pascual), xingguo.han@wsl.ch (X. Han), jbramond@uc.cl (J.-B. Ramond), aline.frossard@wsl.ch (A. Frossard), anna.romani@udg.edu (A.M. Romaní), aprieme@bio.ku.dk (A. Priemé).

https://doi.org/10.1016/j.ecoinf.2024.102817
Received 31 May 2024; Received in revised form 6 August 2024; Accepted 7 September 2024
Available online 10 September 2024

and nutrient cycling, relevant for example to understand soil feedback to climate change or soil fertility. For instance, microbial C-cycling models postulate mechanisms based on traits (Allison and Goulden, 2017; Kaiser et al., 2014), but it remains challenging to validate such mechanisms empirically.

An advantage of marker gene-based analyses of microbial communities is the recovery of rare species, while metagenomic analyses only cover the most abundant species. Moreover, marker gene-based analyses are less costly and require less computational knowledge. Conversely, metagenomics involves complex and time-consuming bioinformatics such as assembly and binning, and annotations are often based on a single best hit without accounting for similarly well matching alternatives (Huson et al., 2011). Thus, inferring traits from marker gene sequencing is an attractive alternative to investigate microbial links to C-cycling. Several approaches have been developed to infer microbial function from marker genes. PICRUSt (Douglas et al., 2020; Langille et al., 2013) places environmental sequences in a phylogenetic tree of 16S rRNA sequences of taxa with sequenced genomes and computes a functional gene profile using extended ancestral state reconstruction. Tax4Fun (Aßhauer et al., 2015) uses a taxonomic profile of the prokaryotes in KEGG (Kanehisa et al., 2016) and estimates the functional gene profile of environmental marker gene sequences using a precomputed matrix of taxonomy - gene content associations. For eukaryotic microorganisms, tools like FUNGuild (Nguyen et al., 2016) and NINJA (Sieriebriennikov et al., 2014) enable trait classifications by matching the taxonomic classification of an environmental sequence at the species or genus level with a trait database. Cébron et al. (2021) suggested an approach to infer traits for environmental sequences based on their taxonomic classification, averaging trait values if there are multiple values within the assigned taxonomic group. These methods have several shortcomings, however. PICRUSt and Tax4Fun, like metagenomes, come with a large number of functional genes that do not translate directly into ecologically meaningful traits. FUNGuild and NINJA are not available for prokaryotes and trait annotations depend on taxonomic annotations. Similarly, the approach by Cébron et al. (2021) depends on taxonomy and does not allow to estimate confidence of trait annotations. Thus, a trait inference workflow that allows for direct classification of prokaryotic traits and that supports confidence estimates is currently lacking.

Recently, efforts have been made to create databases with systematic annotation of microbial traits based on descriptions of isolates (Barberán et al., 2017; Cébron et al., 2021; Madin et al., 2020) as well as on text mining and genomic data (Brbić et al., 2016) offering valuable resources for the development of trait-inference workflows for uncultured, environmental taxa. Edgar (2017) developed the SINAPS algorithm to classify traits from a sequence database with trait annotations. SINAPS uses the algorithm of the taxonomy classifier SINTAX (Edgar, 2018), which is based on shared words between query and reference sequences. SINAPS was tested for energy metabolism, Gram stain, presence of a flagellum, V4 primer mismatches, and 16S rRNA gene copy number (Edgar, 2017). Compared to taxonomy-based trait classification, SINAPS has the advantage of using a bootstrap procedure to estimate traits classification confidence. Thus, SINAPS constitutes a promising trait inference tool, but a comprehensive trait-sequence reference database is currently lacking.

To elucidate the role of microorganisms in soil processes, it is crucial to investigate the distribution of microbial traits through space and time to create a mechanistic understanding of the range of conditions under which microorganisms can thrive, enabling predictions for future environmental conditions. Spatial modelling is well established for macroorganisms (Guisan and Zimmermann, 2000), but less commonly used for microorganisms. Thus, comprehensive frameworks to build and cross-validate models, combining the strengths of multiple model types have been implemented in R packages (Di Cola et al., 2017; Thuiller et al., 2009), but are based on present-absence data and are therefore not directly applicable to continuous measures such as community-weighted trait means (CWM).

To address these gaps, here, we present a complete and flexible workflow to model the spatial distribution of prokaryotic traits inferred from amplicon sequence data in a single framework. We leveraged a phenotypic trait-database (Madin et al., 2020) combined with sequences from the SILVA small subunit (SSU) rRNA database (Quast et al., 2013) to create the trait sequence database *ampliconTraits*. *ampliconTraits* allows for seamless integration with SINAPS to classify environmental marker gene sequences for multiple phenotypic traits. The R package *MicEnvMod* provides functionality to combine the strengths of different model types, to examine variable importance and responses to specific predictors and to cross-validate model performance. We cross-validated trait classifications with *ampliconTraits* and SINAPS and evaluated their robustness with data from a large trans-continental study. Using functions from *MicEnvMod*, we modelled CWMs of genome size as an example trait with a combination of random forest (RF) models and generalized linear models (GLM) and evaluated model performance by repeated split sampling. To demonstrate extrapolation to areas beyond

Table 1
Number of Species and sequences in the reference database for each trait, interval numbers for continuous traits and categories for categorical traits.

	Cell diameter (lower) [µm]	Cell diameter (upper) [µm]	Cell length (lower) [µm]	Cell length (upper) [µm]	Doubling time [h]	Genome size [bp]	pH_{opt}	T_{opt} [°C]	16S rRNA gene copy number
Species	5645	3044	5223	3091	879	8554	3893	6389	2389
Sequences	33,069	18,731	30,886	18,925	14,994	59,416	16,498	46,754	33,216
Interval numbers	5, 10, 20, 30	5, 10, 20, 30	5, 10, 20, 30	5, 10, 20, 30	5, 10, 20, 30, 40, 50	5, 10, 20	5, 10, 20	10, 20	5, 10, exact gene copy numbers

	Cell shape	Gram stain	Oxygen preference	Motility	Salinity preference	Temperature preference	Sporulation
Species	6801	9834	9531	6518	515	2533	5832
Sequences	53,852	67,051	62,806	47,134	13,631	33,160	46,490
Categories	coccus; bacillus; coccobacillus; spiral filament; vibrio; pleomorphic; irregular; disc; star; fusiform; spindle; branced; square	negative; positive	microaerophilic; anaerobic; aerobic; facultative; obligate aerobic; obligate anaerobic	gliding; flagella; yes; axial filament	moderate-halophilic; extreme-halophilic; non-halophilic; halophilic; stenohaline; halotolerant; euryhaline	thermophilic; mesophilic; psychrophilic; extreme thermophilic; facultative psychrophilic; psychrotolerant; thermotolerant	no; yes

the study sites, we predicted genome sizes for the Iberian Peninsula.

2. Methods

2.1. Classification of traits from amplicon sequences

To create *ampliconTraits* trait sequence databases, we used a phenotypic trait database (Madin et al., 2020) combined with the SILVA sequence database (Quast et al., 2013). All code for database creation and classification of environmental sequences along with detailed documentation is available on GitHub (https://github.com/jdonhauser/ampliconTraits). First, we created an amplicon specific version of SILVA SSURef v138 for the V3-V4 region of the 16S rRNA gene (341–806) using RESCRIPt (Ii et al., 2021). We then used traits aggregated at the species level provided as condensed_species_NCBI.csv (Madin, 2021) and obtained 16S rRNA gene sequences for each species by cross mapping the NCBI taxid to the SILVA database using the file taxmap_embl-ebi_ena_ssu_ref_138.txt obtained from the SILVA homepage (https://www.arb-silva.de/fileadmin/silva_databases/release_138/Exports/taxonomy/taxmap_embl-ebi_ena_ssu_ref_138.txt.gz). 13,426 out of 14,893 species in the trait database could be mapped using the NCBI taxid. Of the remaining 1467 species, an additional 495 species could be cross mapped using the species name. In these cases, the taxid in SILVA corresponded to a strain of the species in the trait database and therefore did not match. We implemented the majority of traits present in the phenotypic trait database (Madin et al., 2020). A few traits were omitted due to anticipated methodological issues, because a similar, better represented trait was present or because it was deemed less ecologically relevant. In principle any trait linked with an NCBI taxid or a species name could be implemented in *ampliconTraits*. For instance, the number of coding genes is highly correlated to genome size but had much lower coverage and was therefore not included. Carbon substrates and pathways were omitted because they were represented as a list of terms rather than a single term and because we expected shallow phylogenetic conservation in many cases (Martiny et al., 2012, 2015). Not all species had annotations for all traits (Table 1).

We used SINAPS (Edgar, 2017) to classify environmental amplicon sequence variants (ASVs) against the sequence reference database for each trait using the *-sinaps* command in usearch v11.0.667 (Edgar, 2010). SINAPS works similar to a taxonomy classifier comparing shared k-mers between query and reference sequences and uses bootstrapping for confidence estimates. To classify continuous traits, we binned them into discrete intervals using the function *cut* in R specifying the number of breaks, which divide the whole range of values into n intervals of equal size. We tested different numbers of equal size intervals between 5 and 50, counting only bins that contained values. That is if for a given number of breaks, some intervals were empty, we increased the number of breaks until we obtained the desired number of intervals where each interval contained at least one value.

Then, we created a reference database for each categorical trait and each trait-interval combination for continuous traits. We subset the sequence database to the sequences for which a trait was annotated using seqkit (Shen et al., 2016) and then added the trait and its value to the fasta header in the format trait = value as required by SINAPS. Moreover, we determined the sequence identity between each query and the closest reference sequence using the *usearch_global* command (Edgar, 2010) as an additional measure for the quality of the classification. We used the following traits and interval numbers (terms in parentheses indicate the name of the trait in the downloaded table): cell diameter (lower) (d1_lo; 5, 10, 20 and 30 intervals), cell diameter (upper) (d1_up; 5, 10, 20 and 30 intervals), cell length (lower) (d2_lo; 5, 10, 20, 30 intervals), cell length (upper) (d2_up; 5, 10, 20, 30 intervals), doubling time (doubling_h; 5, 10, 20, 30, 40, 50 intervals), genome size (5, 10, 20 intervals), pH_{opt} (optimum_pH; 5, 10, 20 intervals), T_{opt} (optimum_tmp; 10, 20 intervals), 16S rRNA gene copy number (rRNA16S_genes; 5, 10 intervals, exact number), oxygen preference (metabolism), motility, salinity preference (salinity_range) and sporulation.

2.2. Dataset used to model traits

We used 16S rRNA gene amplicon sequencing data and metadata from a comprehensive soil dataset including 80 sites across Greenland, Europe and South Africa (Fig. S1) to evaluate the performance of our workflow. For evaluation of trait classifications, the full dataset of was used, for modelling trait – environment relationships, 10 sites from Greenland that were sampled to represent small-scale microclimatic heterogeneity were removed. The sites encompassed mean annual temperatures (MAT) from −18.1 °C to 22.4 °C, mean annual precipitation (MAP) from 45 to 1635 mm, soil organic matter (SOM) contents from 0.34 to 59 % and pH from 2.6 to 8.1 (Fig. S1). A detailed description of the sampling procedure, DNA isolation, PCR, amplicon sequencing, inference of ASVs and measurement of soil physicochemical properties is available in the supplementary information. Briefly, we sequenced the V3-V4 region of the prokaryotic16S rRNA gene using the primers 341F and 801R (Frey et al., 2016) and paired-end Illumina Miseq technology and we used DADA2 (Callahan et al., 2016) implemented in qiime2 (Bolyen et al., 2019) to denoise raw sequences and infer ASVs. Measured environmental variables in the dataset include pH, SOM, total organic C (TOC), total C (TC) and N (TN), soil C:N, total litter, litter C, litter N, litter C:N, soil texture (sand, silt, clay), water activity (a_w) and in situ soil temperature at the time of sampling. In addition, we extracted further bioclimatic variables from the worldclim database (Fick and Hijmans, 2017): BIO1 (MAT), BIO5 (maximum temperature warmest month), BIO7 (temperature, annual range; maximum temperature of warmest month minus minimum temperature of coldest month), BIO12 (MAP), BIO15 (precipitation seasonality; ratio of the standard deviation of the monthly total precipitation to the mean monthly total precipitation). The aridity index was extracted from the global aridity and PET database (Zomer et al., 2008). Moreover, we extracted land cover classification according the International Geosphere-Biosphere Programme classification using the MODIS product MCD12Q1_LC1 (Friedl et al., 2010) for the year 2020 and revised it manually according to photos of the study sites. Finally, we extracted the soil water holding capacity (WHC) from the ISRIC-WISE30sec data set (Batjes, 2016). Missing values in the metadata were imputed based on principal components using the *estim_ncpPCA* and *imputePCA* function from the package *missMDA* (Josse and Husson, 2016). All plots and statistical analyses were produced in R version 4.1.3 (R Core Team, 2022). Raw sequences were deposited in the NCBI Sequence Read Archive under the accession number PRJNA1073882.

2.3. Cross validation of trait inference

We assessed the predictive accuracy for each trait as a function of the sequence identity between query and reference sequence (Edgar, 2018). This cross validation approach allows to evaluate accuracy in a range of scenarios where a closely related reference sequence is not available due to limited size and/or phylogenetic coverage of the reference database. A markdown specifying the details of this cross validation is available in the supplementary information. Using the usearch command *distmx_split_identity*, we split the database in a test and a training set where for each sequence in the test set the most similar sequence in the training set has a maximum sequence identity of x % (Edgar, 2018; described in more detail at https://www.drive5.com/usearch/manual/cvi.html). This involves creating a distance matrix of sequence identity based on pairwise alignments of all sequences in the database. Based on the distance matrix, sequences are split into complementary sets of sequences where the most similar sequences in the second set share a maximum of x % sequence identity with all sequences in the first set. Then the sequences in one of these sets can be used as test set and classified against the second set. We did this for sequence identity thresholds of 97, 95, 90, 85 and 80 %. Subsequently, we classified the sequences in the test set

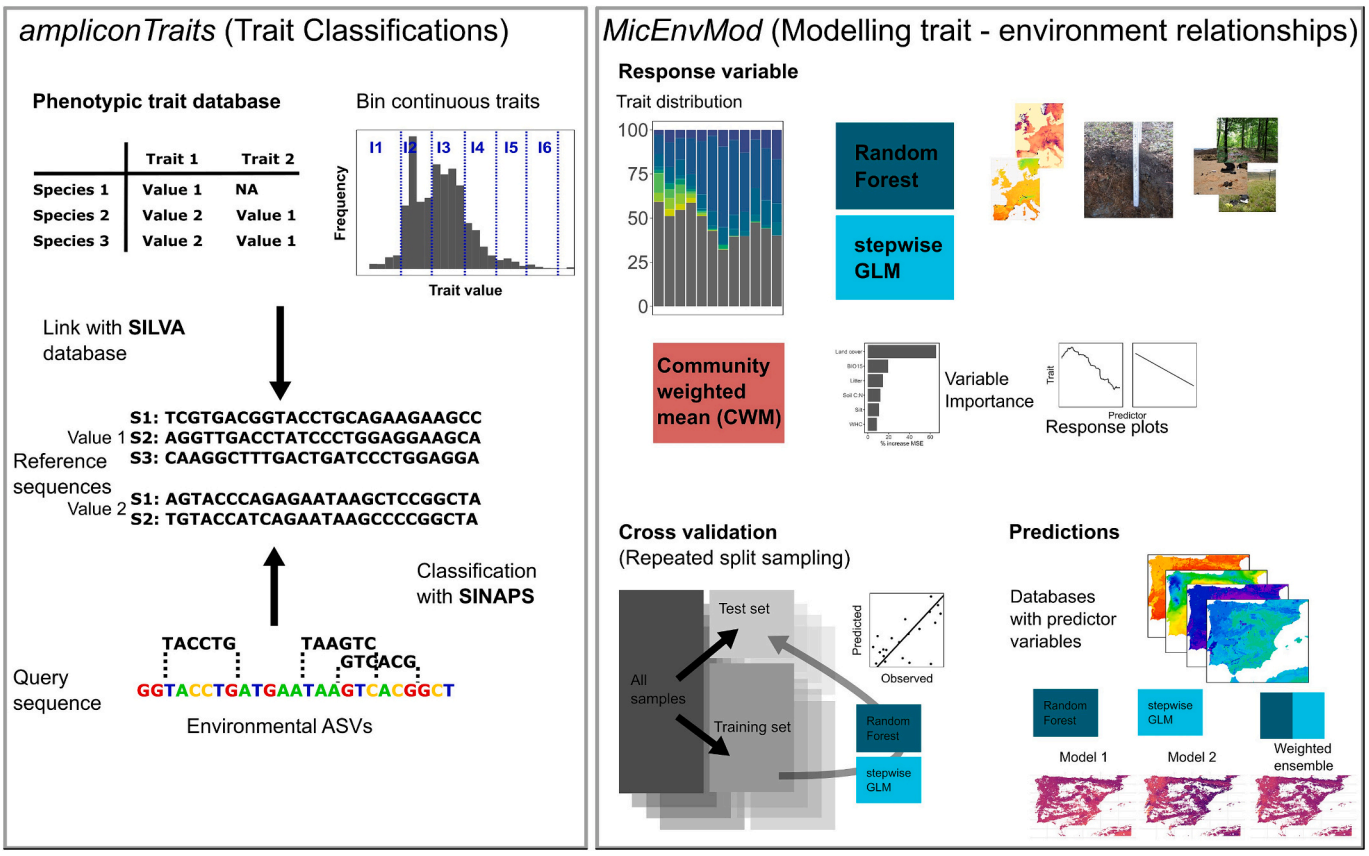


Fig. 1. Overview of the workflow. *ampliconTraits* trait sequence databases were created by linking species from a phenotypic trait database with the SILVA database. Continuous traits were binned into discrete intervals to obtain categories for classification. Environmental sequences can be classified with SINAPS enabling investigation of abundance distributions of the levels of each trait. The R package *MicEnvMod* provides functions to model CWMs obtained from trait distributions. CWMs can be assessed with two types of models (random forest and stepwise GLMs) allowing to identify the most important environmental predictors and to evaluate the trait responses to specific predictors. Individual models and weighted ensemble models can be cross validated by repeated split sampling. Validated models can then be used to predict traits for regions beyond the study sites using georeferenced databases of predictor variables. ASV = amplicon sequence variant.

and determined predictive accuracy by comparing the predictions to the true values. For continuous traits, we calculated the deviation from true values based on the mean of the intervals, while for categorical values, we calculated the percentage of correct predictions as a metric.

We then classified environmental ASVs and tested how the classification of traits was affected by the choice of the number of intervals and the bootstrap cutoff to consider an ASV as classified. A markdown for all analyses with environmental ASVs is provided in the supplementary information. We evaluated the distribution of bootstrap values as well as sequence identities with the reference database across the dataset. To this end, we calculated CWMs as an index of community level trait distribution, as commonly used in macroecology (Daou et al., 2021; Garnier et al., 2004). We compared CWMs for different numbers of intervals as well as for different bootstrap cutoffs creating a correlation matrix with Pearson correlation coefficients using the function *ecospat.cor.plot* in the *ecospat* package (Broennimann et al., 2023). For downstream analyses, we considered trait predictions with a bootstrap value >70 and a sequence identity of >80 % with the top hit as classified.

2.4. Modelling traits with environmental predictors using MicEnvMod

To model genome size with environmental predictors (see supplementary information for code), we created a set of ecologically relevant climatic, edaphic and vegetation-related properties based on measured and database-extracted data. From the full set of predictor variables, we then removed collinear variables with a variance inflation factor > 10 using the function *vifstep* from the *usdm* package (Naimi et al., 2014) resulting in a set of 16 variables. We used RF models, calculated with the function *randomForest* from the package *randomForest* (Liaw and Wiener, 2022) with 1000 trees. Moreover, we used stepwise generalized linear models (GLM) with the Akaike information criterion (AIC) as stopping criterion and selection in both directions using the function stepAIC from the package *MASS* (Venables and Ripley, 2002). For the response variable (CWM of genome size, calculated as abundance weighted average), we used the average of five biological replicates. For the GLM, we assessed the distribution of the response variable using the function *fitdist* from the package *fitdistrplus* (Delignette-Muller and Dutang, 2015) based on the AIC and diagnostic plots.

We then evaluated the importance of each predictor variable based on the increase in mean squared error (MSE) when permuting the variable of interest. For RF, variable importance is implemented in the *randomForest* package. For the GLM, we created the function *VarImp.glm* to obtain an analogous output. To evaluate the direction and magnitude of the response to each predictor variable, we calculated response plots where all predictor variables except the variable of interest are fixed to their mean (continuous variables) or the first level (categorical variables; Elith et al., 2005). To this end, we created the functions *respMono* and *respBi* that take as input a model, a table of predictor variables as well as graphical settings for the plots. The function *respBi* creates bivariate response plots for all combinations of two predictor variables as heatmaps.

To evaluate model accuracy, we created the function *crossVal* to perform a repeated split sampling procedure (Thuiller et al., 2009, 2023), which we applied with default settings. The data is split into a training and a test set at a user-defined ratio (by default 70 to 30 % of the values), the model is trained on the training set and model predictions

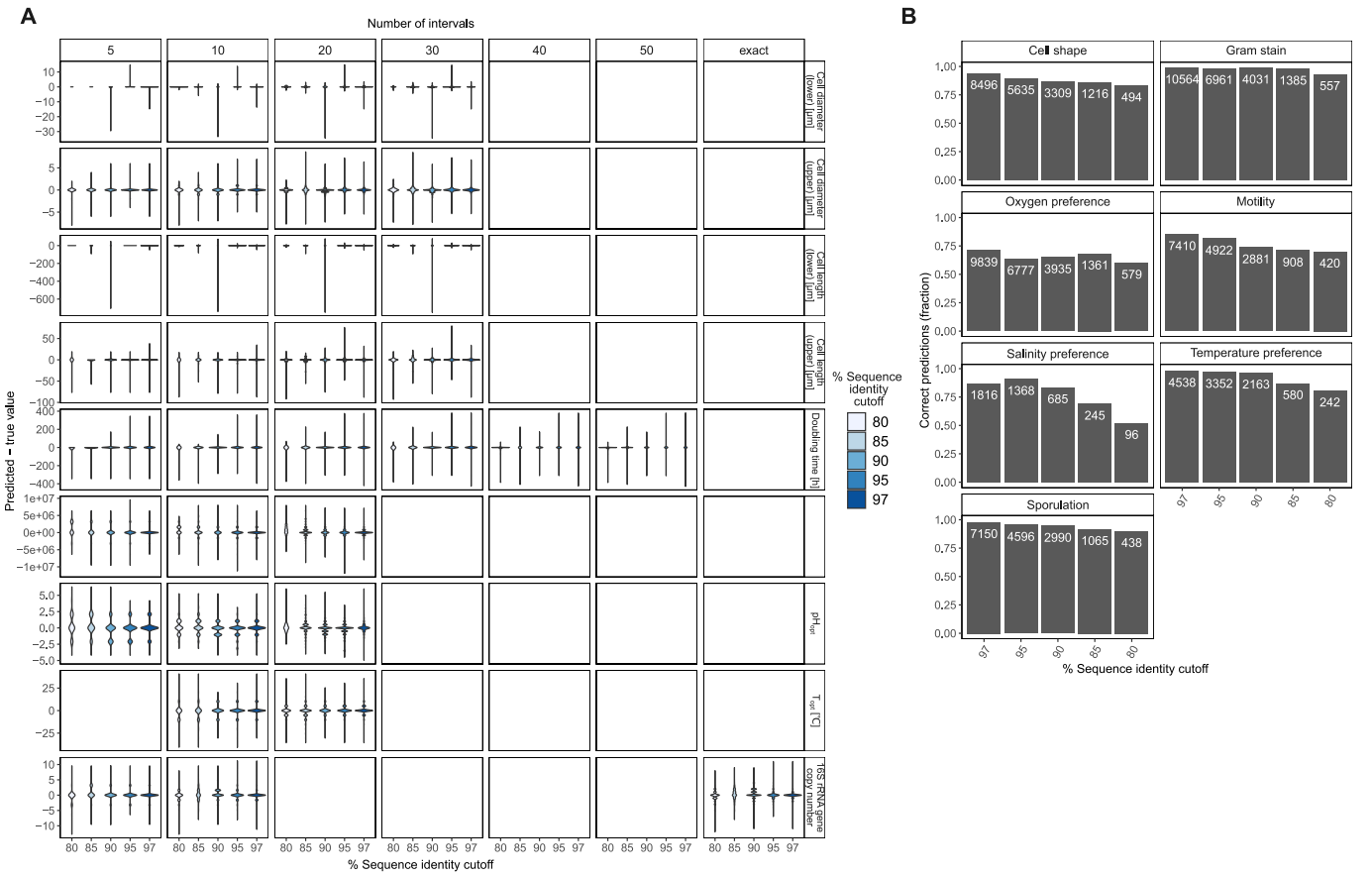


Fig. 2. Cross validation by identity (accuracy of trait prediction as a function of sequence identity with the top hit in the reference database). **A** Frequency distribution of accuracy for different numbers of intervals for continuous traits. **B** Fraction of correct predictions for categorical traits. White numbers indicate the number of comparisons. pH_{opt} = optimum pH, T_{opt} = optimum temperature.

for the test set are compared with the true values. Available metrics of accuracy are correlations between observed and predicted values as well as MSEs. The procedure is repeated multiple times (default 200) and averaged. If the model contains categorical variables, the test set cannot contain levels that are not present in the training set. In that case, for each round, the random subsampling of the data is repeated until all levels in the test set are also present in the training set.

A problem with stepwise model selection is that different combinations of variables can result in similar model fit, making the selection of variables arbitrary (Harrell, 2015). Therefore, with the function *crossVal.step,* we also implemented a split sampling procedure where the stepwise selection is conducted in each round of split sampling allowing to assess the dependence of variable selection on the dataset being used. The fraction of cross validation runs where the variable appears in the model is used as an indicator of the variables' robustness across different datasets.

Implemented via the function *crossVal.ensemble*, we evaluated the performance of a combination of RF and stepwise GLM predictions, weighted by the mean Pearson correlation coefficient between observed and predicted values from the individual cross validation runs in a repeated split sampling procedure as for the individual models.

For the RF model, the number of predictors (16) was relatively high for the number of data points (70 sites). Therefore, in addition to the model with all predictors, we created a model with only the most important predictors where we chose the number of predictors based on the number of variables in the GLM after stepwise selection. This model performed better than the model with all variables, therefore the ensemble model and predictions were calculated only with the RF model with reduced number of predictors.

To predict genome size to new regions, we repeated the modelling procedure including only variables that can be obtained from databases. These variables included TOC, soil C:N ratio, pH, silt and clay content, MAP, BIO5, BIO7, BIO15, WHC and land cover. To demonstrate the process, we predicted genome size for the Iberian Peninsula, which was chosen because it contained a high number of data points. We calculated predictions for the RF model with reduced number of variables, the stepwise GLM, an ensemble model between the two as well as the standard deviation between the two models. Raster layers were cropped to the desired extent, aligned and/or reprojected to the same resolution and projection using the functions *resample* and *projectRaster* from the *raster* package (Hijmans, 2022). In particular, land cover from the MODIS product MCD12Q1_LC1 was reprojected from sinusoidal to World Geodetic System 1984 (WGS84) projection.

3. Results

3.1. Description of the workflow

We created a workflow to model the spatial distribution of prokaryotic traits inferred from amplicon sequence data (see Fig. 1 for an overview). The first part of the workflow (*ampliconTraits)* enables trait classification of environmental sequences. *ampliconTraits* includes preformatted trait sequence databases for the amplicon 341F – 806R of the V3 – V4 region of the 16S rRNA gene enabling classification with SINAPS (Edgar, 2017, 2018). Currently supported traits are cell diameter and length, doubling time, pH optimum, temperature optimum, 16S rRNA gene copy numbers, genome size, oxygen preference, salinity preference, gram stain, sporulation, cell shape and motility (Table 1).

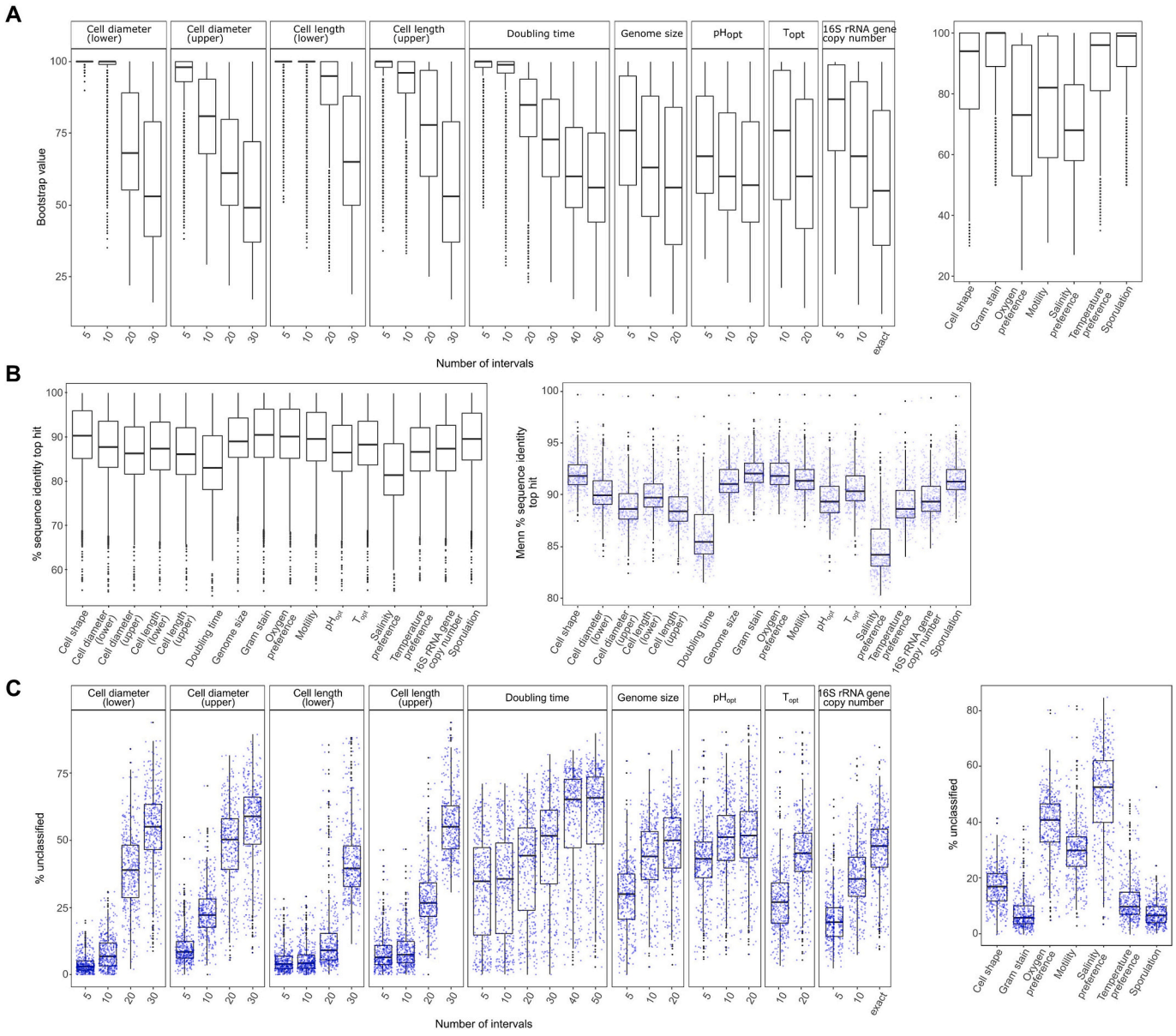


Fig. 3. **A** Bootstrap value across all ASVs in the dataset for all traits **B** Sequence identity between environmental sequences and the top hit in the reference database for all ASVs in the dataset (left) and average per sample (right) **C** Fraction of unclassified sequences for each trait. ASVs with a bootstrap value >70 and sequence identity with the top hit >80 % were considered classified. For continuous traits different numbers of intervals are shown.

Continuous traits were binned into discrete intervals to enable classification and we implemented 2–5 versions of the database with different interval numbers for each continuous trait. Table 1 shows the number of species and sequences available for each trait. We provide detailed documentation on the construction of *ampliconTraits* (https://github.com/jdonhauser/ampliconTraits), enabling users to create customized databases, for instance for a different amplicon or with customized trait annotations. The second part of the workflow is the R package *MicEnvMod* (https://github.com/jdonhauser/MicEnvMod). *MicEnvMod* provides functions to model relationships between a continuous measure of community-level traits and environmental predictors, combining the strengths of multiple model types. Currently, *MicEnvMod* supports RF models and stepwise GLMs. *MicEnvMod* functions allow to assess the importance of predictor variables as well as the response of a dependent variable to a specific predictor variable. Moreover, models can be cross validated by repeated split sampling and stability of variable selection by stepwise GLMs can be examined. Validated models can be used to predict CWMs under different environmental conditions such as under future climate change or different areas. The package includes example data and code for all functions.

3.2. Cross validation of trait predictions from ampliconTraits

First, we assessed the performance of trait classifications with *ampliconTraits*. We cross validated trait predictions as a function of similarity with the reference database by splitting the trait reference databases in a test and a training set at different cutoffs for sequence identity with the most similar reference sequence (Edgar, 2018). This approach enables evaluating accuracy under a range of realistic scenarios with limited similarity between environmental sequences and reference sequences and thus takes into account the size and phylogenetic coverage of the database. For continuous traits, we also compared different interval numbers. For all continuous traits, the accuracy (predicted value minus true value) was similar for all interval numbers and decreased with decreasing sequence identity between query and reference sequence (Fig. 2A). Nonetheless, even at 80 % sequence identity,

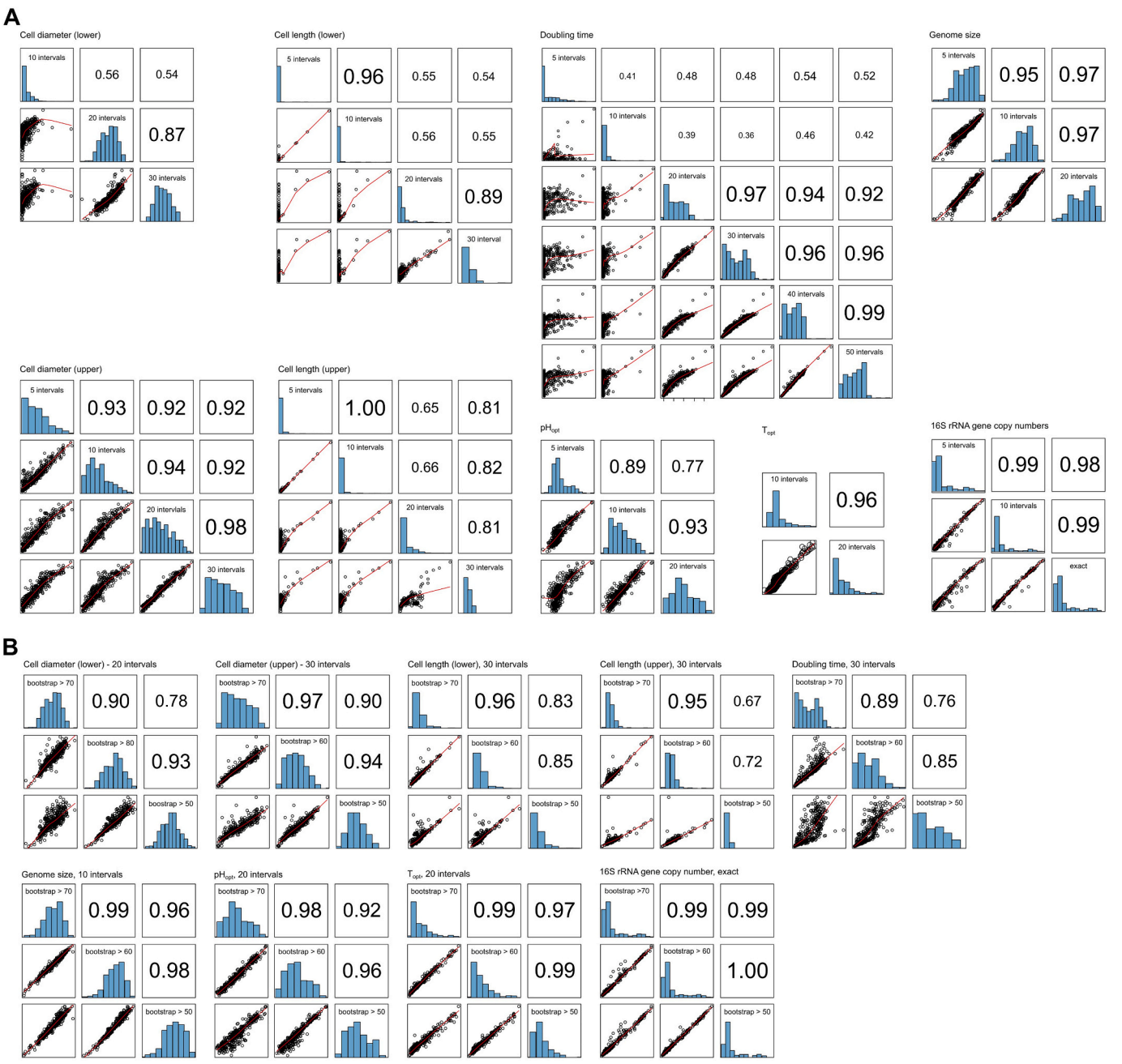


Fig. 4. **A** Correlation matrices for the abundance weighted average for different numbers of intervals for each continuous trait where ASVs with a bootstrap value >70 and > 80 % sequence identity with the top hit were considered as classified. **B** Correlation matrices for the abundance weighted average for different bootstrap cutoffs to consider and ASV as classified.

the greatest fraction of predictions was in the correct interval for all trait-interval combinations. For categorical traits (oxygen preference, motility, salinity preference and sporulation), the number of correct predictions decreased with decreasing sequence identity with the strongest decrease for salinity preference, where 52 % of the predictions were correct at 80 % sequence identity (Fig. 2B). Collectively, trait classification was highly accurate for both continuous and categorical traits, demonstrating the suitability of the approach for environmental sequences.

3.3. Validation of trait classifications with real-world data

Next, we examined the performance of *ampliconTraits* classifications on environmental amplicon sequences, using a large trans-continental soil dataset with 80 sites. The dataset contained 4,479,657 high quality 16S rRNA gene amplicon sequences (11,227 ± 3445 per sample) that formed 24,173 ASVs (149 ± 46 per sample). We used this dataset to evaluate how the choice of classification settings such as the bootstrap cutoff to consider a sequence classified and the number of intervals for continuous traits affects downstream results. For all traits, bootstrap values decreased with increasing number of intervals and variability across the ASVs in the dataset increased (Fig. 3A). For all traits, the ASVs showed a large range of sequence identity with the top hit in the reference database, with a median between 85 and 90 % for most traits (Fig. 3B).

Next, we examined how closely the query sequences were related to the database sequences, as this would impact the classification success. For doubling time and salinity preference, which were the traits with the

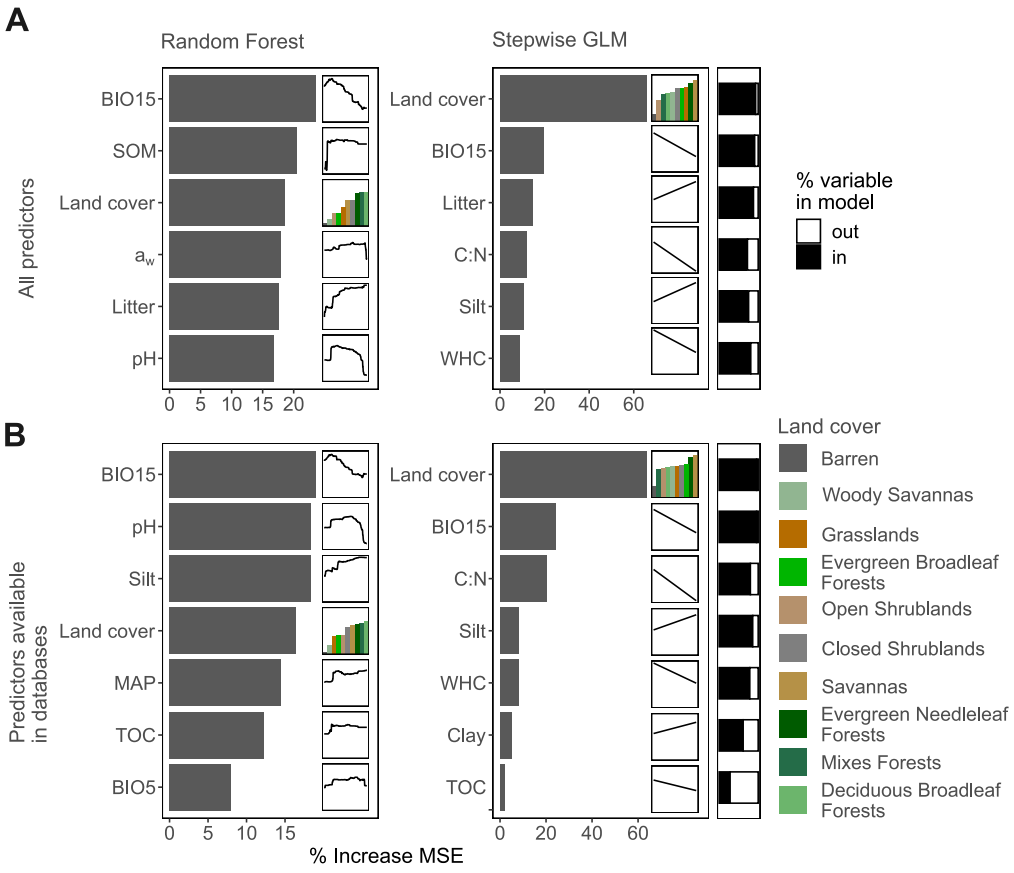


Fig. 5. Variable importance and response plots for a random forest model with the most important variables and a stepwise GLM. **A** For all non-colinear predictors. **B** For predictors available in databases. For the RF model, results from the model with the most important predictors are shown. The small panels indicate how the response variable (genome size, y-axis) responds to a particular predictor (x-axis). For all response plots for the same model, the y-axis has the same scale. For the GLM, the black fraction of the rectangles on the right indicates the fraction of cross validation runs with stepwise selection in which the variable appeared in the model. This indicates how stable the model is across different subsets of the data. MSE = mean squared error, a_w = water activity, MAP = mean annual precipitation, BIO5 = maximum temperature warmest month, BIO15 = precipitation seasonality, WHC = water holding capacity, SOM = soil organic matter, TOC = total organic carbon.

smallest reference databases, sequence identities with the top hit were lower compared to the other traits, suggesting that these traits were more difficult to classify. Average sequence identity with the top hit per sample varied considerably across the dataset, spanning a range of ~10 % for all traits (Fig. 3B), indicating that representation of environmental sequences by the reference database depends on the sample type. We then tested how bootstrap values (i.e. robustness of the classifications) depended on sequence identity and if there was a threshold of sequence identities below which classifications became more unreliable. As expected, we found increasing bootstrap values with increasing sequence identity between query and reference sequences for all traits except for cell length and width as well as doubling time with 5 and 10 intervals where bootstrap values were close to 100 even for very dissimilar sequences (Fig. S2A). Bootstrap values displayed, however, a considerable variation at similar sequence identities (Fig. S2A, B). For most traits, the relationship between bootstrap values and sequence identities showed a plateau with high bootstrap values for sequence identities >85 % (Fig. S2B).

We then evaluated classification success for all traits and intervals where we considered ASVs with a bootstrap support >70 classified, which is in the range of cutoffs commonly used for taxonomic classifications (Bokulich et al., 2018). To consider that a classification may be unambiguous (i.e. has a high bootstrap value) because the correct alternative is not present in the database, we also used the sequence identity between query and reference as an indicator for the quality of the classification and considered ASVs with a sequence identity >80 % as classified. The fraction of unclassified sequences varied strongly across traits and samples and increased with increasing numbers of intervals for continuous traits (Fig. 3C). Our findings indicate a trade-off between the number of intervals (i.e. resolution of the classification) and the number of classified sequences. That is a lower number of intervals results in a higher fraction of classified ASVs, but with coarser resolution.

To address the impact on downstream analyses, we next examined how the number of intervals for continuous traits affected trait distributions among the microbial community across samples. To this end, we calculated CWMs as an index of ecosystem function, which we then compared for different numbers of intervals (Fig. 4A). For doubling time, cell diameter and cell length, 5 and 10 intervals resulted in a skewed distribution of CWMs across samples, with most values in the lowest interval, indicating insufficient resolution. For these three traits, CWMs for 5 and 10 intervals showed low correlation with those for higher number of intervals. For all traits, with higher numbers of intervals, CWMs were highly correlated. For genome size, optimum pH, optimum temperature and 16S rRNA gene copy numbers, also CWMs for 10 intervals were strongly correlated with those for higher number of intervals and for genome size and 16S rRNA gene copy numbers also CWMs for 5 intervals. Our findings indicate that the number of intervals does not affect the outcome of downstream analyses provided that it is high enough to allow for sufficient resolution. Given the large variation of bootstrap values and fractions of unclassified sequences across samples, the range of intervals where CWMs are robust to the choice of

Table 2
Cross validation by repeated split sampling. Pearson correlation coefficients between observed and predicted genome size (mean ± standard deviation of 200 split sampling runs). Ensemble indicates a combined model including the stepwise GLM and the random forest model with the most important variables with predictions weighted by the cross validation results for the individual models.

	Random forest (all non-colinear variables)	Random forest (most important variables)	Stepwise GLM	Ensemble
All predictors	0.59 ± 0.12	0.66 ± 0.12	0.70 ± 0.10	0.70 ± 0.10
Predictors available in databases	0.56 ± 0.12	0.57 ± 0.12	0.66 ± 0.11	0.67 ± 0.10

interval number may however depend on the dataset.

We then tested if the choice of the bootstrap value for considering an ASV as classified affected CWMs comparing cutoffs of 50, 60 and 70 while we kept the number of intervals constant. For genome size, T_{opt} and 16S rRNA gene copy numbers, CWMs were highly correlated across different bootstrap cutoffs ($r > 0.97$; Fig. 4). For cell diameter (upper) and pH_{opt}, some smaller differences were found comparing bootstrap cutoffs of 50 and 70 ($r = 0.9$ and 0.92, respectively). For the remaining traits (cell diameter (lower), cell length (lower), cell length (upper) and doubling time), CWMs for different bootstrap values were still clearly correlated, but showed visible differences, particularly when comparing bootstrap cutoffs of 50 and 70 (r between 0.67 and 0.83). We also compared CWMs for each bootstrap cutoff with and without the additional criteria of sequence identity >80 %, which were highly correlated (all $r = 1.00$, except for doubling time where the smallest r was 0.97). As it is unclear if less stringent annotations (i.e. with lower bootstrap values) or a larger fraction of unclassified sequences cause more uncertainty in the downstream analyses, we kept the bootstrap cutoff of 70 and the sequence identity cutoff of 80 % for all further analyses.

3.4. Modelling genome size with environmental predictors with *MicEnvMod*

The second part of our workflow is to identify environmental predictors of community-level microbial traits. We used genome size as an example to demonstrate environmental modelling with *MicEnvMod* because this trait has widely been used in the ecological literature (e.g. Hessen et al., 2010; Lear et al., 2017; Liu et al., 2023; Sabath et al., 2013). Predictors included ecologically relevant climatic variables, vegetation properties and soil physicochemical parameters obtained from databases or measured from the same soil samples that were used for DNA isolation. *MicEnvMod* leverages the strengths of combining different model types (currently RF and stepwise GLMs). After removing colinear variables, the following predictors were retained: in situ soil temperature at the time of sampling, a_w, total litter, litter total C, litter C:N, soil C:N, pH, silt content, clay content, MAP, BIO5, BIO7, BIO15, WHC, SOM and land cover.

First, we identified key predictors of genome size. RF analysis revealed BIO15, SOM, land cover, a_w and pH as the most important predictors, with relatively small differences in importance between these variables (Fig. 5). In the stepwise GLM, land cover was by far the most important predictor, followed by BIO15, litter, soil C:N, silt content and WHC. These variables appeared in the model in most cross validation runs when the stepwise selection was performed during the cross validation, indicating that the variables were robust across different subsets of the data. Both models predicted decreasing genome sizes with increasing precipitation seasonality and with decreasing litter content. Barren soils harbored the smallest genomes while different types of forests harbored the largest genomes, although with different order between the two model types. Subsequently, we evaluated the models' performance to predict new data. Cross validation by repeated split sampling resulted in an average Pearson correlation coefficient between observed and predicted values of 0.59 ± 0.12 for the RF model with all variables and 0.70 ± 0.10 for the GLM (Table 2). An RF model with only the most important variables (same number of variables as in the stepwise GLM) performed better than the one with all non-colinear variables. A weighted ensemble of the RF model with the most important variables and the GLM performed equally well as the GLM alone (Table 2). Collectively, we show that both RFs and GLMs performed well

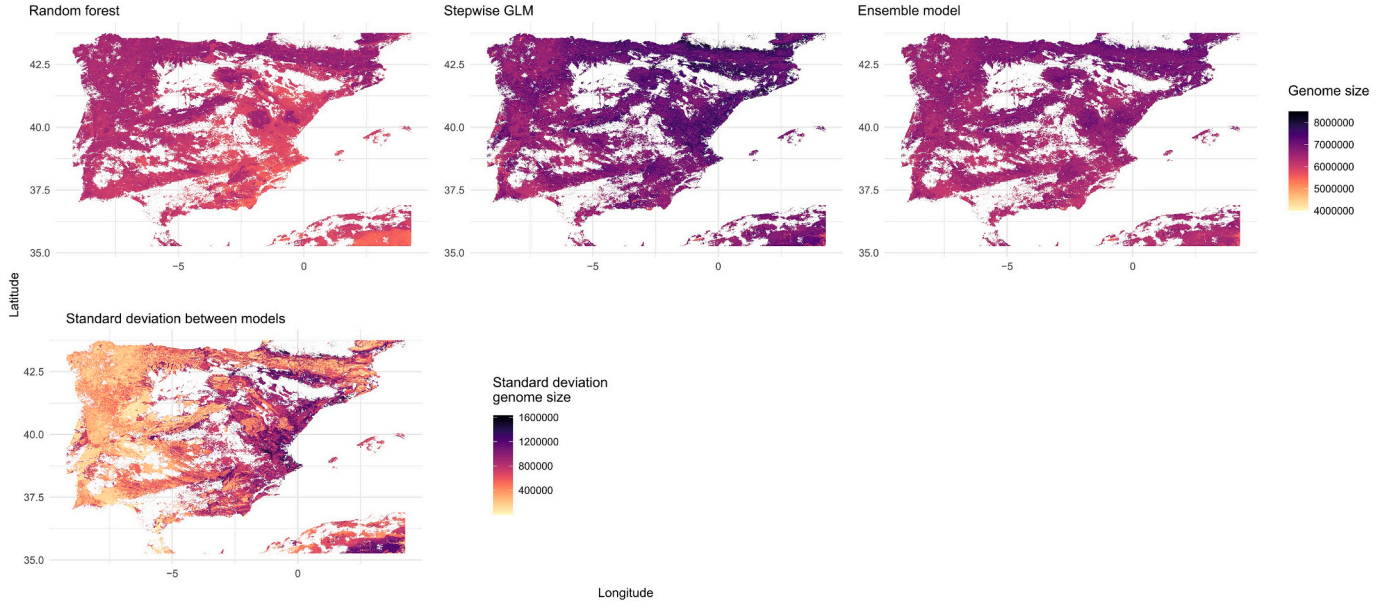


Fig. 6. Predictions of soil prokaryotic genome size for the Iberian Peninsula for the RF model, the stepwise GLM and a weighted ensemble between the two (top row) as well as the deviation between the two model types (bottom row). White areas correspond to land cover classes that were not present in the dataset used to build the models and therefore could not be predicted.

in predicting and genome sizes and that performance of the random forest model can be further improved by reducing the number of predictor variables and thus reducing overfitting.

Besides revealing the drivers of microbial traits such as genome size, we aimed to predict traits in regions beyond the study area. Therefore, we repeated the modelling procedure described above with environmental variables available in databases. After exclusion of colinear variables, these variables were: soil C:N, TOC, pH, silt content, clay content, MAP, BIO5, BIO7, BIO15, WHC and land cover. Like for the models with all variables, BIO15 and land cover were the most important variables for both the RF model and the stepwise GLM. Cross validation showed slightly lower correlation between observed and predicted values compared to the models with all variables (Table 2). A weighted ensemble between the RF model with the most important variables and the GLM performed marginally better than each of the models individually. To demonstrate extrapolation of model results to new regions and environmental conditions, we predicted genome size for the Iberian Peninsula based on the RF model with the most important variables, the GLM and the ensemble model. Both models predicted the largest genomes in the north, which is the region with the lowest precipitation seasonality (Fig. 6). The RF model generally tended to predict smaller genomes than the GLM and predicted the smallest genomes at the east coast, which is also the region with largest discrepancy between the models. Predictions by the RF model for this region are likely driven by pH for which it predicts a hump shaped relationship with small genomes at high pH values and which is not present in the GLM. Conversely, the GLM predicted the smallest genomes in the southwest, which is the region with the highest precipitation seasonality. Collectively, we demonstrate how the combination of multiple models can strengthen predictions and inform of predictive uncertainties.

4. Discussion

With *ampliconTraits,* we present a marker gene trait sequence database that enables users to infer ecologically relevant information from marker genes for multiple prokaryotic phenotypic traits. We showed that using *ampliconTraits* with SINAPS (Edgar, 2017) enables accurate predictions for amplicon sequences with similarity to reference sequences as low as 80 %, which roughly corresponds to order level (Yarza et al., 2014). This confirms that our approach is reliable to infer traits for environmental DNA or cDNA sequences where often no closely related reference sequence is available and indicates that the traits assessed here are deeply conserved in the phylogeny of prokaryotes. Previous studies indicated that the phylogenetic conservation of traits depends on their complexity, i.e. the number of genes involved and thus the number of mutations required to change the value of the trait, with more complex traits being more deeply conserved in the phylogeny (Martiny et al., 2012, 2015). However, we found a gradual decrease in accuracy with decreasing sequence identity. For continuous or ordinal traits, this suggests that the value of the trait changed gradually throughout evolution as genetic alterations accumulated. For discrete traits where no gradual transition across values of the trait is possible, phylogenetic conservation may vary across different groups and thus change gradually on average.

We validated trait inference of environmental sequences with *ampliconTraits* with a large dataset based on 16S rRNA gene amplicon sequences and showed that CWMs are robust to settings such as the number of intervals for continuous traits and the bootstrap cutoff where an ASV is considered classified. We were able to classify 40–60 % of the sequences on average, despite the still relatively small databases and the high dissimilarity between environmental sequences and reference sequence for a large fraction of the dataset. We found a large range of bootstrap values and sequence identities with the top hit and consequently a large range in classification success both within and across samples. These findings indicate that representation of the prokaryotic community by the reference database strongly depends on the type of sample. A further limitation is currently the relatively small size of the trait databases, which comprise ~500–10,000 species per trait compared to taxonomic databases like the SILVA SSU database, which contains ~35,000 prokaryotic species. However, trait classifications will significantly improve as reference databases grow. Particularly classifications for the traits doubling time and salinity preference with <1000 species will benefit from more comprehensive databases.

An important advantage of our method compared to existing approaches is that robustness of the classification is evaluated by bootstrapping, that it allows for cross validation of trait inferences, and that it does not depend on the representation of phylogenetic relationships by taxonomy. For instance, a previous amplicon-based trait inference approach annotated environmental DNA sequences by cross-mapping the taxonomic affiliation of the sequences to the taxa in the trait database (Cébron et al., 2021). If there were multiple values for a taxonomic group assigned to an environmental sequence, the authors averaged trait values across all taxa. Opposed to this taxonomy-based averaging approach, our approach avoids classifying sequences beyond the level of phylogenetic conservation of the trait.

Another study using amplicon-based trait inference that was limited to few traits however (genome size, 16S rRNA gene copy numbers, oxygen requirement and motility; Gravuer and Eskelinen, 2017) employed phylogenetically independent contrasts ancestral state reconstruction (Garland Jr. and Ives, 2000; Kembel et al., 2012). This approach measures uncertainty of the trait estimation based on the length of the connecting branches in the phylogenetic tree and allows to assess accuracy by leave-one out cross-validation. An advantage of ancestral state reconstruction is that it directly estimates trait values for continuous traits without the necessity to bin them into discrete intervals. In contrast, an advantage of the word-based classification approach used here is that it does not depend on phylogenetic trees and associated evolution models. Moreover, opposed to the leave-one-out cross-validation implemented for ancestral state reconstruction, the cross-validation approach used here considers the similarity between query and reference sequences.

Metagenomics constitute an alternative approach to infer soil microbial traits. An advantage of *ampliconTraits* compared to metagenomics is the large number of taxa recovered. For instance, our dataset contained >20,000 ASVs, while the number of MAGs from soil metagenomes typically range from 30 to 500, depending on sequencing depth, diversity of the sample and quality cutoffs (Kroeger et al., 2018; Sipes et al., 2021; Wu et al., 2022, 2023). In the case of genome size, which is one of the most widely studied microbial traits, most previous studies were based on complete genomes or metagenomes (Chuckran et al., 2021, 2022; Rodríguez-Gijón et al., 2022; Sabath et al., 2013), sometimes combining several thousand of genomes from multiple resources but with limited metadata (Chuckran et al., 2021; Rodríguez-Gijón et al., 2022). With *ampliconTraits* and SINAPS, we were able to classify 10,171 ASVs and were able to draw on comprehensive metadata to identify environmental drivers of genome size in downstream analyses. An advantage of the metagenome-based approach is, however, that it does not depend on primers causing biases in the representation of taxa. Further, it does not depend on the phylogenetic conservation of the traits while the amplicon-based approach requires reference sequences that are more similar than the level of phylogenetic conservation. Nonetheless, both approaches depend on reference databases and a significant fraction of metagenomics sequences can currently not be classified (Choi et al., 2017; Donhauser et al., 2021).

Classification of multiple traits with *ampliconTraits* further allows for identification of trade-offs among traits by evaluating trait co-occurrences within taxa. Trade-offs are a key concept in ecology explaining species co-occurrence (Kneitel and Chase, 2004) and are linked to ecosystem C-cycling (Malik et al., 2019). For instance, a trade-off between fast growth and efficient growth has been linked to microbial incorporation of C versus release to the atmosphere (Roller et al., 2016). Thus, *ampliconTraits* has the potential to link microbial

community dynamics with ecosystem C-cycling. Alternatively, trait trade-offs can be inferred from MAGs (Karaoz and Brodie, 2022) based on the presence of many functional genes. This results in complex datasets and may be complicated by incomplete MAGs. Conversely, the amplicon-based approach directly provides trait values.

To link microbial community dynamics to ecosystem functioning it is pivotal to understand adaptation of microorganisms to specific environmental conditions, mediated by their traits. *MicEnvMod* provides functionality to identify key predictors of microbial traits, to examine specific trait-predictor relationships and to evaluate model performance. Demonstrating the application of *MicEnvMod* for prokaryotic genome size, we revealed precipitation seasonality and land cover as the most important drivers. Soils with high precipitation seasonality, which are likely exposed to prolonged periods of drought each year, as well as barren soils harbored the smallest genomes. These findings are in accordance with previous studies that associated small genomes with aridity, low net primary production, nutrient poor conditions and deserts (Chuckran et al., 2022; Gravuer and Eskelinen, 2017; Liu et al., 2023; Simonsen, 2022). Small genomes may serve to reduce the energetic cost in nutrient-poor and physiologically challenging environments and enable fast replication with the onset of favorable conditions such as precipitation after a long drought. In line with this notion, Sabath et al. (2013) found streamlined genomes and low replication times in thermophilic prokaryotes.

(Liu et al., 2023) used joint species distribution models (JSDM) as an alternative modelling approach to assess genome size – environment relationships. JSDMs predict species occurrences or abundances based on environmental parameters and test if species niches across environmental gradients (represented by the regression coefficients from these models) are explained by species-specific traits (Ovaskainen et al., 2017; Tikhonov et al., 2020). A strength of JSDMs compared to our method is that they can account for phylogenetic dependencies that may affect species distributions rather than environmental selection on traits. However, because each taxon is modelled individually, JSDMs for microbial communities involve a very large number of models representing taxa with little a priori knowledge about their environmental niche. In the study by Liu et al. (2023), this approach was feasible because they aggregated ASVs at the genus level excluding unclassified taxa resulting in 143 genera and because they used presence-absence data. To apply this approach in our study at the ASV level based on abundances, we would need to model 10,171 ASVs. Under these circumstances, it becomes complicated to implement a suitable distribution for sequencing-derived count data (Love et al., 2014), to choose environmental predictors, to compare alternative models with different sets of predictors or different model types, and to cross-validate models. Therefore, due to the complexity of microbial communities, we chose to model CWMs as an aggregated index of species traits, allowing to extend the framework easily to other model types and to perform cross-validation of individual and ensemble models.

MicEnvMod currently supports two different types of models and implements cross validation of weighted ensemble models based on metrics suitable for continuous data. Different model types capture different properties of the data and ensemble models can therefore outperform individual models (Araújo and New, 2007). For instance, GLMs are restricted to linear relationships between predictor and response variables (or other defined polynomials), while RF models can fit any kind of relationship. This is exemplified by the relationship between genome size and pH in our dataset where RF analysis revealed a hump shaped curve with an abrupt drop at high values while pH was not among predictors in the stepwise GLM. In the future, further model types such as generalized additive models, which allow a more flexible relationship between response and predictor variables than GLMs, as well as further machine learning algorithms will be added to *MicEnvMod*. To comprehensively evaluate soil ecosystem functioning, it is important to create comprehensive data resources of microbial indicators such as the traits in *ampliconTraits* through a large range of climatic conditions in space and time. Using georeferenced databases of climatic and soil physico-chemical properties, models validated with *MicEnvMod* can be used to predict prokaryotic traits through space and assess differences between predictions from different models, as demonstrated for genome size across the Iberian Peninsula. RF models and GLMs agreed on the relationship between genome size and the most important predictor precipitation seasonality as well as on smaller genomes in barren compared to vegetated soils and consistently predicted the largest genomes in the northeast. A caveat for building models with variables derived from databases is that the predictors may come with considerable uncertainty themselves. Due to limited spatial resolution and/or heterogeneity of the terrain, the value of a grid cell may poorly capture the properties at the specific sampling location. This is particularly the case for soil properties, while climate is expected to be more constant across space. It is important to note that predictions from the two model types varied considerably in some regions due to different implementation of predictors as outlined above for pH. This highlights the importance of considering multiple models to substantiate predictions and estimate uncertainty and suggests that combination of further model classes beyond RFs and GLMs may be needed to explore uncertainty across different models and space. Beyond different types of models, the principle of ensemble modelling can be extended to averaging models trained with different initial conditions such as subsets of the data used the cross validation (Araújo and New, 2007; Thuiller et al., 2017). Functionality to make ensemble predictions from all models within a cross validation run will be included in future versions of *MicEnvMod*.

In conclusion, we describe a novel and robust approach to infer prokaryotic traits from 16S rRNA gene amplicon data, to identify environmental predictors of trait distributions, and to predict community level traits to new conditions. *ampliconTraits* significantly contributes to overcoming the limited ecological interpretability of sequencing-based investigations of microbial communities by integrating a comprehensive trait sequence database with a state-of-the-art classification algorithm. Currently, a key limitation of *ampliconTraits* is the size of the trait databases. Thus, future work should invest into measuring traits in a larger number of isolates with a particular focus on understudied environments like soil. Further phenotypic traits such as carbon use efficiency would be highly relevant for ecosystem C-cycling and could be implemented into *ampliconTraits*. In addition, eukaryotic microorganisms, particularly fungi, play an important role in soil processes. Thus, an analogous approach would be desirable. Although the *ampliconTraits* database is formatted for use with SINAPS it can be used with other algorithms in principle and thus provides a resource for further development of trait inference tools. *ampliconTraits* and *MicEnvMod* may be harnessed for instance to inform and validate trait-based biogeochemical models that represent the effect of microbial community dynamics on soil C-cycling processes. Thus, by improving such models, our workflow may contribute to quantifying CO_2 emissions from soil and thus feedback to climate change, which is currently neglected in earth system climate models. Trait inference with *ampliconTraits* and environmental modelling with *MicEnvMod* further allow to study microbial niche adaptation. Beyond soil C cycling, strategies that enable microorganisms to colonize a habitat play a role in a large range of environments and processes, such as interactions with host organisms ranging from plants to humans. Thus, our trait-inference workflow has the potential to contribute to answering a broad spectrum of research questions in microbial ecology and biogeochemistry.

CRediT authorship contribution statement

Jonathan Donhauser: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization, Writing – review & editing. **Anna Doménech-Pascual:** Writing – review & editing, Investigation. **Xingguo Han:** Writing – review & editing, Investigation. **Karen Jordaan:** Writing – review & editing, Investigation. **Jean-**

Baptiste Ramond: Writing – review & editing, Investigation, Funding acquisition. **Aline Frossard:** Writing – review & editing, Investigation, Funding acquisition. **Anna M. Romaní:** Writing – review & editing, Investigation, Funding acquisition. **Anders Priemé:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

ampliconTraits trait sequence databases and files for database construction are available at https://erda.ku.dk/archives/f5d4b1d41f74ba3d6f73b212dbb11591/published-archive.html. Code to create databases and documentation for *ampliconTraits* are hosted at https://github.com/jdonhauser/ampliconTraits. The R package *MicEncMod* is available at https://github.com/jdonhauser/MicEnvMod. A markdown for all analyses in this manuscript is available in the supplementary information. Raw sequences were deposited in the NCBI Sequence Read Archive under the accession number PRJNA1073882.

Acknowledgements

We thank George Stoletov at the University of Copenhagen for extracting the DNA and Dr. Joseph Nesme for support with denoising amplicon data. We thank Dr. Joan Pere Casas-Ruiz for contributing to sampling soil from the Spanish sites. We acknowledge Pallas core facilities at the university of Copenhagen and the Danish National Life Science Supercomputing Center Computerome for providing access to high-performance computing facilities. Further, we acknowledge the contribution of scientists at the McGill University and Génome Québec Innovation Center, Montréal, Canada, for paired-end Illumina MiSeq sequencing. Funding: This work was supported by the Swiss National Science Foundation [grant number SNF 31BD30_193667]; the Spanish State Research Agency [grant number PCI2020-120702-2/AEI/10.13039/501100011033]; the Innovation Fund Denmark [grant number BiodivClim-76 GRADCATCH]; and the Department of Science and Innovation of the Republic of South Africa [grant number GRADCATCH], through the 2019-2020 BiodivERsA joint call for research proposals, under the BiodivClim ERA-Net COFUND programme.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102817.

References

Allison, S.D., Goulden, M.L., 2017. Consequences of drought tolerance traits for microbial decomposition in the DEMENT model. Soil Biol. Biochem. 107, 104–113. https://doi.org/10.1016/j.soilbio.2017.01.001.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22, 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

Aßhauer, K.P., Wemheuer, B., Daniel, R., Meinicke, P., 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31, 2882–2884. https://doi.org/10.1093/bioinformatics/btv287.

Barberán, A., Caceres Velazquez, H., Jones, S., Fierer, N., 2017. Hiding in plain sight: mining bacterial species records for phenotypic trait information. mSphere 2. https://doi.org/10.1128/mSphere.00237-17 e00237–17.

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma 269, 61–68. https://doi.org/10.1016/j.geoderma.2016.01.034.

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., Gregory Caporaso, J., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 1–17. https://doi.org/10.1186/S40168-018-0470-Z/TABLES/3.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., Supek, F., 2016. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 44, 10074–10090. https://doi.org/10.1093/nar/gkw964.

Broennimann, O., Cola, V.D., Guisan, A., 2023. ecospat: spatial ecology miscellaneous methods. R package version 3.5.1. https://CRAN.R-project.org/package=ecospat.

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

Cébron, A., Zeghal, E., Usseglio-Polatera, P., Meyer, A., Bauda, P., Lemmel, F., Leyval, C., Maunoury-Danger, F., 2021. BactoTraits – a functional trait database to evaluate how natural and man-induced changes influence the assembly of bacterial communities. Ecol. Indic. 130, 108047 https://doi.org/10.1016/j.ecolind.2021.108047.

Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Flater, J., Tiedje, J.M., Hofmockel, K.S., Gelder, B., Howe, A., 2017. Strategies to improve reference databases for soil microbiomes. ISME J. 11, 829–834. https://doi.org/10.1038/ismej.2016.168.

Chuckran, P.F., Hungate, B.A., Schwartz, E., Dijkstra, P., 2021. Variation in genomic traits of microbial communities among ecosystems. FEMS Microbes 2, xtab020. https://doi.org/10.1093/femsmc/xtab020.

Chuckran, P.F., Flagg, C., Propster, J., Rutherford, W.A., Sieradzki, E., Blazewicz, S.J., Hungate, B., Pett-Ridge, J., Schwartz, E., Dijkstra, P., 2022. Edaphic controls on genome size and GC content of bacteria in soil microbial communities. https://doi.org/10.1101/2021.11.17.469016.

Daou, L., Garnier, É., Shipley, B., 2021. Quantifying the relationship linking the community-weighted means of plant traits and soil fertility. Ecology 102, e03454. https://doi.org/10.1002/ecy.3454.

Delignette-Muller, M.L., Dutang, C., 2015. Fitdistrplus: an R package for fitting distributions. J. Stat. Softw. 64, 1–34. https://doi.org/10.18637/jss.v064.i04.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F.T., D'Amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R.G., Hordijk, W., Salamin, N., Guisan, A., 2017. Ecospat: an R package to support spatial analyses and modeling of species niches and distributions. Ecography 40, 774–787. https://doi.org/10.1111/ecog.02671.

Donhauser, J., Niklaus, P.A., Rousk, J., Larose, C., Frey, B., 2020. Temperatures beyond the community optimum promote the dominance of heat-adapted, fast growing and stress resistant bacteria in alpine soils. Soil Biol. Biochem. 148, 107873 https://doi.org/10.1016/j.soilbio.2020.107873.

Donhauser, J., Qi, W., Bergk-Pinto, B., Frey, B., 2021. High temperatures enhance the microbial genetic potential to recycle C and N from necromass in high-mountain soils. Glob. Chang. Biol. 27, 1365–1386. https://doi.org/10.1111/gcb.15492.

Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., Langille, M.G.I., 2020. PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. 38, 685–688. https://doi.org/10.1038/s41587-020-0548-6.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

Edgar, R.C., 2017. SINAPS: Prediction of microbial traits from marker gene sequences. bioRxiv 124156. https://doi.org/10.1101/124156.

Edgar, R.C., 2018. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 6, e4652. https://doi.org/10.7717/peerj.4652.

Elith, J., Ferrier, S., Huettmann, F., Leathwick, J., 2005. The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. Ecol. Model. 186, 280–289. https://doi.org/10.1016/j.ecolmodel.2004.12.007.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

Frey, B., Rime, T., Phillips, M., Stierli, B., Hajdas, I., Widmer, F., Hartmann, M., 2016. Microbial diversity in European alpine permafrost and active layers. FEMS Microbiol. Ecol. 92, fiw018. https://doi.org/10.1093/femsec/fiw018.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS collection 5 global land cover: algorithm refinements and characterization of new datasets. Remote Sens. Environ. 114, 168–182. https://doi.org/10.1016/j.rse.2009.08.016.

Garland Jr., T., Ives, A.R., 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155, 346–364. https://doi.org/10.1086/303327.

Garnier, E., Cortez, J., Billès, G., Navas, M.-L., Roumet, C., Debussche, M., Laurent, G., Blanchard, A., Aubry, D., Bellmann, A., Neill, C., Toussaint, J.-P., 2004. Plant functional markers capture ecosystem properties during secondary succession. Ecology 85, 2630–2637. https://doi.org/10.1890/03-0799.

Gravuer, K., Eskelinen, A., 2017. Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. Front. Microbiol. 8.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9.

Harrell, F.E., 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer Series in Statistics. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-19425-7.

Hessen, D.O., Jeyasingh, P.D., Neiman, M., Weider, L.J., 2010. Genome streamlining and the elemental costs of growth. Trends Ecol. Evol. 25, 75–80. https://doi.org/10.1016/j.tree.2009.08.004.

Hijmans, R.J., 2022. raster: geographic data analysis and modeling. R package version 3.5–29. https://CRAN.R-project.org/package=raster.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., Schuster, S.C., 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21, 1552–1560. https://doi.org/10.1101/gr.120618.111.

Ii, M.S.R., O'Rourke, D.R., Kaehler, B.D., Ziemski, M., Dillon, M.R., Foster, J.T., Bokulich, N.A., 2021. RESCRIPt: Reproducible sequence taxonomy reference database management. PLoS Comput. Biol. 17, e1009581 https://doi.org/10.1371/journal.pcbi.1009581.

Josse, J., Husson, F., 2016. missMDA: a package for handling missing values in multivariate data analysis. J. Stat. Softw. 70, 1–31. https://doi.org/10.18637/jss.v070.i01.

Kaiser, C., Franklin, O., Dieckmann, U., Richter, A., 2014. Microbial community dynamics alleviate stoichiometric constraints during litter decay. Ecol. Lett. 17, 680–690. https://doi.org/10.1111/ele.12269.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462. https://doi.org/10.1093/nar/gkv1070.

Karaoz, U., Brodie, E.L., 2022. microTrait: a toolset for a trait-based representation of microbial genomes. Front. Bioinform. 2.

Kembel, S.W., Wu, M., Eisen, J.A., Green, J.L., 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comput. Biol. 8, e1002743 https://doi.org/10.1371/journal.pcbi.1002743.

Kneitel, J.M., Chase, J.M., 2004. Trade-offs in community ecology: linking spatial scales and species coexistence. Ecol. Lett. 7, 69–80. https://doi.org/10.1046/j.1461-0248.2003.00551.x.

Kroeger, M.E., Delmont, T.O., Eren, A.M., Meyer, K.M., Guo, J., Khan, K., Rodrigues, J.L. M., Bohannan, B.J.M., Tringe, S.G., Borges, C.D., Tiedje, J.M., Tsai, S.M., Nüsslein, K., 2018. New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. Front. Microbiol. 9.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., Huttenhower, C., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31, 814–821. https://doi.org/10.1038/nbt.2676.

Lear, G., Lau, K., Perchec, A.-M., Buckley, H.L., Case, B.S., Neale, M., Fierer, N., Leff, J. W., Handley, K.M., Lewis, G., 2017. Following Rapoport's rule: the geographic range and genome size of bacterial taxa decline at warmer latitudes. Environ. Microbiol. 19, 3152–3162. https://doi.org/10.1111/1462-2920.13797.

Liaw, A., Wiener, M., 2022. Classification and regression by randomForest. R News 2, 18–22.

Liu, H., Zhang, H., Powell, J., Delgado-Baquerizo, M., Wang, J., Singh, B., 2023. Warmer and drier ecosystems select for smaller bacterial genomes in global soils. iMeta 2, e70. https://doi.org/10.1002/imt2.70.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8.

Madin, J.S., 2021. A synthesis of bacterial and archaeal phenotypic trait data. Figshare. https://doi.org/10.6084/m9.figshare.c.4843290.v3.

Madin, J.S., Nielsen, D.A., Brbic, M., Corkrey, R., Danko, D., Edwards, K., Engqvist, M.K. M., Fierer, N., Geoghegan, J.L., Gillings, M., Kyrpides, N.C., Litchman, E., Mason, C. E., Moore, L., Nielsen, S.L., Paulsen, I.T., Price, N.D., Reddy, T.B.K., Richards, M.A., Rocha, E.P.C., Schmidt, T.M., Shaaban, H., Shukla, M., Supek, F., Tetu, S.G., Vieira-Silva, S., Wattam, A.R., Westfall, D.A., Westoby, M., 2020. A synthesis of bacterial and archaeal phenotypic trait data. Sci. Data 7, 170. https://doi.org/10.1038/s41597-020-0497-4.

Malik, A.A., Martiny, J.B.H., Brodie, E.L., Martiny, A.C., Treseder, K.K., Allison, S.D., 2019. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. ISME J. https://doi.org/10.1038/s41396-019-0510-0.

Martiny, A.C., Treseder, K., Pusch, G., 2012. Phylogenetic conservatism of functional traits in microorganisms. ISME J. 7, 830. https://doi.org/10.1038/ismej.2012.160.

Martiny, J.B.H., Jones, S.E., Lennon, J.T., Martiny, A.C., 2015. Microbiomes in light of traits: a phylogenetic perspective. Science 350, aac9323. https://doi.org/10.1126/science.aac9323.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., Toxopeus, A.G., 2014. Where is positional uncertainty a problem for species distribution modelling. Ecography 37, 191–203. https://doi.org/10.1111/j.1600-0587.2013.00205.x.

Nguyen, N.H., Song, Z., Bates, S.T., Branco, S., Tedersoo, L., Menke, J., Schilling, J.S., Kennedy, P.G., 2016. FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecol. 20, 241–248. https://doi.org/10.1016/j.funeco.2015.06.006.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., Abrego, N., 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. 20, 561–576. https://doi.org/10.1111/ele.12757.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596. https://doi.org/10.1093/nar/gks1219.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez-Gijón, A., Nuy, J.K., Mehrshad, M., Buck, M., Schulz, F., Woyke, T., Garcia, S. L., 2022. A genomic perspective across Earth's microbiomes reveals that genome size in Archaea and Bacteria is linked to ecosystem type and trophic strategy. Front. Microbiol. 12.

Roller, B.R.K., Stoddard, S.F., Schmidt, T.M., 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat. Microbiol. 1, 16160. https://doi.org/10.1038/nmicrobiol.2016.160. https://www.nature.com/articles/nmicrobiol2016160#supplementary-information.

Sabath, N., Ferrada, E., Barve, A., Wagner, A., 2013. Growth temperature and genome size in Bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biol. Evol. 5, 966–977. https://doi.org/10.1093/gbe/evt050.

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11, e0163962. https://doi.org/10.1371/journal.pone.0163962.

Sieriebriennikov, B., Ferris, H., Goede, R.G.M.D., 2014. European journal of soil biology short communication NINJA : an automated calculation system for nematode-based biological monitoring. Eur. J. Soil Biol. 61, 90–93. https://doi.org/10.1016/j.ejsobi.2014.02.004.

Simonsen, A.K., 2022. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. ISME J. 16, 423–434. https://doi.org/10.1038/s41396-021-01082-x.

Sipes, K., Almatari, A., Eddie, A., Williams, D., Spirina, E., Rivkina, E., Liang, R., Onstott, T.C., Vishnivetskaya, T.A., Lloyd, K.G., 2021. Eight metagenome-assembled genomes provide evidence for microbial adaptation in 20,000- to 1,000,000-year-old siberian permafrost. Appl. Environ. Microbiol. 87 https://doi.org/10.1128/AEM.00972-21 e00972–21.

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD - a platform for ensemble forecasting of species distributions. Ecography 32, 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x.

Thuiller, W., Guisan, A., Zimmermann, N.E., 2017. Habitat suitability and distribution models: 14 ensemble modeling and model averaging. In: Habitat Suitability and Distribution Models with Applications in R, Ecology. Biodiversity and Conservation. Cambridge University Press, Cambridge, pp. 224–236.

Thuiller, W., Georges, D., Gueguen, M., Engler, R., Breiner, F., Lafourcade, B., Patin, R., 2023. biomod2: ensemble platform for species distribution modeling. 2023. R package version 4.2–4. https://CRAN.R-project.org/package=biomod2.

Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehikoinen, A., de Jonge, M.M.J., Oksanen, J., Ovaskainen, O., 2020. Joint species distribution modelling with the r-package Hmsc. Methods Ecol. Evol. 11, 442–447. https://doi.org/10.1111/2041-210X.13345.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.

Wu, X., Cui, Z., Peng, J., Zhang, F., Liesack, W., 2022. Genome-resolved metagenomics identifies the particular genetic traits of phosphate-solubilizing bacteria in agricultural soil. ISME Commun. 2, 1–4. https://doi.org/10.1038/s43705-022-00100-z.

Wu, X., Bei, S., Zhou, X., Luo, Y., He, Z., Song, C., Yuan, H., Pivato, B., Liesack, W., Peng, J., 2023. Metagenomic insights into genetic factors driving bacterial niche differentiation between bulk and rhizosphere soils. Sci. Total Environ. 891, 164221 https://doi.org/10.1016/j.scitotenv.2023.164221.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W.B., Euzéby, J., Amann, R., Rosselló-Móra, R., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645. https://doi.org/10.1038/nrmicro3330.

Zomer, R.J., Trabucco, A., Bossio, D.A., Verchot, L.V., 2008. Climate change mitigation: a spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agricult. Ecosyst. Environ. Int. Agricult. Res. Clim. Change: Focus Trop. Syst. 126, 67–80. https://doi.org/10.1016/j.agee.2008.01.014.