Deliverable D2.1

Priorities for future isolation-based approaches of microbiota

Work package number and title	*WP2 Novel ecological concepts for microbiome biobanking.*
Work package leader	*HMGU*
Participants	*DSMZ, EMBL, CABI, AIT, MUG, SU*
Relevant Task	*T2.1*
Lead contributor to Deliverable	*HMGU*
Dissemination level	*PU*
Due date (month)	*M24*
Version	*1*

Table of contents

Summary

This document provides quality criteria for the construction of microbial co-occurrence networks, which are the basis for the definition of keystone taxa based on metabarcoding data For this work, we used soil samples, which represent an ecosystem with high microbial biomass and diversity. We applied these criteria to identify keystone taxa from the “soil use case” which is used by many partners of the MICROBE project (mostly in WP1, 2 and 3).

Introduction

The specific objectives of WP2 are: (i) to make use of ecological concepts for the definition of novel isolation and cultivation strategies of microbiota from different ecosystems; (ii) to define strategies for the reconstruction of core microbiomes based on existing and novel isolates to reproduce given functionalities. This deliverable is the result of the T2.1 *Definition of “core microbiomes” and “microbial keystone taxa”*: Identifying microbial keystone taxa might help define priorities in isolating that microbiota from their environment.

Targeted isolation requires a preselection of the important microorganisms from a healthy environment, which is especially relevant when considering the vast microbial diversity and their consequent broadly diverse eco-physiological requirements. The microbial keystone taxa concept might help to prioritize those important microorganisms for targeted isolation. However, there is currently no consensus about the conceptual definition of microbial keystone taxa. Hence, it is important to integrate different perspectives to set up a unified definition to make the research outcomes of this field globally scalable and to face current sustainability threats with microbiome-based solutions. We have worked on the theoretical definition of this concept preparing a publication where we propose an extended 8-point operational definition of microbial keystone taxa. The current definition of microbial keystone taxa refers to specific microorganisms within a microbial community that play critical roles in maintaining the overall structure, function, and stability of the community (Banerjee et al., 2018). We define a microbial keystone taxa as a microbial strain that: (1) has a high impact on the ecosystem independently of its abundance, (2) plays critical functions in the ecosystem and organismal health, (3) can regulate microbial community dynamics and prevents the overgrowth of other species populations, (4) has a very unique ecological niche specialization, (5) contributes to the stability and recovery of microbial communities following environmental disturbances, (6) can modify the ecosystem at different levels, (7) has genome containing key genetic material that can be passed to other members of the microbial community, and (8) can uptake genetic material from other members of the microbial community. These criteria were recently submitted for publication to provide an operational definition for keystone taxa, which is also the basis for the provided technical criteria in this document (Hernandez et al., submitted). Based on these criteria we defined a keystone taxon as a microorganism which determines network morphologies of microbial networks by providing interactions with other microbiota.

To test for the robustness of network analysis, this deliverable had the aim to define quality criteria for metabarcoding sequencing data (here 16S rRNA gene amplicon data to characterize bacterial diversity). The tested parameters included a) sample size, b) microbial diversity, c) sequencing read length, d) sequencing depth, e) ASV inference methods, and f) association methods to identify microbial keystone taxa based on network analyses. As a test data set, we used existing, high-quality metabarcoding data from a soil study "Jülich Priority Experiment" project. We further used the defined pipeline together with the quality criteria provided in this study to define keystone taxa from the "soil use case", which is part of the MiCROBE project, to make our results comparable to WP1 data.

Phase I: Benchmarking a pipeline for the identification of microbial keystone taxa

Methods

Soil samples

We used soil sample data from the "Jülich Priority Experiment". This field experiment is located southeast of Jülich (Germany, altitude 94 m, – 50° 53'51.53'' N, 6°25'21.09'' E); a detailed description of the experiment can be found in Delory et al., 2019, and Weidlich et al., 2017. A subset of 16 grassland plots was selected from 2012 dataset under a high-density treatment (HD). This selection of the plots is important to enable a powerful network analysis, which requires several biological replicates in a relatively homogeneous environment.

Amplicon sequencing

DNA was extracted with DNeasy PowerSoil Pro Kit (Qiagen, Germany) according to the manufacturer's instructions. We targeted the V4 region of the 16S rRNA gene with the 515F/806R primers (Apprill et al., 2015; Parada et al., 2016). The amplicon sequencing process includes two consecutive polymerase chain reactions (PCRs) followed by cleaning and output quantification. The first PCR starts with 1 µl of DNA at 5 ng/µL concentration, 12.5 µl NEB Next High Fidelity Master Mix, 2.5 µl 3% BSA, 8 µl MiliQ H_2O, and 0.5 µl of each primer at 10 pmol/L. The PCR conditions were 28 cycles of denaturation at 98°C for 10 s, annealing at 55°C for 30 s, and extension at 72°C for 30 s, finalizing with 5 min at 72 °C and cooling down to 4°C. Amplicon sizes were verified on a 1% agarose gel, running for 30 min at 120 V, and subsequently cleaned up with MagSi-NGSprep Plus Beads at a 0,8 beads:sample ratio. The function of the second PCR is to add the barcodes to the targeted region, starting with 1 µl of DNA produced in the first PCR at 10 ng/µL concentration, 12.5 µl NEB Next High Fidelity Master Mix, 6.5 µl DEPC H2O, and 2.5 µl of each indexing primer, using Nexters XT Index Kit v2 Set A, B, C, or D. The PCR conditions were 30 s at 98°C, 8 cycles of denaturation at 98°C for 10 s, annealing at 55°C for 30 s, and extension at 72°C for 30 s, finalizing with 5 min at 72 °C and cooling down to 4°C. The PCR products were subsequently cleaned up with MagSi-NGSprep Plus Beads (ratio 0,8 beads: 1 sample), and the quality was checked by DNF-473 Standard Sensitivity NGS Fragment Analysis Kit (1-6000 bp) in the ProSize Data Analysis Software v.5.0. Pooled libraries at 4nM were sequenced in 2 × 250 bp paired-end reads on Illumina MiSeq

Reagent v3 (600 Cycle) (MS-10216 S-3003) instrument. Following sequencing, samples were demultiplexed using the GenerateFASTQ module on the Illumina™ MiSeq instrument.

Initial quality filter

In most cases, the number of reads per sample ranged from 40k to 100k reads (Figure 1a). The length of the reads was on average higher than 200 bp for forward (R1) and reverse (R2) reads, and the standard cut at 20 quality score was sufficient to keep the quality of the data (Figure 1b). The fastq files of the 16S rRNA gene sequencing data were trimmed from adapters, low-quality regions, and "Ns" by trimgalore/0.6.10 (Krueger et al., 2023), and quality was checked by fastQC (Figure 1b). To only retain bases with quality scores ≥ 20, non "Ns", and with a minimum length of 200 bases, we used these parameters: --paired --trim-n --max_n 0 --length 200 --quality 20 –fastqc.

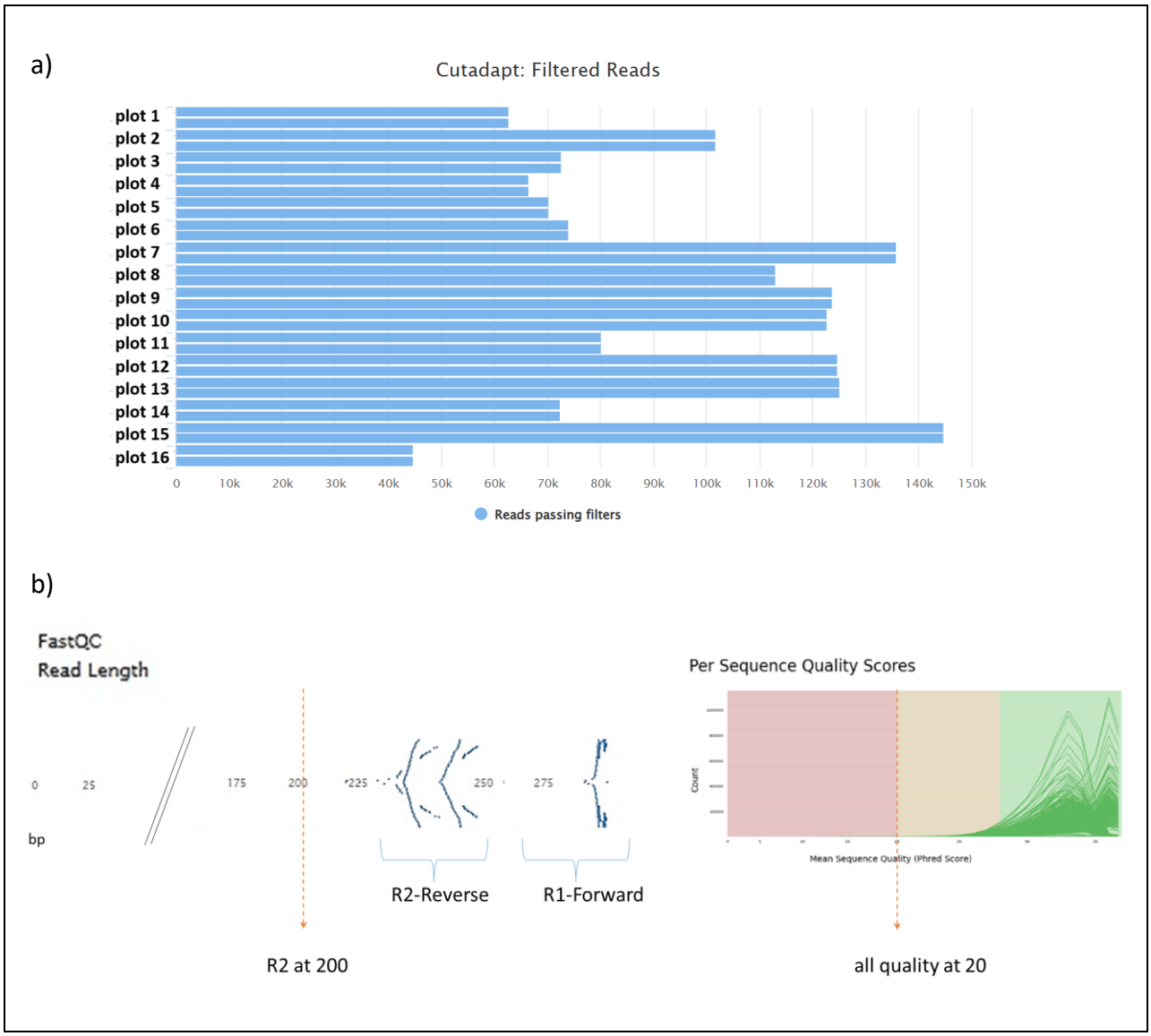

Figure 1. Quality check by FastQC. a) the number of reads passing filtering criteria across different plots. b) The length of the Forward (R1) reads is higher than the Reverse (R2) reads, but all are over 200 bp, and the standard cut at 20 quality score.

Benchmark design

We designed a full factorial benchmark assessment for the identification of microbial keystone taxa based on network analyses (Figure 2). Our dataset was large enough to allow us to subset real data to simulate, for example, lower sample size or microbial diversity. We resampled the data, producing several artificial new data sets, to evaluate the effect of:

a) Sample size

We considered every of the 16 selected plots as a biological replicate, and those were subsequently randomized and subsampled to create sample-size subsets in R.

NetCoMi_Input [sample(nrow(NetCoMi_Input), n),]

where n is the sample size, and n = 3, 5, 8, 12, 16.

b) Microbial diversity

Also using R script, the taxa were randomized and subsampled to create subsets.

i) all taxa: includes all ASVs present in at least 80% of the samples.

ii) low-diversity: defined as low richness by filtering a small random group of ASVs.

NetCoMi_Input [, sample(ncol(NetCoMi_Input), 300)]

iii) high-diversity: defined as high variance and set in NetCoMi:

```
netConstruct(NetCoMi_Input,
        filtTax="highestVar", filtTaxPar=list(highestVar= 800), …)
```

iv) low-abundance: defined as taxa with low read counts.

NetInput_3plots[,which(colSums(NetCoMi_Input)<= 20)]

v) high-abundance: defined as taxa whose number of reads is at least x% of the total number of reads. and set in NetCoMi:

```
netConstruct(NetCoMi_Input,
        filtTax = "relFreq",
        filtTaxPar = list(relFreq = 0.0005), …)
```

c) Sequencing read length

We artificially created data sets with reduced sequencing read lengths by trimming the raw reads randomly, keeping read lengths at 50pb, 100pb, 150pb, 200pb, and 250pb, using *seqtk* (Li, 2012) set as:

seqtk trimfq -L 50 - 100 - 150 - 200 – 250

d) Sequencing depth

We artificially created data sets with reduced depths by resampling the raw reads randomly, keeping the 20%, 40%, 60%, 80%, and 99% of the reads, using seqtk (Li, 2012) set as:

seqtk sample 0.2 - 0.4 - 0.6 - 0.8 - 0.99

e) ASV inference method

The clean reads were then imported in DADA2 v. 1.16 (Callahan et al., 2016) as fastaq files to infer the abundance of amplicon sequence variant (ASV) per sample. We tested the three ASV inference methods in DADA2, not pooling (default), pseudo-pool, and full-pool of the samples. After that, the taxonomic assignment was done according to the Silva Database (http://www.arb-silva.de/).

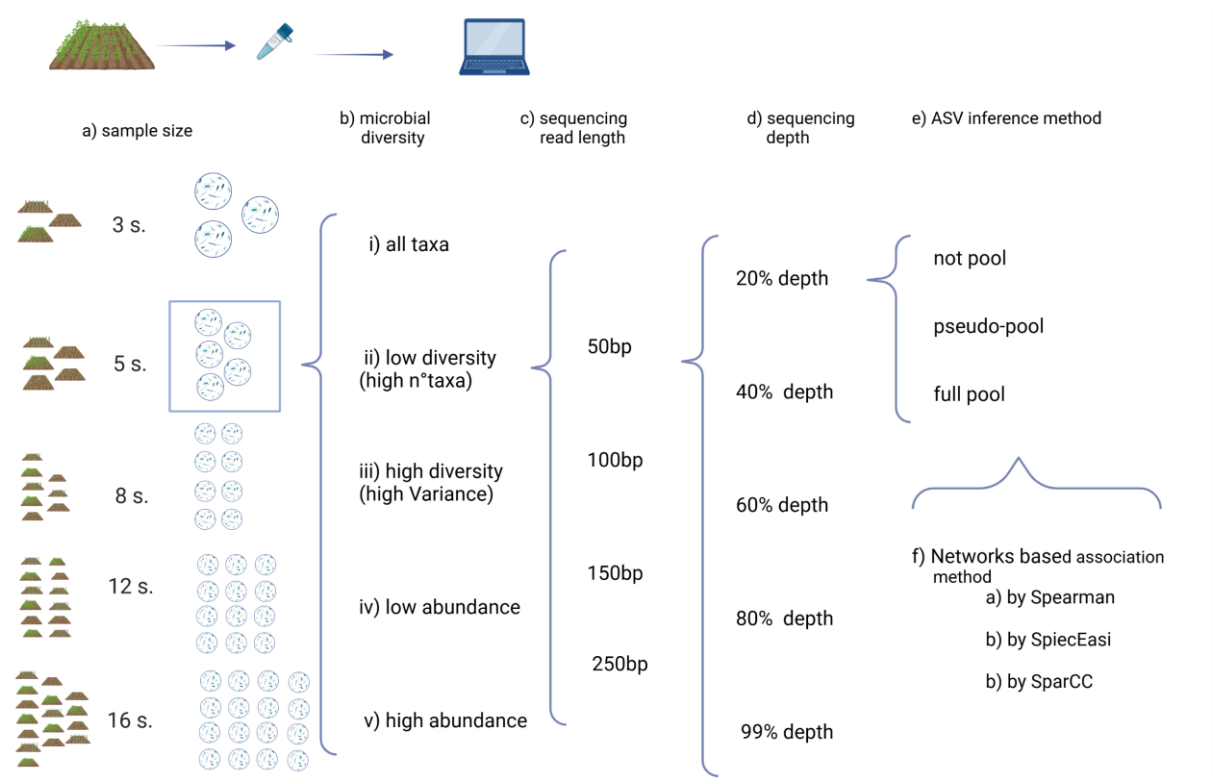

Figure 2. Full factorial analysis of key factors influencing microbial network analyses. (a) sample size, (b) microbial diversity, (c) sequencing read length, (d) sequencing depth, (e) ASV inference method, and (f) network-based association methods. Created in BioRender.com.

f) Association methods to construct microbial networks

Finally, all data sets were tested using different association methods, including Spearman Correlation, SpiecEasi, and SparCC as implemented in NetCoMi (Peschel et al., 2021). The input for NetCoMi, corresponds to the ASVs abundance table after all filters.

netConstruct() setting for the different association methods:

measure="**spearman**", dissFunc= "unsigned", normMethod= "clr",
zeroMethod="pseudo",sparsMethod= "threshold", thresh=0.3

measure = "**spieceasi**", nlambda = 20, and dissFunc = "unsigned".

measure = "**sparcc**" , nlambda = 20, and dissFunc = "unsigned".

We estimated the basic global properties of the networks for each year separately, with netAnalyze() from NetCoMi, using clustMethod = "cluster_fast_greedy," hubPar = "degree," hubQuant = 0.90, lnormFit = TRUE.

Cross-validation of the microbial networks

After we selected the optimal conditions for the network analysis from the previous step, we applied them to a deeper network parameter assessment. The robustness of microbial networks can be assessed by systematically removing individual ASVs and evaluating the impact on network stability. If the removal of a single ASV causes a significant decrease in network stability (e.g., reduced modularity, clustering, or connectivity), it indicates that the ASV plays a critical role in maintaining microbial interactions. This robustness test provides insights into the ecological importance of specific taxa and helps disentangle the contributions of rare versus abundant species to microbial community dynamics. For that, we used cross-validation by the susceptible-infected-recovered (SIR) model with leave-one-out-cross validation (Salavaty et al., 2020). We used *Igraph* (Csardi & Nepusz, 2006) and *influential* (Salavaty et al., 2020) R packages for this post-processing step of the networks. Due to the fact that the *influential* package can only work with unweighted adjacency matrix, we had to transform the weighted adjacency matrix of *NetCoMI* output into an adjacency matrix, replacing the non-zero abundances by presence. Then *igraph* package was used to reconstruct an unweighted graph. We imported the unweighted graph into *influential* to estimate centrality measures, including degree and betweenness centrality, neighborhood connectivity, and local h-index. Based on all these estimations *influential* package can identify the most influential network nodes based on Integrated Value of Influence (IVI) and a range-score hubness. The Integrated Value of Influence (IVI) allows the prioritization of the taxa by ranking the HUBs according to their centrality and influence on the network structure; this is called “HUBness”.

Results

Sample size

When number of replicates was three plots or less, the network construction based on *SpiecEasi* produced, in some cases, empty networks (Figure 3). Further, global properties such as MODULARITY indicate that networks' structure based on three replicates cannot be differentiated from randomness. Hence, no further calculation was possible under that scenario. In cases with more replicates, a network could be established. However, we could show that there is an optimal maximum sample size because there is a tradeoff between increasing the sample size and the number of taxa that remain prevalent (Figure 5). Thus, the number of samples should be higher than five plots and less than 20 plots.

Microbial diversity

Networks based on low-diversity data sets are characterized by a low number of detected hubs, low modularity, higher clustering, and higher edge density. Samples with a) small sample size, b) low microbial diversity, c) shorter sequencing reads, and d) low sequencing depth produce “low-diversity-like” data sets (Figure 4). In those four cases, the microbial abundance table, which is the input for the network analyses, contains a small number of ASVs. As a consequence, the network construction with *SpiecEasi* produces empty networks, with no possibilities for further calculation.

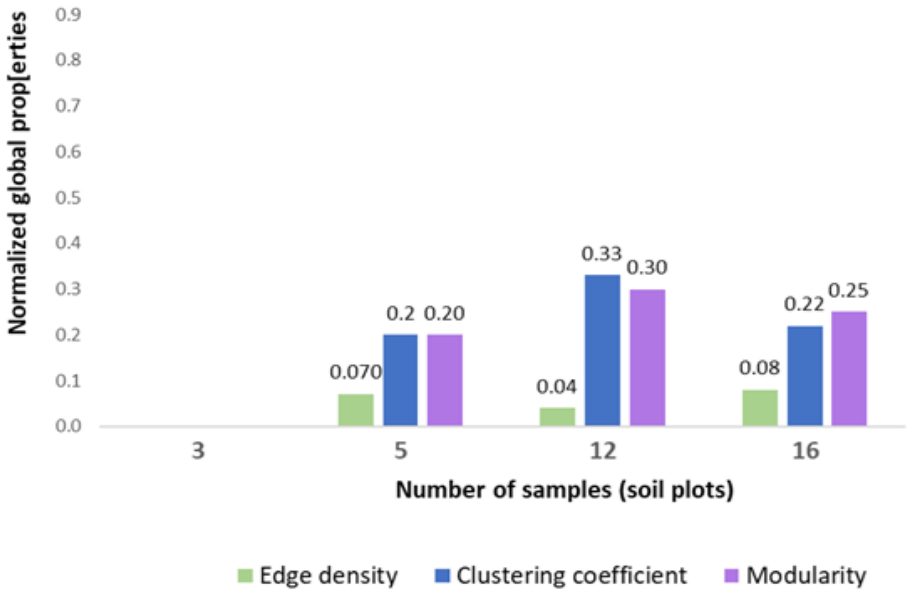

Figure 3. Normalized global network properties by *SpiecEasi* (edge density, clustering coefficient, and modularity) for different numbers of soil samples. Increasing the sample size improves the power of the network construction. Notably, non-calculation is possible at 3 replicates.

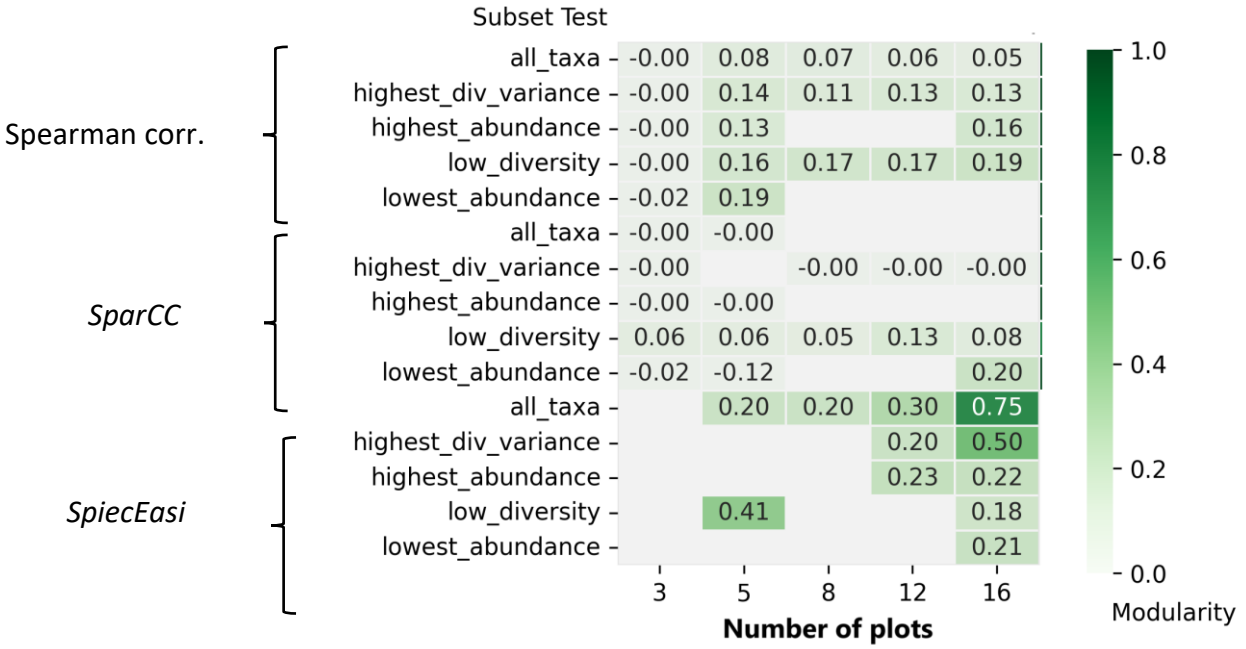

Figure 4. Heatmap of modularity across different sample sizes and microbial diversity, calculated by three association methods. In general, higher modularity is observed in subsets with greater microbial diversity and sample size when calculated by *SpiecEasi*.

Sequencing read length

Table 1 shows results based on the datasets designed to systematically test the effects of sequencing read length on the network's global properties. Here, we show the modularity as an example that illustrates the association between the network construction and the number of taxa on the dataset. Modularity cannot be calculated at 50 bp, and it is 0.089 at 100 bp,

which is a value that likely indicates no modular structure detected or missing data. The second option is more likely because the number of detected and retained ASVs decreases at shorter sequencing reads. Longer sequencing reads allowed for the retention of more than double the taxa. As 250 bp is the longer sequencing read that we could produce by Illumina sequencing, we consider that this length should be the standard for this purpose.

Table 1. Effect of sequencing read length on modularity and the number of taxa removed/remaining in a dataset.

Subset test	Modularity	Taxa_removed	Taxa_remaining
50bp	-	315	157
100bp	0.08898	4284	316
150bp	0.45406	6892	352
200bp	0.43275	6540	338
250bp	0.38796	6575	315

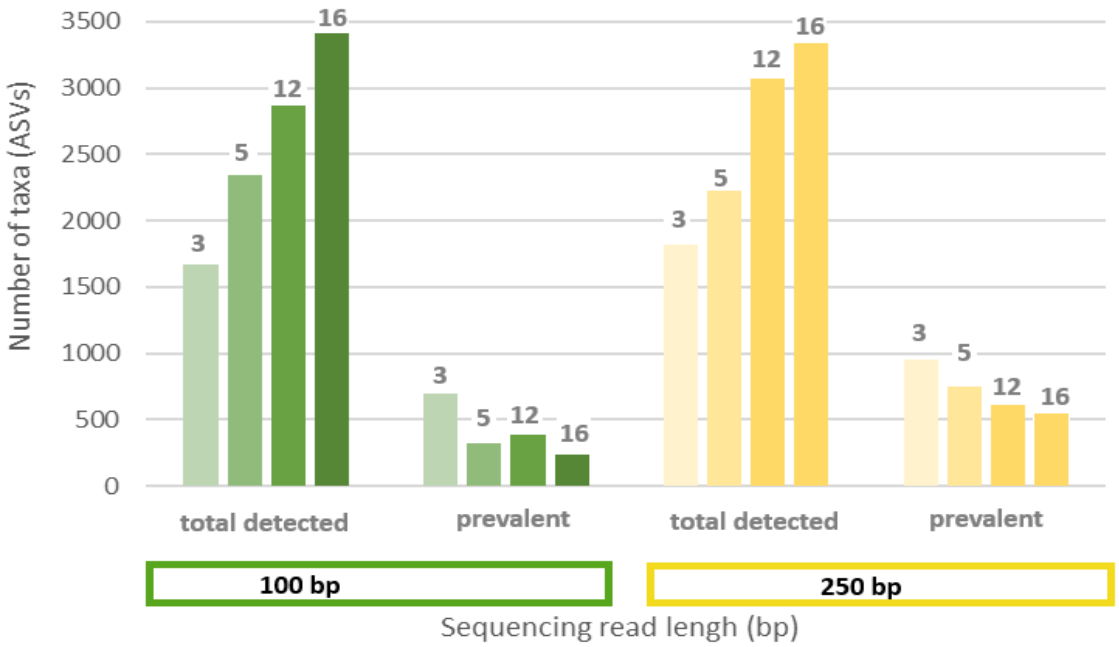

Figure 5. Number of taxa in ASVs in relation to the length of the sequenced reads (bp) and the number of samples (soil plots). Total number of taxa and prevalent number of taxa in at least 80% of the samples.

Sequencing depth

Table 2 shows the effect of the artificial reduction of the sequencing depth on the network's global properties. Here, we show the modularity estimated by SpiecEasi as an example that illustrates that the sequencing depth does not influence modularity in the same way as read length. This suggests that rare taxa are more likely to be excluded at lower sequencing depth. This seems to affect the network construction, especially when based on the SpiecEasi method, probably due to its incorporated robustness test during network construction. Our results suggest that the depth should not be less than the subset at 80% depth, and it should be equivalent to the depth of our original real data or higher. That is, a sequencing depth of at least 50,000 reads per sample (Figure 1a).

Table 2. Effect of sequencing depth on the network construction by SpiecEasi.

Subset test	Modularity	Taxa_removed	Taxa_remaining
20%_depth	-	879	0
40%_depth	-	1663	15
60%_depth	-	2419	48
80%_depth	0.07	3230	110
99%_depth	0.219	3903	284

ASV inference method

We found that the ASVs inference is a crucial downstream bioinformatic step, especially when the aim is to identify taxa that can be at low numbers as the keystone taxa (Figure 6). We found that the abundance of ASV tables produced by DADA2 is insufficient for calculating robust networks when the ASV inference was made at each sample independently. This happened because the low abundance ASVs are not detected when using the default DADA2 parameters (not-pooling). As shown in Figure 6, the global network properties are only consistent when pooling the samples for ASV inference. In general, this happened because *SpiecEasi* has an incorporated robustness test during network construction. In this step, some ASVs are randomly removed, and the associations are recalculated. However, if the sample size is too small or has low diversity, then the removal of some ASVs might cause insufficient data to calculate the association matrix. Thus, the full-pool mode is the recommended setting for ASV inference with DADA2.

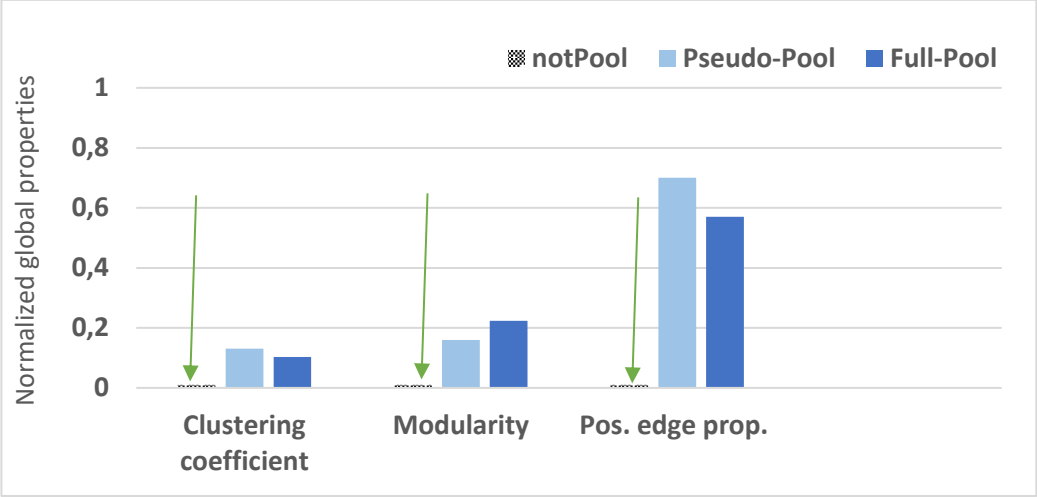

Figure 6. Normalized global network properties (clustering coefficient, modularity, and proportion of positive edges) based on *Spiec-Easi* across ASV inference methods (DADA2: notPool, Pseudo-Pool, Full-Pool). The Full-Pool method can produce works with higher modularity and positive edge proportions. In contrast to pseudo-pooling, which seems to be a suitable alternative to the full-pool, not-pooling does not produce enough data to calculate networks based on *Spiec-Easi*.

Association methods to construct microbial networks

As shown in Table 3 and Figure 7, the networks calculated with the Spearman method had a higher clustering coefficient and longer average path length and diameter, indicating a strong community structure and a more complex network. In Figure 7 can be seen that networks according to (a) *SpiecEasi* and (b) *SparCC* appear visually similar, but the one by (c) Spearman Correlation shows two very clear clusters, which is very likely an artifact due to the suboptimal treatment of the sparsity of the data by the Spearman. The networks calculated with *SpiecEasi* method had a lower clustering coefficient and shorter paths with smaller diameters, suggesting a compact structure and moderate efficiency. The networks calculated with the *SparCC* method were compact with very short paths and minimal clustering, indicating high connectivity but less structured communities. Hence, the networks by *SpiecEasi* have an optimal balance between structure and connectivity

Table 3. Global network properties according three different association methods.

	Spearman	**SpiecEasi**	**SparCC**
Clustering Coefficient	0.4940	0.0866	0.1180
Average Path Length	0.0967	0.0002	0.00001
Diameter	0.1435	0.0097	0.0001

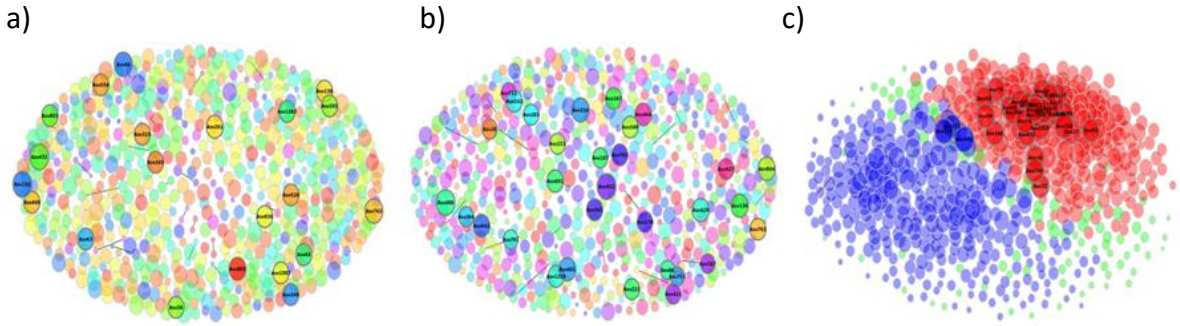

Figure 7. Network visualization based on a) SpiecEasi, b) SparCC, and c) Spearman Correlation, as implemented in NetCoMi (Peschel et al., 2021).

Conclusions

For identifying microbial keystone taxa based on network analyses, the following recommendations are derived from the presented data:

Sample size

A sample size of five (5) to 20 biological replicates is optimal to balance the statistical power to construct a network without compromising the possible prevalence across samples.

Microbial Diversity

The microbial diversity of the target environment influences the required sample size for network construction. High microbial diversity environments require fewer samples, as a single sample can capture much of the diversity.

Sequencing read length

A read length of 250 bp paired-end reads is recommended. Paired-end sequencing ensures overlap between forward and reverse reads, enhancing the accuracy of read assembly and alignment.

Sequencing Depth

A sequencing depth of at least 50,000 reads per sample. The highest possible depth is advised to maximize the detection of low-abundance taxa, which may play critical roles as keystone taxa in microbial networks.

ASV inference method

DADA2's full-pool method for inferring ASVs from sequencing data. Full-pool inference improves resolution, distinguishing true biological sequences from sequencing errors. This is especially relevant for detecting rare taxa that contribute to network associations.

Network association method

SpiecEasi is the most suitable association method for constructing microbial networks because it accounts for sparsity and compositional effects, reducing false positive associations.

These parameters were set for soils, which are considered as ecosystems with a very high microbial diversity and biomass. Parameter settings may differ for other ecosystems where diversity and biomass are lower.

Phase II: Implementation of the pipeline on the soil use case

Methods

Sampling

Soil samples were taken from grassland fields of the National Park Hainich (51°05'48"N, 10°22'27"E), Thuringia, Germany, during the last four Soil Sample Campaigns from 2014 – 2023 in collaboration with the German Biodiversity Exploratories (Fischer, Bossdorf, et al., 2010; Fischer, Kalko, et al., 2010).

The plots have 50 × 50 m each, and were in total 200 plots; 50 EP plots in grassland each year, HEG SSC 2014, 2017, 2023. The so-called "EP plots" are very diverse regarding the soil type and the land-use-intensity index (LUI). The LUI was calculated as a regional mean of grassland management for the Hainich region overall for the years 2014 to 2023, according to Blüthgen et al. 2012, based on information from the land owners on mowing, grazing, and fertilization Vogt et al. 2019 using the LUI calculation tool Ostrowski et al. 2020 implemented in BExIS (http://doi.org/10.17616/R32P9Q). We selected 16 EP plots that have the same soil type and moderated LUI intensity; Cambisol and LUI values from 1 to 2, respectively ((a) sample size: 16 EP plots, based on the typical high (b) microbial diversity of soil).

Amplicon sequencing

DNA was extracted at UFZ Halle, using DNeasy PowerSoil Pro Kit (Qiagen, Germany) according to the manufacturer's instructions. Amplicon sequencing libraries were prepared according to phase I standards, and sequenced on Illumina MiSeq Reagent v3 (600 Cycle) (MS-102-3003) ((c) sequencing read length: 250 bp paired-end reads, (d) sequencing depth: at least 50,000 reads per sample).

Inferring ASVs from sequencing reads

The fastq files of the 16S rRNA gene sequencing were trimmed by trimgalore/0.6.10 (Krueger et al., 2023) and quality was checked by fastQC. We pooled together all samples for ASV inference, following phase I standards ((e) ASV inference method: DADA2 default parameters except for pool=TRUE). Taxonomic assignment according to Silva Database (http://www.arb-silva.de/).

Co-occurrence networks construction

We used the ASVs relative abundance table from the 16 EP plots, including all prevalent taxa *i.e.*, ASVs present in 80% of the samples. Network construction and analysis were performed using the NetCoMi (Peschel et al., 2021) R package, following phase I standards ((f) network-based association method: *SpiecEasi*). Hence, we set NetCoMi parameters to:

-netConstruct() with measure = "spieceasi", nlambda = 20, and dissFunc = "unsigned".

-netAnalyze() for each year separately, using clustMethod = "cluster_fast_greedy", hubPar = "degree", hubQuant = 0.90, lnormFit = TRUE.

Networks post-processing

The post-processing of the networks was done by *Igraph* (Csardi & Nepusz, 2006) to reconstruct an unweighted graph, and *influential* (Salavaty et al., 2020) R packages to identify the most influential network nodes based on Integrated Value of Influence (IVI) and a range-score hubness.

Results

Inferring ASVs from very diverse environmental samples

Based on our benchmarked pipeline from phase I, we sequence at the highest possible quality to obtain a representative sample for the characteristic high diversity of soil. After all filters, each year, around 920 ASVs are prevalent in 80% of the samples (when inferred by DADA2 in the full pool mode). As shown in Figure 8, most of the HUBs were from the *Actinomycetota*, *Thermoproteota*, and *Verrucomicrobiota* phyla.

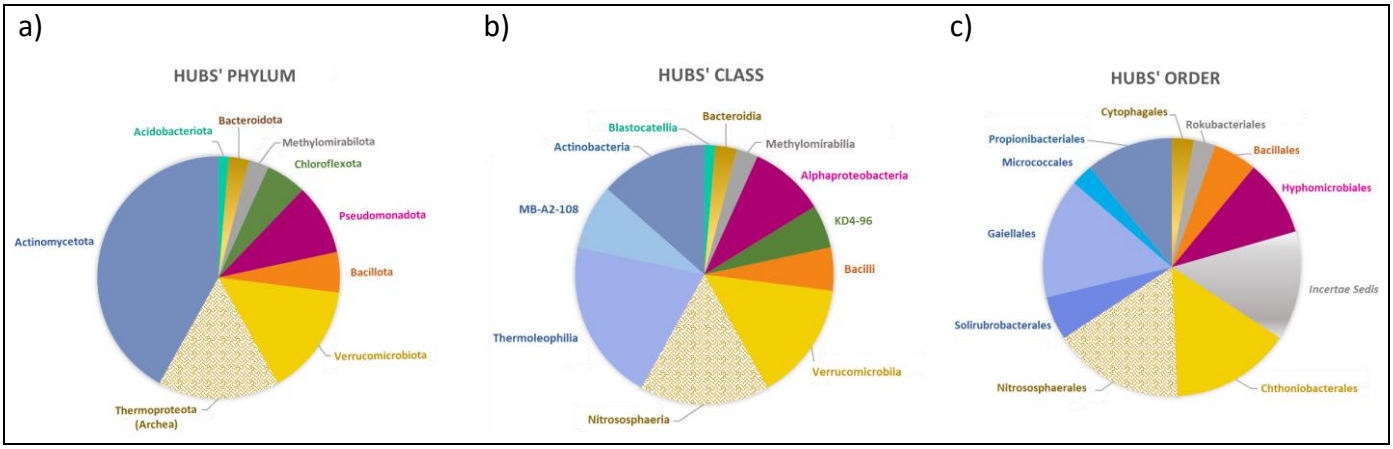

Figure 8. Microbial community in grassland soil. It shows the keystone taxa at different taxonomic composition levels: (a) phylum, (b) class, and (c) order.

Cross-validation of the microbial networks: identifying candidate keystone taxa from the highly influential HUBs of the networks

The result from the network analysis is a list of candidate keystone taxa based on co-occurrence association models (Annex table, corresponding to Milestone Ms8). Visualization based on Integrated Value of Influence (IVI) from the *influential* (Salavaty et al., 2020) clearly

shows the ranking of the hubs, which allow the priorization (Figure 9). We found 92 HUBs in 2014, 112 HUBs in 2017, 115 HUBs in 2021, 90 HUBs in 2023, and 74 HUBs that overlap across all years (Annex: List of candidate microbial keystone taxa). These 4-year HUBs are already candidates. We further selected among them 15 HUBs with high priority, based on four criteria: (i) they are part of the core microbiome, (ii) they have higher HUBnes scores (see "Mean HUBness in 4 years" in Annex table) with (iii) lower abundance Standard Deviation of their taxonomic clade (see "SD HUBness In 4 years" in Annex table), (iv) one or two HUBs were selected per taxonomic clade (Figure 10, and Annex table, corresponding to milestone eight). We chose two HUBs instead of one in the clades where some of the HUBs were identified as the metagenomes of the same samples.

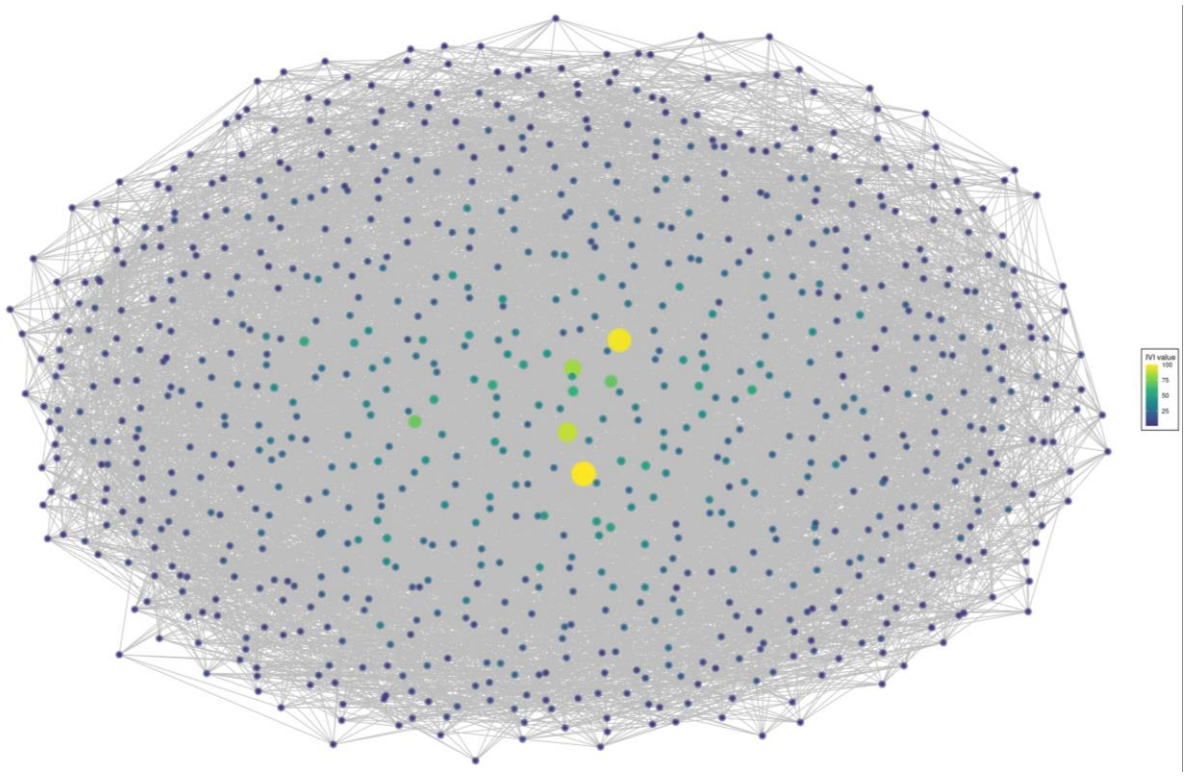

Figure 9. Microbial network visualization. The yellow circles represent the most influential network nodes based on the Integrated Value of Influence (IVI) from the *influential* (Salavaty et al., 2020) R package.

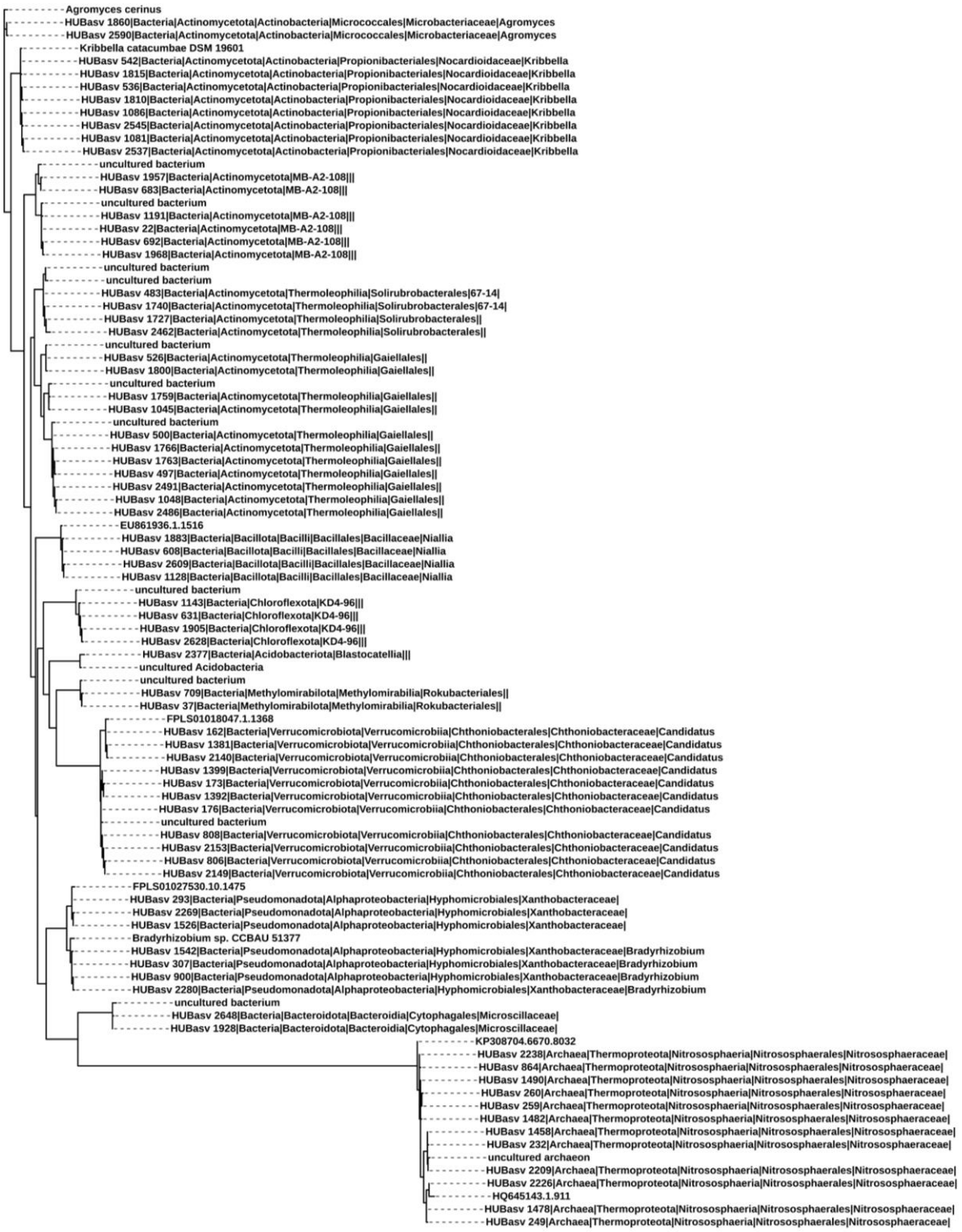

Figure 10. Phylogenetic tree of ASVs based on 16S rRNA sequences, including one reference neighbor per clade. Showing taxonomic relationships across bacterial and archaeal taxa. ASVs are labeled with their corresponding taxonomic

Next steps

In the next steps, the identified keystone taxa will be further characterized by mapping the metabarcoding sequence data to the metagenome-assembled genomes, and the isolates obtained in T2.3. We will use metagenome sequence analysis to functionally describe the microbial community and to predict the targeted isolation condition of the candidate keystone taxa based on their genomes' functional annotation. We will use the isolates of the identified keystone taxa to develop and establish Synthetic Communities. Complementing and mapping the results from T2.1, the definition of microbial keystone taxa, the identification of priorities for isolation, and the metagenome functional analyses is the connection between this deliverable, T2.2 *Inferring functional characteristics from metagenomic data,* and with the WP1 *Technical solutions for biobanking of microbiome samples*.

Currently, we have sequenced two representative samples by short reads Illumina NovaSeq. The raw reads were trimmed from adapters, low-quality regions, and N-rish regions by trimgalore/0.6.10 (Krueger et al., 2023) and quality was checked by fastQC. To only retain bases with quality scores ≥ 20, maximum 10 "Ns", and with a minimum length of 100 bases, we used these parameters: --paired --trim-n --max_n 10 --length 100 --quality 20 –fastqc

The assembly and functional annotation of the metagenomes were done using the SqueezeMeta pipeline. Based on these preliminary metagenome results, we have found four ASVs from the "List of candidate microbial keystone taxa", and we are working to fine-tune this pipeline.

Availability of datasets and tools

The two datasets used for validation and analysis will be available in the SRA. The code will be available on GitHub. The exact accession numbers and GitHub repository address will be updated shortly.

Annex: List of candidate microbial keystone taxa

List of priorities for isolation, corresponding to Milestone Ms8 “Synthetic consortium candidate species identified”. Organisms marked in red represent keystone taxa with highest priority of isolation

HUBasv_id	Mean HUBness in 4 years	SD HUBness In 4 years	inCoreMic	Priority per Clade	Domain	Phylum	Class	Order	Family	Genus
HUBasv_2209	53.90	6.31	inCoreMic	1	Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_2377	34.90	14.05	inCoreMic	1	Bacteria	Acidobacteriota	Blastocatellia		*Incertae Sedis*	*Incertae Sedis*
HUBasv_2590	45.40	11.39	inCoreMic	1	Bacteria	Actinomycetota	Actinobacteria	Micrococcales	Microbacteriaceae	Agromyces
HUBasv_2545	20.40	6.69	inCoreMic	1	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_683	38.70	10.24	inCoreMic	1	Bacteria	Actinomycetota	MB-A2-108	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1759	50.20	8.70	inCoreMic	1	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales		
HUBasv_2462	60.40	21.29	inCoreMic	1	Bacteria	Actinomycetota	Thermoleophilia	Solirubrobacterales		
HUBasv_608	28.40	13.12	inCoreMic	1	Bacteria	Bacillota	Bacilli	Bacillales	Bacillaceae	Niallia
HUBasv_1128	43.30	16.41	inCoreMic	1	Bacteria	Bacillota	Bacilli	Bacillales	Bacillaceae	Niallia
HUBasv_1928	41.00	21.07	inCoreMic	1	Bacteria	Bacteroidota	Bacteroidia	Cytophagales	Microscillaceae	
HUBasv_1143	49.80	16.79	inCoreMic	1	Bacteria	Chloroflexota	KD4-96	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_709	36.90	11.03	inCoreMic	1	Bacteria	Methylomirabilota	Methylomirabilia	Rokubacteriales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1542	41.90	2.77	inCoreMic	1	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	Bradyrhizobium
HUBasv_293	19.30	5.69	inCoreMic	1	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	*Incertae Sedis*
HUBasv_162	21.50	18.74	inCoreMic	1	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_1458	41.10	16.33	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_249	34.60	5.73	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	
HUBasv_232	33.40	2.63	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_2226	26.80	17.09	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	
HUBasv_1482	22.20	12.36	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_2238	19.80	11.58	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_864	17.80	9.90	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_1490	17.50	8.17	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_1478	16.70	5.28	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	
HUBasv_260	16.10	7.45	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*
HUBasv_259	9.90	7.87	inCoreMic		Archaea	Thermoproteota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae	*Incertae Sedis*

HUBasv_1860	36.10	23.05	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Micrococcales	Microbacteriaceae	Agromyces
HUBasv_2537	26.70	12.60	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_1086	24.80	11.04	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_1081	24.60	10.27	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_542	22.00	15.23	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_1810	19.90	9.05	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_1815	18.50	10.83	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_536	18.20	8.70	inCoreMic	Bacteria	Actinomycetota	Actinobacteria	Propionibacteriales	Nocardioidaceae	Kribbella
HUBasv_1957	40.30	19.24	inCoreMic	Bacteria	Actinomycetota	MB-A2-108	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1191	32.10	19.59	inCoreMic	Bacteria	Actinomycetota	MB-A2-108	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_22	24.20	14.41	inCoreMic	Bacteria	Actinomycetota	MB-A2-108	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_692	20.20	10.26	inCoreMic	Bacteria	Actinomycetota	MB-A2-108	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1968	17.50	5.23	inCoreMic	Bacteria	Actinomycetota	MB-A2-108	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1048	49.40	15.78	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_500	44.30	20.48	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_2491	42.40	12.39	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_497	40.90	13.80	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1045	40.40	4.48	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales		
HUBasv_1763	40.30	17.00	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1766	37.00	24.00	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_2486	36.60	14.59	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_526	28.50	9.86	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1800	25.60	9.19	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Gaiellales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_1740	44.60	30.99	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Solirubrobacterales	67-14	*Incertae Sedis*
HUBasv_483	34.90	14.58	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Solirubrobacterales	67-14	*Incertae Sedis*
HUBasv_1727	31.90	8.31	inCoreMic	Bacteria	Actinomycetota	Thermoleophilia	Solirubrobacterales		
HUBasv_1883	34.70	22.93	inCoreMic	Bacteria	Bacillota	Bacilli	Bacillales	Bacillaceae	Niallia
HUBasv_2609	21.80	18.51	inCoreMic	Bacteria	Bacillota	Bacilli	Bacillales	Bacillaceae	Niallia
HUBasv_2648	41.90	14.29	inCoreMic	Bacteria	Bacteroidota	Bacteroidia	Cytophagales	Microscillaceae	
HUBasv_1905	36.70	9.29	inCoreMic	Bacteria	Chloroflexota	KD4-96	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_2628	29.70	14.27	inCoreMic	Bacteria	Chloroflexota	KD4-96	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*
HUBasv_631	39.30	13.47	inCoreMic	Bacteria	Chloroflexota	KD4-96	*Incertae Sedis*	*Incertae Sedis*	*Incertae Sedis*

HUBasv_37	30.20	9.19	inCoreMic	Bacteria	Methylomirabilota	Methylomirabilia	Rokubacteriales	*Incertae Sedis*	*Incertae Sedis*
HUBasv_2280	21.90	10.59	inCoreMic	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	Bradyrhizobium
HUBasv_307	42.80	37.96	inCoreMic	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	Bradyrhizobium
HUBasv_900	35.30	11.26	inCoreMic	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	Bradyrhizobium
HUBasv_1526	32.30	19.12	inCoreMic	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	*Incertae Sedis*
HUBasv_2269	20.90	6.45	inCoreMic	Bacteria	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	*Incertae Sedis*
HUBasv_808	36.60	14.91	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_173	28.00	16.55	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_176	27.90	12.91	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_1392	26.10	11.08	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_2153	25.90	6.47	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_2140	23.40	10.01	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_1399	22.10	5.36	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_2149	21.50	9.01	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_806	17.90	5.01	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter
HUBasv_1381	17.60	4.30	inCoreMic	Bacteria	Verrucomicrobiota	Verrucomicrobiia	Chthoniobacterales	Chthoniobacteraceae	Candidatus Udaeobacter

References

Apprill, A., Mcnally, S., Parsons, R., & Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, *75*(2), 129–137. https://doi.org/10.3354/ame01753

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*, *16*(9), 567–576. https://doi.org/10.1038/s41579-018-0024-1

Blüthgen, N., Dormann, C. F., Prati, D., Klaus, V. H., Kleinebecker, T., Hölzel, N., Alt, F., Boch, S., Gockel, S., Hemp, A., Müller, J., Nieschulze, J., Renner, S. C., Schöning, I., Schumacher, U., Socher, S. A., Wells, K., Birkhofer, K., Buscot, F., … Weisser, W. W. (2012). A quantitative index of land-use intensity in grasslands: Integrating mowing, grazing and fertilization. *Basic and Applied Ecology*, *13*(3), 207–220. https://doi.org/10.1016/j.baae.2012.04.001

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. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Delory, B. M., Weidlich, E. W. A., von Gillhaussen, P., & Temperton, V. M. (2019). When history matters: The overlooked role of priority effects in grassland overyielding. *Functional Ecology*, *33*(12), 2369–2380. https://doi.org/10.1111/1365-2435.13455

Fischer, M., Bossdorf, O., Gockel, S., Hänsel, F., Hemp, A., Hessenmöller, D., Korte, G., Nieschulze, J., Pfeiffer, S., Prati, D., Renner, S., Schöning, I., Schumacher, U., Wells, K., Buscot, F., Kalko, E. K. V., Linsenmair, K. E., Schulze, E. D., & Weisser, W. W. (2010). Implementing large-scale and long-term functional biodiversity research: The Biodiversity Exploratories. *Basic and Applied Ecology*, *11*(6), 473–485. https://doi.org/10.1016/j.baae.2010.07.009

Fischer, M., Kalko, E. K. V., Linsenmair, K. E., Pfeiffer, S., Prati, D., Schulze, E. D., & Weisser, W. W. (2010). Exploratories for large-scale and long-term functional biodiversity research. In *Long-Term Ecological Research: Between Theory and Application* (pp. 429–443). Springer Netherlands. https://doi.org/10.1007/978-90-481-8782-9_29

Krueger, F., James, F., Ewels, P., Afyounian, E., Weinstein, M., & Schuster-Boeckler, B. (2023). *TrimGalore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ file* (0.6.10). Babraham Institute. https://github.com/FelixKrueger/TrimGalore

Li, H. (2012). *Seqtk: a fast and lightweight tool for processing sequences in the FASTA or FASTQ format*. https://github.com/lh3/seqtk

Ostrowski, A., Lorenzen, K., Petzold, E., & Schindler, S. (2020). *Land use intensity index (LUI) calculation tool of the Biodiversity Exploratories project for grassland survey data from three different regions in Germany since 2006, BEXIS 2 module* (2.0.0). https://zenodo.org/records/3865579

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403–1414. https://doi.org/10.1111/1462-2920.13023

Peschel, S., Müller, C. L., Von Mutius, E., Boulesteix, A. L., & Depner, M. (2021). NetCoMi: Network construction and comparison for microbiome data in R. *Briefings in Bioinformatics*, *22*(4), 1–18. https://doi.org/10.1093/bib/bbaa290

Salavaty, A., Ramialison, M., & Currie, P. D. (2020). Integrated Value of Influence: An Integrative Method for the Identification of the Most Influential Nodes within Networks. *Patterns*, *1*(5). https://doi.org/10.1016/j.patter.2020.100052

Vogt, J., Klaus, V. H., Both, S., Fürstenau, C., Gockel, S., Gossner, M. M., Heinze, J., Hemp, A., Hölzel, N., Jung, K., Kleinebecke, T., Lauterbach, R., Lorenzen, K., Ostrowski, A., Otto, N., Prati, D., Renner, S., Schumacher, U., Seibold, S., … Weisser, W. W. (2019). Eleven years' data of grassland management in Germany. *Biodiversity Data Journal*, *7*, 1–38. https://doi.org/10.3897/BDJ.7.E36387

Weidlich, E. W. A., von Gillhaussen, P., Delory, B. M., Blossfeld, S., Poorter, H., & Temperton, V. M. (2017). The importance of being first: Exploring priority and diversity effects in a grassland field experiment. *Frontiers in Plant Science*, *7*(January), 1–12. https://doi.org/10.3389/fpls.2016.02008