communications biology

A Nature Portfolio journal

https://doi.org/10.1038/s42003-025-07912-4

Definition of the microbial rare biosphere through unsupervised machine learning

Check for updates

Francisco Pascoal[1,2], Paula Branco[3], Luís Torgo[4], Rodrigo Costa[5,6] & Catarina Magalhães[1,2]

The microbial rare biosphere, composed of low-abundance microorganisms in a community, lacks a standardized delineation method for its definition. Currently, most studies rely on arbitrary thresholds to define the microbial rare biosphere (e.g., 0.1% relative abundance per sample), hampering comparisons across studies. To address this challenge, we present *ulrb* (Unsupervised Learning based Definition of the Rare Biosphere), available as an R package. *ulrb* uses unsupervised machine learning to optimally classify taxa into abundance categories (e.g., rare, intermediate, or abundant) within microbial communities. We show that *ulrb* is more consistent than threshold-based approaches and can be applied to data derived from common microbial ecology protocols and non-microbial studies. *ulrb* can be used to identify different types of rarity and is statistically valid for the analysis of various dataset sizes. In conclusion, *ulrb* discerns rare from abundant organisms in a user-independent manner, finding applicability in selected ecological datasets.

Most species in nature are rare[1–5], a trend recognized as early as in the XIX century, by Darwin, in *The Origin of Species*: "rarity is the attribute of a vast number of species"[6]. Generally, the identification of rare species is important for biodiversity conservation, because rare species are often closer to extinction[7]. Within the microbiology field, the rare biosphere[8] is considered a reservoir of genetic diversity[2,3], which is of crucial relevance for the resistance and resilience of ecosystems[4], a source of symbionts shaping host-associated microbiomes[9], and a source of novel biosynthetic genes[10].

The standard computational measure to study the rare biosphere is to order all taxa from the most to the least abundant, in a Rank Abundance Curve (RAC). The RAC can be mathematically described by the power-law[11], whereby a few taxa are abundant, but many are rare in the so-called long tail of the RAC. Most studies define the microbial rare biosphere using relative abundance thresholds such as 0.1% or 0.01% per sample (e.g.,[12–21]), based on early microbial ecology studies of the RAC[2,8,22]. However, threshold-based approaches do not accommodate for differences in sequencing depth obtained by different methodologies. Moreover, different thresholds provide different interpretations of the RAC and most likely none provides consistent results across different methods or communities. Using a specific example, the results obtained by a 0.1% relative abundance, per sample, will be different between using amplicon sequencing of a small region of the 16S rRNA gene or using shotgun metagenome sequencing. This is because the methods produce abundance tables with taxon abundance scores in different orders of magnitude. Thus, a definition of 0.1% relative abundance, per sample, might work well to describe the long RAC tail of a 16S rRNA sequencing dataset. However, this same threshold would yield a very different view of the rare biosphere from the shotgun metagenome sequencing data from the same sample[20]. This is a problem, because it complicates inter-comparability across studies and sequencing methodologies (Supplementary Fig. 1). In summary, threshold-based approaches are flawed, because they are arbitrary.

Previous studies have proposed alternative ways of defining the rare biosphere, for example, by calculating the impact of different thresholds on beta diversity (Multilevel Cutoff Level Analysis, MultiCoLA)[23,24]. However, in a previous study we showed that MultiCoLA did not resolve the arbitrary nature of threshold-based approaches to define the rare biosphere[20]. Other studies have suggested evaluating several thresholds against the RAC[25,26] and recalibrate according to sequencing depth, using the ratio between observed and expected taxa (with Chao index)[25]. Outside the scope of microbial ecology, the utilization of unsupervised learning to define rare and common species has been proposed with the FuzzyQ method[27].

Here, we propose an unsupervised machine learning approach to solve the major issues of the threshold-based methods to define the microbial rare biosphere. We refer to our approach and respective methodology (using default parameters, unless stated otherwise) as Unsupervised Learning based Definition of the Rare Biosphere (*ulrb*).

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal. [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal. [3]School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Ottawa, ON, Canada. [4]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada. [5]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. [6]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. ✉e-mail: rodrigoscosta@tecnico.ulisboa.pt; catarina.magalhaes@fc.up.pt

ulrb clusters all taxa sampled from a biological community using the k-medoids model with the partitioning around medoids algorithm (pam)[28]. The k-medoids model is an unsupervised learning model that partitions points of data into k clusters, minimizing the distance between the points and the centroid of the clusters[28]. Within *ulrb*, the points are the taxa abundance scores in a given sample, and the clusters represent their abundance classifications. The *ulrb* method allows for different numbers of classifications, which can be adapted to the experimental design of the user. As a default parameter, *ulrb* uses three clusters ($k = 3$), corresponding to the classifications "rare", "undetermined" and "abundant". The "undetermined" classification can be interpreted as "intermediate", that is, a state of abundance between "rare" and "abundant". There are metrics that can be used to inspect what is the best number of classifications[29–31] and there is an option to automatically decide the number of classifications in *ulrb* (see Methods).

The introduction of an intermediate classification is optional but recommended to avoid the existence of taxa with very similar abundance scores having opposite classifications ("rare" or "abundant"). Previous studies, using relative abundance thresholds, have also introduced intermediate classifications to provide more comprehensive information[32,33]. The ecological implication of considering intermediate classifications is the acknowledgment that some taxa are neither rare nor abundant, for example, they might be transitioning between being rare and abundant, as conditionally rare taxa[34,35]. The most important aspect of *ulrb* is that it automatically classifies taxa based solely on their abundance score within a community. Furthermore, the method considers that a taxon is not rare/abundant by itself. Instead, a taxon is rare relative to another that is abundant, or vice-versa.

The objective of this study is to present an unsupervised machine learning approach for the definition of the rare biosphere and validate its applicability to a wide range of datasets. Our method can be used for the analysis of any biological community with the R package *ulrb* (Unsupervised Learning based Definition of the Rare Biosphere), which uses open-source code and is available in The Comprehensive R Archive Network (CRAN, https://cloud.r-project.org/web/packages/ulrb/index.html) and GitHub (https://github.com/pascoalf/ulrb) repositories. Additionally, the R package *ulrb* includes a dedicated website with several tutorials and extensive documentation on all functions (https://pascoalf.github.io/ulrb/). *ulrb* was tested against microbial communities obtained from different sequencing and bioinformatics strategies and compared against threshold-based methods for the description of the rare biosphere. Its statistical validity was evaluated against variations in the number of phylogenetic units, samples and sequencing depth. Further, the applicability of *ulrb* for non-microbial (animal and plant) datasets was tested, while also applying the FuzzyQ method to the microbial datasets analyzed in this study. Finally, an *ulrb* extension to identify types of rarity in a host-microbiome context was illustrated.

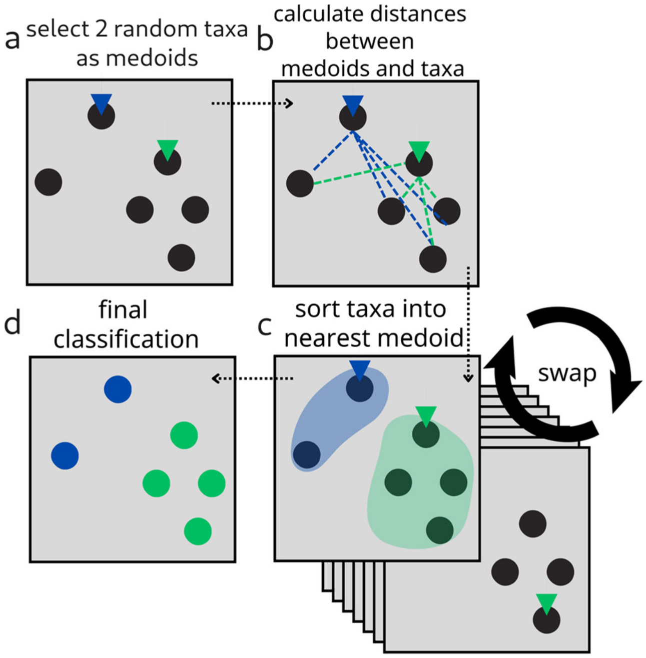
a select 2 random taxa as medoids
b calculate distances between medoids and taxa
d final classification
c sort taxa into nearest medoid
swap

Fig. 1 | Schematic representation of k-medoids. This is a simplified illustration of k-medoids, using 2 dimensions (x and y axis), for two clusters ($k = 2$), represented by different colors. **a** Two random taxa are selected as medoids. **b** The distance between each medoid and all other taxa is calculated. **c** Taxa are sorted in the cluster of the nearest medoid. **d** After a swapping multiple times, a final classification is achieved.

Methods

The *ulrb* algorithm

The unsupervised learning method used by *ulrb* is partitioning around medoids (pam) algorithm[28], based on k-medoids model[36]. In the context of *ulrb*, we apply the pam algorithm for a single feature, which is the abundance scores of taxa in a given sample. Thus, the result obtained in one sample is independent from the result obtained in another sample. The principle of the pam algorithm[28], in *ulrb*, is to divide all taxa into a predefined number of clusters (k), so that taxa within the same cluster are more similar to each other than what they are compared to taxa of other clusters. This is achieved by finding the centroids of clusters (medoids) and maximizing the objective function, which in this case minimizes the distance between taxa and their respective medoid. To do this, the algorithm randomly selects two candidate taxa as medoids, then it calculates the distance between them and all other taxa, attributing all taxa to the nearest medoid (Fig. 1). Then, the algorithm enters into the swap phase, whereby the medoids are replaced and distances are calculated again (Fig. 1). The swap phase is repeated until the total distances between taxa are minimized, and clusters are defined (Fig. 1). For more details on the algorithm, we refer the reader to the *ulrb* package documentation (https://pascoalf.github.io/ulrb/), as well as to Kaufman and Rousseeuw[28] and cluster package documentation[37]. Because *ulrb* explores one dimension of the abundance table (phylotype abundances in a sample), any data transformation will not change the relative distance between points for abundance classification, and thus the method works equally well for compositional and non-compositional data.

ulrb R package construction and utilization

The *ulrb* R package was built using the functionalities of devtools[38]. It includes functions to prepare abundance tables and apply the pam algorithm, and helper functions to verify statistics and for data visualization.

The main function in the *ulrb* package is called define_rb(), which will apply the *ulrb* method and automatically provide a classification of all taxa into "rare", "undetermined" or "abundant". The define_rb() function uses an abundance table as input. This table should include, at least, three columns, indicating the abundance, sample name and phylogenetic unit. Additional variables are allowed and unchanged by the function define_rb(). To apply the pam algorithm[28,39] we used the pam() function, from the cluster package[37]. Besides the default parameters, it is possible to choose a specific number of abundance classifications, but in this case the user needs to manually name them. For example, if the user decides to use $k = 4$, then the abundance classifications will be named "1", "2" and so on, but it is trivial to change those automatic names into user specified terms, *e.g.*, "very rare", "rare", "abundant", and so on.

It is possible to automatically decide k in define_rb() function. For that purpose, we made an additional function, suggest_k(), which will calculate the best k possible, based on either the average Silhouette score[31], Davies-Bouldin index[30] or Calinski-Harabasz index[29] (more details below). To calculate the average Silhouette score we used the pam() function from the cluster R package[37] and to calculate the Davies-Bouldin and Calinski-Harabasz indices we used the clusterSim R package[40]. By default, suggest_k() will use the average Silhouette score. Independently of using default or user specified parameters, the define_rb() function will throw a warning for

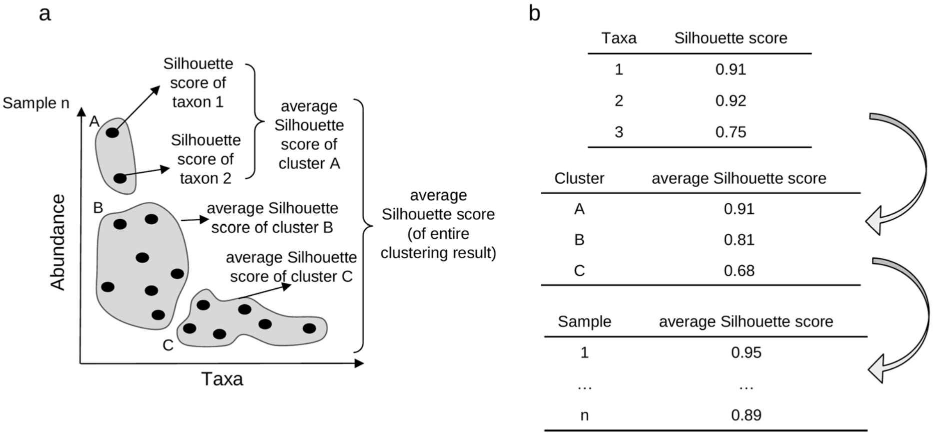
a
Sample n
Abundance
Taxa
A
B
C
Silhouette score of taxon 1
Silhouette score of taxon 2
average Silhouette score of cluster A
average Silhouette score of cluster B
average Silhouette score of cluster C
average Silhouette score (of entire clustering result)
b
Taxa | Silhouette score
1 | 0.91
2 | 0.92
3 | 0.75
Cluster | average Silhouette score
A | 0.91
B | 0.81
C | 0.68
Sample | average Silhouette score
1 | 0.95
... | ...
n | 0.89

Fig. 2 | Schematic representation of the information that the Silhouette scores can provide. **a** Visual representation of taxa grouped in three different clusters, each taxon with a Silhouette score. **b** The average Silhouette scores of all taxa within a cluster can be calculated to deliver a single quality score regarding that specific cluster, or the Silhouette scores of the taxa from all clusters can also be calculated, to obtain a single average Silhouette score for the entire sample. The values shown are arbitrary, for the purpose of illustrating the possible ways of grouping the data.

samples with low Silhouette scores. To do that, define_rb() identifies clusters, across all samples, where at least half the taxa correspond to a Silhouette score below 0.5. Even if this warning appears, the user can proceed normally, being aware that it might be possible to improve the clustering performance. However, the fact that a specific cluster got a bad average score does not imply that the structure of the entire clustering result is artificial. An artificial cluster is a cluster produced through a human method without prior assumptions on the data and that may have an unknown or currently unobservable meaning when looking at the properties of the data. We warn the user, however, that if different studies use different numbers of classifications, to accommodate the best Silhouette scores, then comparability is hindered.

The function suggest_k() provides the best k value for all samples used as input, by default. However, suggest_k() can alternatively return a detailed result, which provides a list with a report on the behavior of the three different indices (average Silhouette score, Davies-Bouldin and Calinski-Harabasz indices) across different values of k. The values of k that are tested by default range from 3 to 10. This range of k values can be changed, but more than 10 clusters might erode the purpose of using unsupervised learning methods to define the rare biosphere (and other domain-related abundance classifications, like "abundant"), because the more clusters there are, the less information they provide. The user can use any range of allowed values of k, from two up to the total number of different abundance scores in a given sample. Note that if more than one sample is tested at the same time, then the maximum k will be the lowest maximum k across all samples tested. A tutorial is available on the *ulrb* R package website illustrating the impact of extreme k values on abundance classifications (https://pascoalf.github.io/ulrb/articles/explore-classifications.html).

To help the users format their dataset for *ulrb* package functions, we provide the prepare_tidy_data() function, which can transform common abundance table formats into the required format. Specifically, taxa by rows, with samples as columns; or vice versa.

Additional functions used within the major functions described in here were illustrated in the package tutorials, available online (https://pascoalf.github.io/ulrb/index.html).

Unsupervised learning statistics

The package *ulrb* includes three main statistics to evaluate the quality of the clustering, which are the average Silhouette score[31], Davies-Bouldin index[30] or Calinski-Harabasz index[29]. To evaluate the quality of the clusters obtained from *ulrb* results in this study, we relied on the Silhouette score[31]. However, depending on the user's needs, one of the other statistics might be more useful. Briefly, the average Silhouette score measures cluster definition and separation, the Calinski-Harabasz index measures cluster separation and density, and Davies-Bouldin measures cluster separation. Below, we describe the Silhouette score in more detail, because it was the index used to evaluate the results presented here. For more details on Calinski-Harabasz and Davies-Bouldin indices, see Supplementary Methods.

Silhouette score

The Silhouette score calculates how close a taxon is to its own cluster relative to the next closest cluster. The Silhouette score of a given taxon, $S(i)$, is given by Eq. 1,

$$S(i) = \frac{(b - a)}{\max(a, b)} \tag{1}$$

where a is the mean distance between the ith taxon and all other taxa on the same cluster, and b is the mean distance between all taxa in the cluster of the ith taxa and the centroid of the next closest cluster. It follows that $-1 \leq S(i) \leq +1$. By convention, $S(i) = 0$ means that the ith taxon is as close to its own cluster as it is to the next closest cluster; $S(i) = -1$ means that the ith taxon is better positioned in the next closest cluster, instead of its own cluster; and $S(i) = +1$ means that the ith taxon is in the center of its own cluster[31]. Note that a perfect score might indicate an artificial cluster in the case of an outlier group[31], but we address this issue in the Discussion section and accept clusters of outliers as valid.

Based on Kaufman and Rousseeuw[39], we interpreted the average Silhouette score as: >0.71 strong cluster; >0.51 reasonable cluster; ≥0.26 weak cluster; and values below 0.26 indicate a potentially artificial cluster.

The Silhouette score is calculated for each taxon, but it can provide information on a specific cluster or all clusters (Fig. 2). Thus, the average Silhouette score of all clusters provides a statistic of quality of the clustering method, which is comparable with other methods.

Datasets used to validate *ulrb*

To validate *ulrb* we used an original dataset presented in this article (Environmental Monitoring of Svalbard and Jan Mayer, MOSJ 2016–2020), along with publicly available datasets emulating diverse ecological contexts, to strengthen the validation of *ulrb* and cover a representative range of methodologies. The public datasets are: Norwegian Young Sea Ice Expedition (N-ICE), MOSJ 2019, Ants, BCI, and coral microbiome. A summary of the selected datasets and their major features is available in Table 1 (see also Data Availability). Below we provide a short description of the previously published datasets, with additional details on Supplementary Methods, followed by details on the MOSJ 2016–2020 dataset.

Validating *ulrb* for different phylogenetic units: the N-ICE dataset

The N-ICE dataset is composed of samples collected North of Svalbard in 2015[41], which were used for V4V5 16S rRNA gene amplicon sequencing and shotgun metagenomic sequencing[42]. The sequencing results were previously processed, using distinct bioinformatics approaches[20], resulting in amplicon sequence variants (ASVs, n = 9 samples), operational taxonomic units (OTUs, n = 9 samples), and metagenome derived OTUs (mOTUs, n = 9 samples). For extended details on sampling, sequencing and bioinformatics processing, see Supplementary Methods. A summary of the sequencing statistics of N-ICE is available in Supplementary Table 1.

Validating *ulrb* for different amplicon sequencing strategies: the MOSJ 2019 dataset

The MOSJ 2019 dataset is composed of samples collected during an expedition in Svalbard, in the framework of the Environmental Monitoring of Svalbard and Jan Mayer[43] (MOSJ) in 2019. Samples were collected for two different amplicon sequencing approaches[44], specifically: V4V5 16S rRNA gene amplicon sequencing, with Illumina technology, and full-length 16S rRNA gene amplicon sequencing, with Circular Consensus Sequencing PacBio technology. Initially, there were 18 samples available per amplicon sequencing strategy, but after filtering for the samples with high quality in both sequencing strategies, this number was reduced to 6 samples. For extended details sequencing and bioinformatics processing, see Supplementary Methods. Sequencing statistics were summarized in Supplementary Table 2.

Validating *ulrb* across varying sample sizes, sequencing depths and phylogenetic diversity: the MOSJ 2016–2020 dataset

We explored a time series of Arctic seawater samples collected for microbiome analyses, hereby referred to as the "MOSJ 2016–2020 dataset", published in this study, to test the robustness of *ulrb* under varying sample sizes, sequencing effort and phylogenetic diversity. Below we describe the sampling, sequencing, and data processing details of this dataset.

MOSJ 2016–2020: Sampling and sequencing details

Microbiome samples were collected from 2016 to 2020 (n = 119 samples) in a standardized way[45] in the framework of the Environmental Monitoring of Svalbard and Jan Mayer (MOSJ)[43]. Every year, during the summer season, the MOSJ campaign collects samples at several stations from the Kongsfjorden transect, covering the epipelagic, mesopelagic and bathypelagic layers. Details on sampling coordinates and depth for the samples that were used are available in Supplementary Data 1.

Seawater was filtered (mean = 2.9 L, sd = 1.4 L and n = 117 samples, Supplementary Data 1) through cartridge filters (0.22 µm pore size; Sterivex units) and DNA was extracted following the DNeasy PowerWater Sterivex Kit (Qiagen) and best practices from OSD[45]. Based on a previous work, variable filtration volume does not constitute a confounding variable[46]. For the amplification of V4V5 16S rRNA gene, the primers 515YF (5′-GTGYCAGCMGCCGCGGTAA-3′) and 926 R (5′ - CCGYCAATTYMTTTRAGTTT- 3′)[47–50] were used. Sequencing was performed with Illumina technology, on MiSeq platforms (2 x 300bp). This study integrates all 119 samples from MOSJ2016-2020 in a single dataset.

MOSJ 2016–2020: processing of V4V5 16S rRNA gene amplicons

To produce ASVs from V4V5 16S rRNA gene sequencing, we used a bioinformatic protocol based on DADA2[51]. Reads were trimmed at 249 nt (Forward) and 214 nt (Reverse) based on quality profiles of the entire MOSJ dataset (2016 to 2020) (Supplementary Fig. 2). Thus, the trimming criteria were the same for all years, as a compromise to allow standardization of ASV creation. Default parameters were used for the remaining steps of DADA2 protocol[51], which include the creation of an error model for quality filtering, identification of ASVs (*i.e*., the unique sequences), chimera removal and taxonomic assignment with Naive-Bayesian algorithm[52] and the Silva v138 database[53,54].

ASV tables were filtered to remove taxa attributed to unknown domain-level classifications, eukaryotes and organelles, if any; and singletons were removed, if any. ASV tables were rarefied at several rarefaction levels, considering the rarefaction curves (Supplementary Fig. 3), always discarding samples below the rarefaction threshold (n = 117 samples after this step).

Table 1 | Summary of datasets used in this study

Dataset name	Method	Microbial	Phylogenetic unit	Purpose in this study	Samples used	References
N-ICE	Amplicon sequencing of V4V5 16S rRNA gene	Yes	ASV	Validate *ulrb* for different kinds of phylogenetic units	9	20,41,42
N-ICE	Amplicon sequencing of V4V5 16S rRNA gene	Yes	OTU		9	
N-ICE	Shotgun metagenome sequencing	Yes	mOTU		9	
MOSJ 2019	Amplicon sequencing of V4V5 16S rRNA gene	Yes	ASV	Validate *ulrb* for different amplicon sequencing approaches	6	44
MOSJ 2019	Amplicon sequencing of full-length 16S rRNA gene	Yes	ASV		6	
MOSJ 2016–2020	Amplicon sequencing of V4V5 16S rRNA gene	Yes	ASV	Evaluate *ulrb* for different sample size, sequencing depth and number of ASV	117	This study.
Coral microbiome	Shotgun metagenome sequencing	Yes	mOTU	Identify types of rarity	20	55
Ants	Counts	No	Species	Validate *ulrb* for non-microbial data	98	27,56,57
BCI	Counts	No	Species		18	58
Total					292	

This table includes information on the relevant methodology used, if the dataset is microbial or not, the kind of phylogenetic unit used, the purpose of the dataset in this study, and the number of samples used.

For a summary of raw read processing statistics, see Supplementary Table 3.

Examining types of rarity with *ulrb*: the coral microbiome dataset

Depending on how the abundance classification changes, taxa can be grouped in types of rarity[3]. For example, if one taxon oscillates between being rare and abundant, it can be considered conditionally rare[34]. The current version of *ulrb* does not allow for the automatic calculation of the types of rarity. However, once the *ulrb* classification is obtained ("rare", "undetermined" and "abundant" classifications), it is possible to manually inspect how specific taxa change their classification across some variable. To test this possibility, we used the coral microbiome dataset, which includes samples characterized by shotgun metagenomic sequencing to describe coral host associations[55]. Specifically, samples were collected within the coral tissue (n = 13 samples), and in the sediment (n = 3 samples) and seawater (n = 4 samples) surrounding the corals. The corals selected are within the group of octocorals and include the species *Eunicella gazella* (n = 3 samples of healthy tissue, and n = 3 of necrotic tissue), *Eunicella verrucosa* (n = 4 samples of healthy tissue), and *Leptogorgia sarmentosa* (n = 3 samples). The 16S rRNA gene reads from the shotgun metagenomic dataset included 93,589 high-quality reads and 1041 mOTUs defined at a 97% similarity cut-off[55]. For extended details on sampling, sequencing and bioinformatics processing, see Supplementary Methods.

Validating *ulrb* for non-microbiome data: Ants and BCI datasets

The Ants dataset includes 49 different species surveyed at 99 sites[56,57]. This dataset was made available in the FuzzyQ R package[27]. For the purpose of this study, a site is equivalent to a sample. Prior to analysis with *ulrb*, one sample was removed from the Ants dataset (site 95) because of low sampling effort.

The Barro Colorado Island Tree Counts (BCI) is a publicly available dataset[58,59]. The BCI dataset used 50 plots of 1 hectare, surveyed over 35 years. For this study, a subset of the full BCI census was used to make a species abundance table, filtering alive trees and counting the number of species found in each combination of plot and year of survey (sample for our purpose). Then, we filtered samples with, at least, more than two tree species. Our final species abundance table, derived from a BCI subset, includes 327 tree species and 18 samples in a species abundance table.

Statistics and reproducibility

All statistical analyses and plots were produced using R software[60]. Several plots used the package *ulrb* (presented in here) together with ggplot2[61] and gridExtra[62]. Rarefaction was done using the rrarefy() function from the Vegan R package[63], to standardize the total number of reads per sample. When necessary, centrality metrics were used to avoid overlapping of samples in the plots. The centrality metric used was the mean ± standard deviation (sd), with the number of samples (n) indicated in the figure legend. To compare independent groups, we also used boxplots. For an alternative unsupervised approach to classify the rare biosphere, we used the FuzzyQ R package[27], which also calculates Silhouette scores for a statistical evaluation of results. For reproducibility, all source data and code are publicly available (see Source Code, and Data Availability statements). Biological replicates were defined as independent samples representing the properties on each independent group of samples being compared, and the sample size (n) was indicated in each analysis (Table 1). The source code and source data allow full reproduction of our results[64].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Testing *ulrb* for different kinds of phylogenetic units

To test *ulrb* applicability across common kinds of phylogenetic units, we used the N-ICE dataset. Briefly, we considered amplicon sequence variants (ASVs, n = 9 samples) and pre-computed Operational Taxonomic Units (OTUs, n = 9 samples) from V4V5 16S rRNA gene amplicon sequencing, and metagenomic operational taxonomic units (mOTUs) (n = 9 samples) from full-length 16S rRNA genes obtained via shotgun metagenomic sequencing. For ASVs, OTUs, and mOTUs, *ulrb* provided a RAC description of the microbial communities consistent with the classical view of the rare biosphere as the long tail of the RAC (Fig. 3a), showing that it can be used for either kind of phylogenetic unit.

To determine the statistical support of the unsupervised learning results, we calculated Silhouette scores for the datasets obtained with each phylogenetic unit (see Methods). The Silhouette scores were higher for OTUs and mOTUs than ASVs (Fig. 3b), meaning that clustering of phylogenetic units into abundance classifications by *ulrb* was overall more robust for OTUs and mOTUs than for ASVs. More than 75% of OTUs and mOTUs formed strong or reasonable clusters (Supplementary Fig. 4). However, 58% of the abundant ASVs formed weak clusters (Supplementary Fig. 4). OTUs formed strong clusters in all samples; the mOTUs formed either strong or reasonable clusters in all samples; and ASVs formed strong or reasonable clusters for "rare" and "undetermined" classifications, except for one sample (Supplementary Fig. 4). Although some phylogenetic units and clusters had lower Silhouette scores, the average Silhouette score indicated that the clustering structure across the entire dataset was strong or reasonable in all samples. This was consistent for all tested phylogenetic units, including ASVs (Supplementary Fig. 4). The fully automatic alternative of *ulrb* selected three clusters for ASVs and OTUs, but four clusters for mOTUs. Thus, based on average Silhouette scores (default settings), the ASV clustering could not be improved any further by using any other value of k. One possible reason why abundant ASVs were more difficult to cluster with *ulrb*, might be that it included more different abundance values than OTUs (216 ASVs vs 192 OTUs) and more extreme values (ASV maximum abundance = 5613 reads; OTU maximum abundance = 4825 reads). Note that this comparison refers to the *ulrb* statistical robustness against using different phylogenetic units (OTUs, ASVs, and mOTUs). It does not, however, imply any recommendation regarding which phylogenetic unit should be used in specific studies.

To verify if *ulrb* provides more consistent abundance classifications for different phylogenetic units in comparison with threshold-based methods, we examined the alpha diversity (number of ASVs/OTUs/mOTUs) within each classification obtained (rare, undetermined and abundant) when using *ulrb* and two threshold-based approaches (Fig. 4). Results obtained with *ulrb* showed a consistent trend for all phylogenetic units tested, revealing, in all cases, that the rare biosphere consisted of a larger richness of phylogenetic units than that of undetermined or abundant phylogenetic units (Fig. 4). Thus, *ulrb* reflected the shape of the RAC with better consistency than threshold-based definitions, which presented distinct patterns for each phylogenetic unit approach (Fig. 4). The absolute values of the response variable (number of ASVs/OTUs/mOTUs) are different, because the methodology is different, but they are consistent, since they have the same relationship. Thus, when using *ulrb*, the definition of rarity (and, by extension, the definition of "abundant" and "undetermined" classifications) had the same interpretation across phylogenetic units.

For perspective, we applied the same analysis using an alternative unsupervised learning approach to define the rare biosphere, using FuzzyQ (Supplementary Fig. 5). The FuzzyQ method worked as expected (Supplementary Fig. 5), presenting generally good quality clusters (Supplementary Fig. 5). Similar to *ulrb*, the phylogenetic unit with worse quality clusters was the ASVs (lower Silhouette scores, Supplementary Fig. 5).

Testing *ulrb* for different amplicon sequencing strategies

To test *ulrb* applicability to different amplicon sequencing strategies, we used the MOSJ 2019 dataset, which includes samples from short-reads (V4V5 region of the 16S gene, n = 6 samples), and long-reads (full-length 16S rRNA gene, n = 6 samples). For either sequencing approach, *ulrb* was able to characterize the classical RAC in a way that shows a long-tail of rare

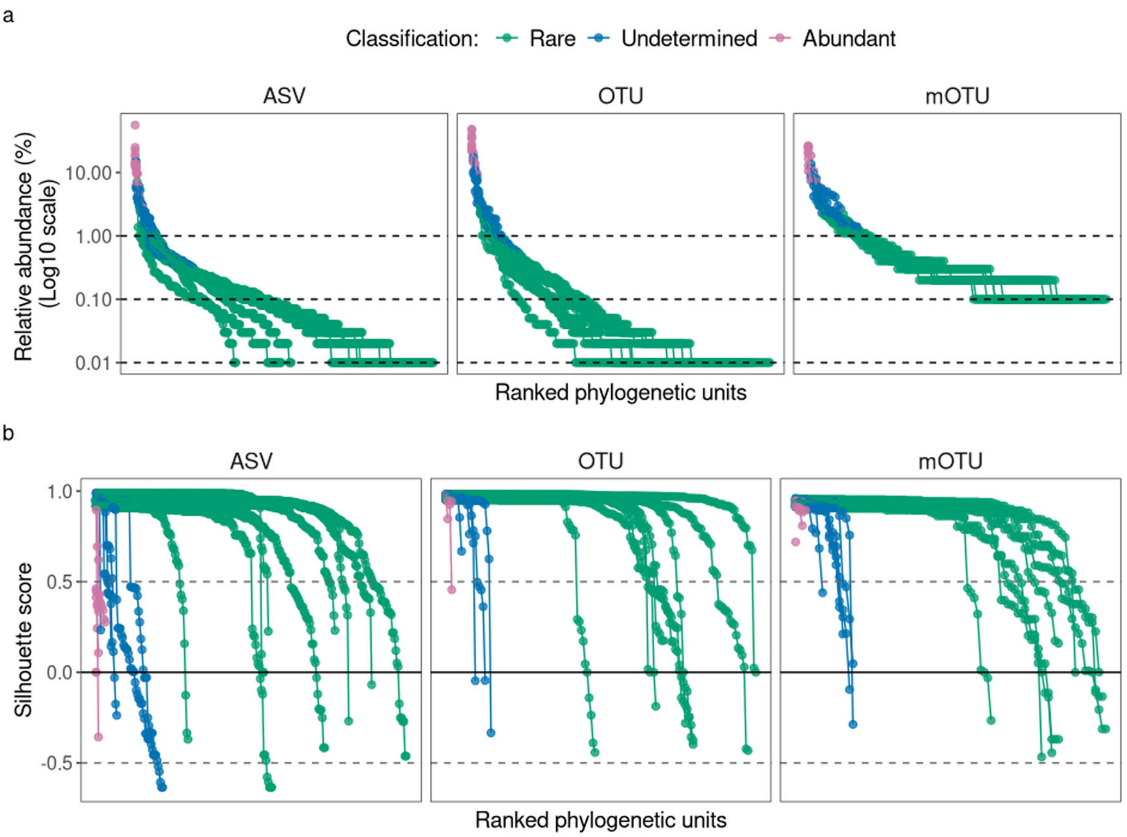

Fig. 3 | RAC and Silhouette score plots for N-ICE dataset. a Relative abundance of each phylogenetic unit (either ASV, OTU or mOTU) from all samples. For each sample, phylogenetic units were ordered in the x axis from the most to the least abundant. The y axis shows the abundance score of each phylogenetic unit in Log10 scale. For context, three relative abundance thresholds are highlighted with dashed lines (0.01%, 0.10% and 1.00%). **b** Silhouette scores obtained for each phylogenetic unit from all samples. Phylogenetic units are ordered from highest to lowest Silhouette score. For (**a**, **b**), lines were used to group phylogenetic units in the same sample and abundance classification. Default classifications (“rare”, “undetermined” and “abundant”) were colored coded.

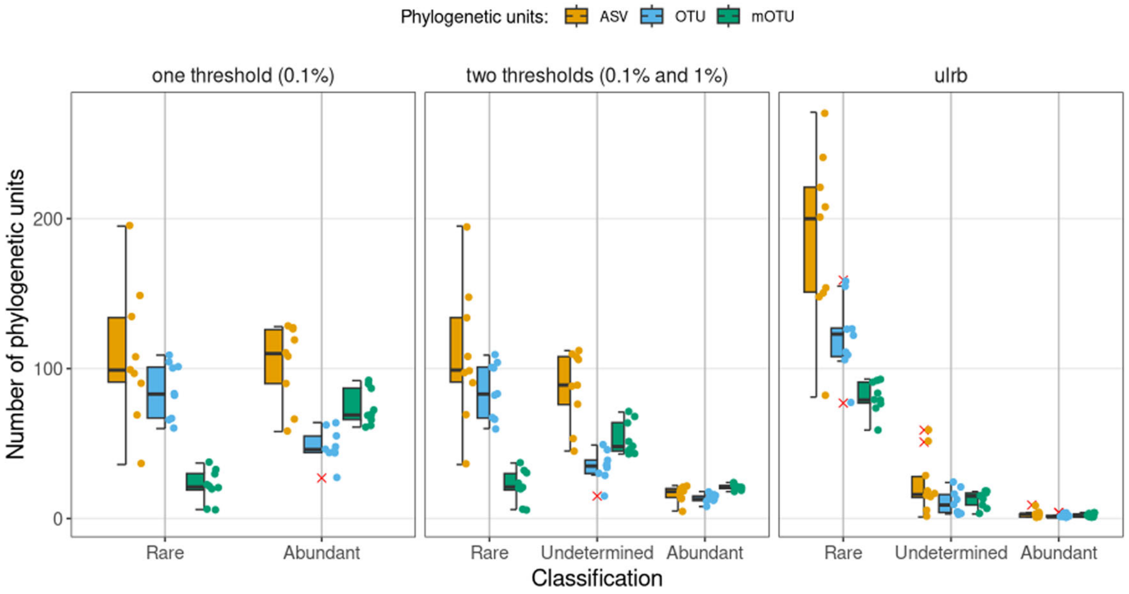

Fig. 4 | Comparison of number of rare, undetermined (if applicable) and abundant ASVs from ASVs, OTUs, and mOTUs. Abundance classifications were estimated with different methods: one single relative abundance threshold (0.1%, per sample), two relative abundance thresholds (0.1% and 1%, per sample) and *ulrb* using default parameters ($k = 3$). The number of ASVs was illustrated with boxplots and points for each sample (n = 6 samples). The outliers were marked with red cross.

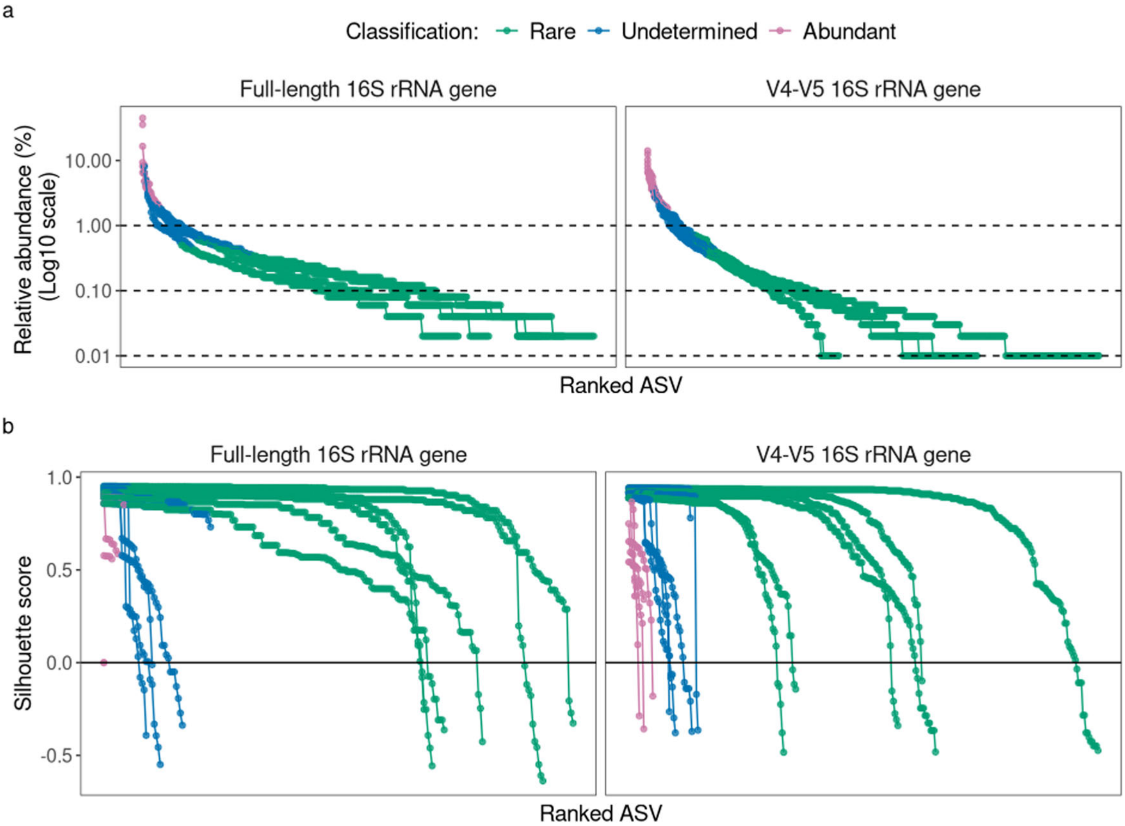

Fig. 5 | Comparison of ASVs derived from V4V5 and full-length 16S rRNA gene. **a** Relative abundance of each ASV in Log10 scale, ordered in the x axis from the most to the least abundant. For context, three relative abundance thresholds are highlighted with dashed lines (0.01%, 0.10% and 1.00%). **b** Silhouette score of each ASV. For both (**a**, **b**), lines were used to group ASVS in the same sample and abundance classification. ASVs were colored coded by abundance classification ("rare", "undetermined" and "abundant").

ASVs, followed by an intermediate region of undetermined ASVs and a few ASVs with very high abundance (Fig. 5a).

In terms of statistical quality of the unsupervised learning results, the Silhouette scores below 0.5 were more often attributed to abundant and undetermined than to rare ASVs, especially in the analysis of the full-length 16S rRNA gene approach (Fig. 5b). More than 75% of the rare and abundant ASVs from the full-length 16S rRNA gene approach formed strong or reasonable clusters, in contrast with the undetermined ASVs (with up to 34.5% weak and potentially artificial clustering) (Supplementary Fig. 6). For the V4V5 16S rRNA gene approach, abundant ASVs presented the weakest Silhouette scores (Supplementary Fig. 6b). For the full-length 16S rRNA gene approach, half of samples got either strong or reasonable clusters for any abundance classification (Supplementary Fig. 6). Regarding the V4V5 rRNA gene approach, half of the samples displayed weak clustering for the "abundant" classifications (Supplementary Fig. 6). When all clusters were considered, both approaches (V4V5 and full-length 16S rRNA gene sequencing) had strong or reasonable clustering results (Supplementary Fig. 6). Thus, the average Silhouette score never fell below 0.5, which means that the clusters found were not artificial. We attempted an improvement by using the automatic option of *ulrb*, but the automatic result (relying on average Silhouette scores) also selected three clusters.

We compared the consistency of different definitions of rarity between V4V5 and full-length 16S rRNA gene sequencing (Fig. 6). The most common approach to delineate the rare biosphere (0.1% relative abundance, per sample) resulted in a higher number of rare ASVs than abundant ASVs with the V4V5 region of the 16S rRNA gene, but the opposite was observed for the full-length 16S rRNA gene (Fig. 6). Using two thresholds also resulted in different patterns, with the number of ASVs going up and down from rare to undetermined to abundant for the full-length 16S rRNA gene, but always decreasing when the V4V5 region of the 16S rRNA gene was used (Fig. 6). Finally, the *ulrb* approach was the only one to provide the same pattern with both molecular methods, showing in each case a clearly higher richness of ASVs classified as rare than undetermined or abundant (Fig. 6). Thus, *ulrb* was able to provide a consistent definition of rarity between the two sequencing strategies, while the other two definitions, relying on relative abundance thresholds, failed to do so.

We tested the applicability of FuzzyQ in comparing amplicon V4V5 with full-length 16S rRNA gene sequencing (Supplementary Fig. 7). The method was able to classify taxa into common and rare but using three classifications (instead of two) would have been better for the full-length 16S rRNA gene sequencing data, because some ASVs were grouped near the threshold of 0.5 commonality index (Supplementary Fig. 7). Regardless of the number of clusters, the clustering quality was good for both sequencing strategies, except for a few common ASVs obtained from V4V5 16S rRNA gene sequencing approach (Supplementary Fig. 7).

Verifying robustness of *ulrb* against sample size, sequencing depth and number of taxa

To verify the robustness of *ulrb* we used the MOSJ2016-2020 dataset, which includes up to 117 Arctic seawater samples characterized by 16S rRNA gene sequencing and processed with the DADA2 pipeline for ASV-based diversity assessments (see Methods). We tested the quality of clustering (measured by average Silhouette score) as a function of three variables that distinguish datasets: (1) number of samples (n); (2) number of taxa (ASVs in this context); and (3) sequencing depth, per sample.

To test the effect of sample size (n), we locked the sequencing depth at 10,000 reads (using rarefaction), resulting in a total pool of 114 high-quality samples. Then, we subsampled random samples, without replacement, from the pool of 114 high-quality samples. At each step (from n = 6 to n = 114), we applied *ulrb* to all the samples and then calculated

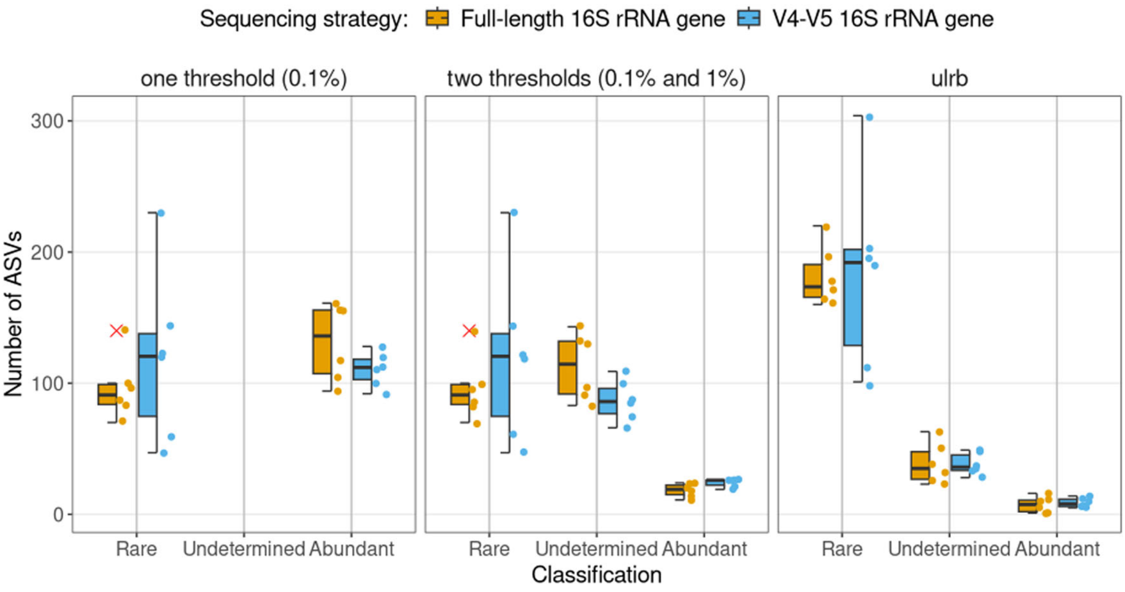

Fig. 6 | Comparison of number of rare, undetermined (if applicable) and abundant ASVs from V4V5 and full-length 16S rRNA gene sequencing. Abundance classifications were estimated with different methods: one single relative abundance threshold (0.1%, per sample), two relative abundance thresholds (0.1% and 1%, per sample) and *ulrb* using default parameters ($k = 3$). The number of ASVs was illustrated with boxplots and points for each sample (n = 6 samples). There were no outliers.

the average Silhouette score of each sample and plotted the mean ± sd (Fig. 7a). Results showed that *ulrb* provided high quality clustering (average Silhouette scores >0.75) for the rare biosphere with low ($n < 30$) and high ($n > 30$) sample size. The "undetermined" and "abundant" classifications, similarly to the previous sections (Fig. 3 and Fig. 5), presented lower quality. However, the "undetermined" classification presented mostly reasonable clusters, and the "abundant" classification varied between weak and reasonable clusters (Fig. 7a). Importantly, the average Silhouette scores presented more random variation at the "undetermined" and "abundant" classification at low sample size ($n < 30$) than at large sample size ($n > 30$). In fact, above 30 samples, *ulrb* results were very robust for all abundance classifications (Fig. 7a).

To test the robustness of *ulrb* against different number of taxa (ASVs in this context), we selected 34 samples and rarefied them to 50,000 reads, to have as many ASVs as possible and at least $n > 30$ samples. Then, we collected random ASVs (from 100 ASVs to up to 4000 ASVs) per sample, without replacement. Figure 7b shows that, for this set of samples, all abundance classifications obtained very good scores (average Silhouette score >0.75). Importantly, the number of ASVs clearly had no effect on the quality of the clustering obtained by *ulrb*. To show that the random selection of ASVs was able to keep the RAC shape and was not exclusively obtaining ASVs of one single abundance classification, we illustrate the RAC obtained by a random selection of 100, 1000, and 3000 ASVs in a random sample (Supplementary Fig. 8).

To test the impact of sequencing depth, we selected the 34 samples with more reads and applied different rarefaction levels to them (from 1000 reads to up to 50,000 reads). Remarkably, *ulrb* was extremely robust for variations in sequencing depth, since the average Silhouette score was almost perfectly constant as a function of sequencing depth (Fig. 7c). As in the sample size analysis, the "rare" classification presented better quality than the "undetermined" and "abundant" classifications (Fig. 7c).

In summary, by applying variations in specific features of a large dataset, we showed that *ulrb* presented robust results for variations in sample size, number of taxa (ASVs in here) and sequencing depth. Since any abundance table will ultimately vary because of a combination of different number of taxa, samples and order of magnitude of the abundance score, we present evidence that *ulrb* is robust for a wide range of abundance tables (Fig. 7).

Finally, we verified the impact of the same variables on the application of FuzzyQ and found that this method generally presented high quality clustering for the "rare" classification, but potentially artificial clusters for the "common" classification (Supplementary Fig. 9). However, FuzzyQ results improved for larger datasets ($n > 30$ and ASVs > 700), and it was not limited by sequencing depth (Supplementary Fig. 9).

Validating *ulrb* for non-microbial datasets

To test if *ulrb* can be applicable for non-microbiome datasets, we applied *ulrb* to animal and plant datasets that were publicly available, the Ants[27,56] and the BCI[59] datasets. *ulrb* was able to classify all ant species into abundance categories, depicting a few species that were abundant, undetermined or rare in different samples and also a long tail of rare species (Fig. 8a). The clustering quality was also good for most species (average Silhouette score >0.75), with very few species in low quality clusters (Fig. 8b).

Similarly, *ulrb* was able to classify rare, undetermined and abundant tree species in the BCI dataset (Fig. 8c, d). Furthermore, the classifications obtained showed a reasonable division between abundance scores, illustrating the applicability of *ulrb* (Fig. 8a). The Silhouette plot reveals that *ulrb* provided robust classifications for most species, with only a few presenting low average Silhouette scores (Fig. 8d).

Using *ulrb* to establish types of rarity

To show that *ulrb* can be used to monitor taxa and, therefore, describe types of rarity, we used a coral microbiome from a shotgun metagenomic sequencing dataset[55]. Specifically, we monitored the classifications obtained for a selected group of mOTUs (561, 559 and 866, based on Keller-Costa et al.[55]) across different coral species, health status and surrounding environment, in a way that effectively described different types of rarity (Fig. 9). The mOTUS (561, 559, and 866) were selected based on previous knowledge of their ecology and adequacy to describe types of rarity. OTU 561 (genus *Anaerospora*) was absent in healthy coral tissue and in sediment but was a member of the seawater rare biosphere and colonized the necrotic coral tissue, becoming rare or undetermined (Fig. 9). Thus, *ulrb* helped identify OTU 561 as a potential necrotic tissue colonizer and established its origin in the seawater rare biosphere. Another example of a necrotic tissue colonizer was OTU 559 (family *Rhodobacteraceae*), which was abundant in necrotic tissue and seawater, but rare or undetermined in healthy octocoral tissue.

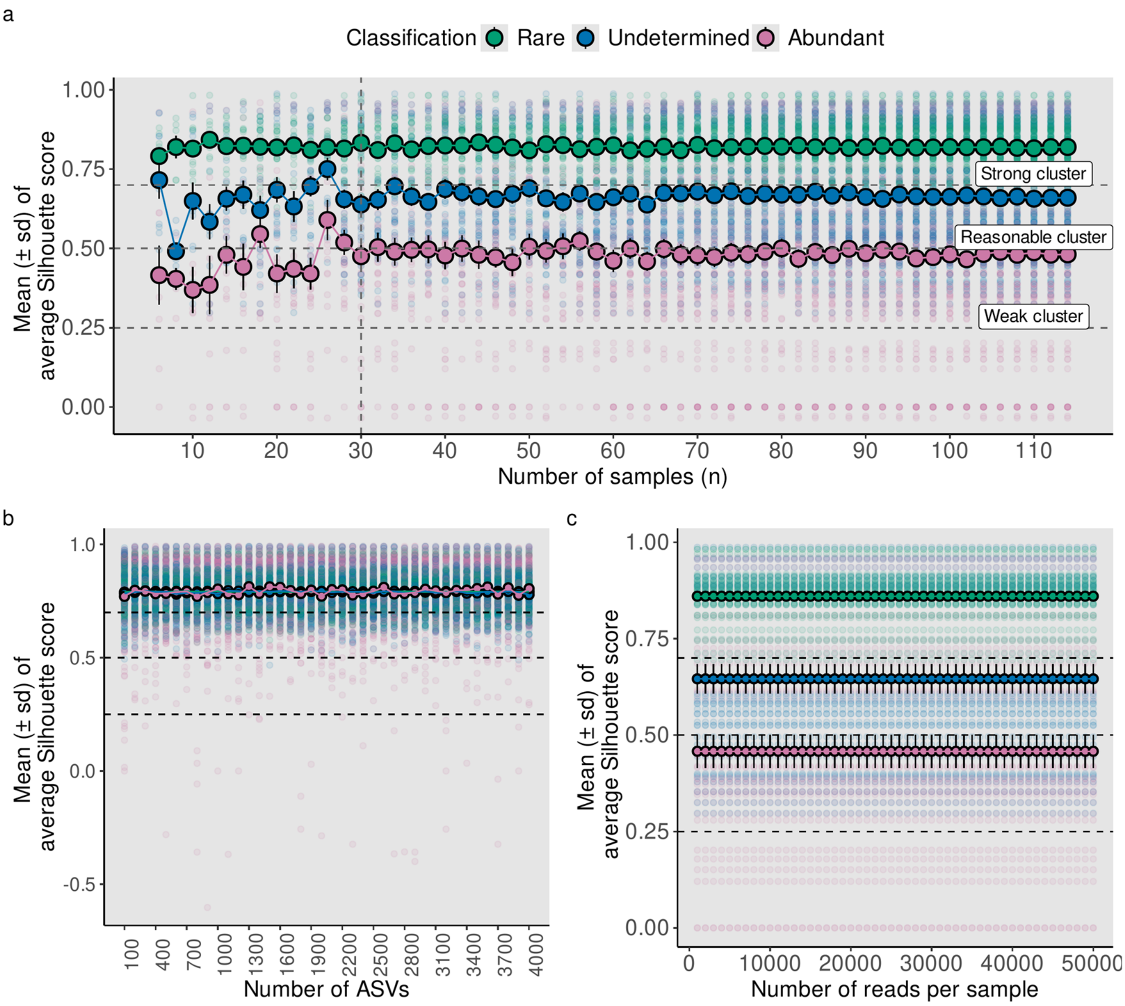

Fig. 7 | Quality of *ulrb* clustering measured by the average Silhouette score as a function of number of samples, ASVs, and sequencing depth. a Number of samples (n), ranging from 6 samples up to 114 samples, in increments of two samples; **b** number of ASVs, ranging from 100 to 4000 ASVs per sample, in increments of 300 ASVs; **c**, number of reads per sample, in increments of 1000 reads. The samples (**a**), ASVs (**b**) and reads (**c**) were randomly selected in each increment, without replacement. For the response variable, the mean (±sd) of the average Silhouette score was used and the classifications were grouped by different colors, as illustrated in the figure legend. The MOSJ2016-2020 dataset was used for this figure.

This result indicates that specific, abundant members of the seawater microbiome (in this case, a *Rhodobacterales* phylotype) may belong to the rare biosphere of healthy, host-associated microbiomes and rapidly colonize decaying host tissue, transitioning from rare to abundant while the symbiotic microbiome enters the dysbiosis state (Fig. 9). A contrary example is OTU 866 (family *Endozoicomonadaceae*), which was abundant in healthy coral tissues, but became rare or undetermined under necrosis (except for EG18_N) and was rare or absent in the sediment and seawater samples (Fig. 9). Thus, *ulrb* indicates that this phylotype in the family *Endozoicomonadaceae* represents a coral symbiont enriched in healthy while depleted in necrotic tissues.

Discussion

Microbial ecology studies usually delineate rare from abundant taxa based on relative abundance thresholds[2,3]. Here we propose the *ulrb* method, which automatically clusters taxa based on the relationship between their abundances in a given sample, without the need of a threshold selection. Thus, the common observation that most taxa are "rare" means that these are within a small range of low abundance values, while the few "abundant" have a disproportionately higher abundance. In microbial ecology, the killing-the-winner hypothesis[65,66], in which lower cell abundance decreases the probability of encountering bacteriophages, is often evoked to explain a possible ecological strategy underlying the existence of so many rare taxa within a community. In addition, some microorganisms might be dormant but keep the ability to grow and become abundant under conditions that are more favorable[34]. Other microorganisms are able to keep high metabolic activity, even though at low abundance[67]. Ecological effects, such as dispersion and drift might also contribute for the emergence of some rare taxa[1,24]. Finally, some rare taxa might be decreasing their abundance towards local extinction[2]. Previous reviews have summarized evidence for these mechanisms[24]. Since *ulrb* is an unsupervised machine learning method, it makes no assumptions about metabolic or ecological mechanisms shaping community composition, *i.e.*, the classification solely depends on the abundance table provided. However, the resulting classifications can be explained by such ecological mechanisms. For example, an abundant taxon that becomes rare could indicate the existence of a top-down factor, while the sudden emergence of a rare taxon previously unreported in a specific environment could indicate dispersion effects.

The most used methods to define the rare biosphere are problematic, because they are based on arbitrary thresholds of relative abundance[20,24]. To provide concrete numbers, we summarize the literature on the microbial rare biosphere, from January 2006 to 2024 (Supplementary Table 4). Of 181 articles, approximately 37% did not provide a clear methodology to define the rare biosphere and, among those that defined the rare biosphere explicitly, approximately 84% used relative abundance thresholds (Supplementary Table 4). Within the studies that relied on relative abundance

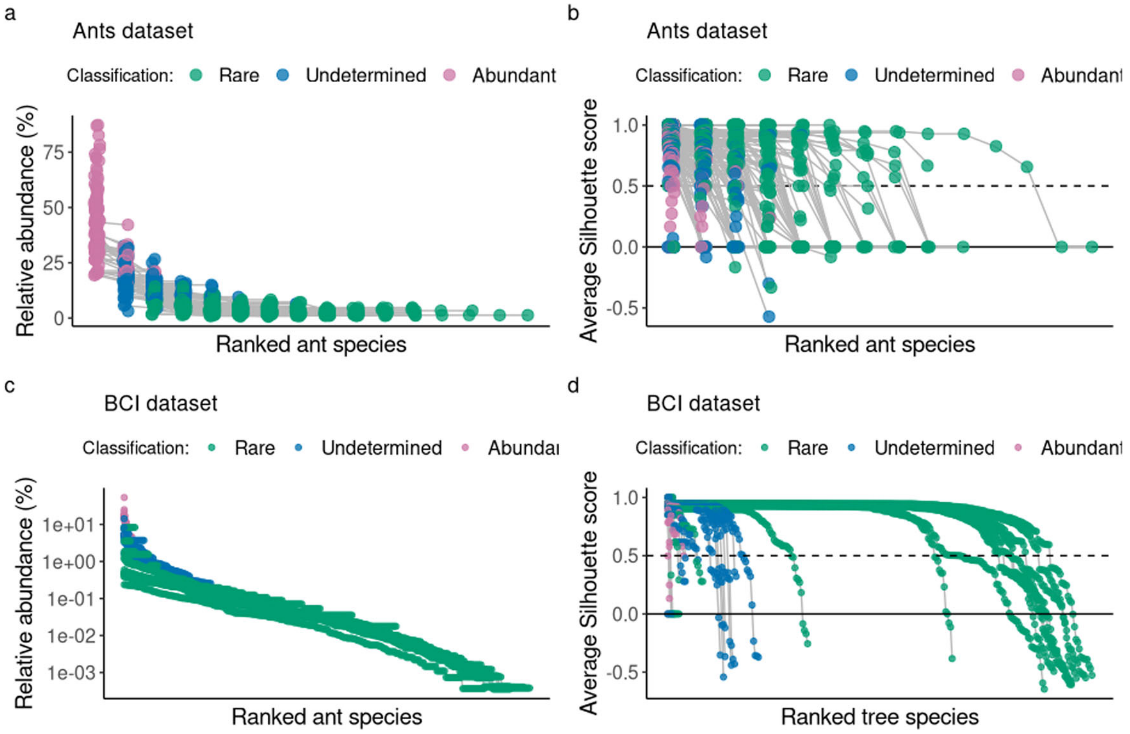

Fig. 8 | Analysis of *ulrb* applicability to Ants and BCI datasets. a Mean ± sd of relative abundance of ranked species from the Ants dataset (n = 98 samples), colored by classification; **b** mean ± sd of average Silhouette score of ranked species from the Ants dataset (n = 98 samples), colored by classification; **c** mean ± sd of relative abundance of ranked species from the BCI dataset (n = 18 samples), colored by classification; **d** mean ± sd of average Silhouette score of ranked species from the BCI dataset (n = 18 samples), colored by classification.

thresholds, approximately 60% used a single threshold, while the remaining used two or more thresholds (Supplementary Table 4). Approximately half of the studies used 0.1% relative abundance to distinguish rare from abundant phylogenetic units within communities, with approximately 70% applying the threshold per sample, instead of applying it to the whole dataset at once (Supplementary Table 4).

We compared two of the most common approaches to define the rare biosphere against *ulrb*, specifically, the utilization of a single threshold of 0.1% relative abundance per sample, and the alternative including an intermediate level of relative abundance ranging from 0.1 to 1%. To do this comparison, we applied the different definitions of rarity to environmental replicates assessed by different methods (V4V5 and full-length 16S rRNA gene sequencing and metagenomics), showing that threshold-based definitions have patterns of diversity that are method-dependent, while *ulrb* provided the same pattern across all methods. More specifically, using threshold-based definitions, the number of rare and abundant taxa was inconsistent across methodologies, but it was very consistent when using *ulrb*. This is because threshold-based methods do not accommodate the differences in sequencing depth and variability of taxa abundance, unlike *ulrb*. *ulrb* captures the rarity concept without the need for arbitrary thresholds in a way that is consistent across datasets, because it solely depends on the relative distance between taxa abundance scores. This ability to capture connections between the abundance of taxa, independently of the order of magnitude of the abundance scores, provides classifications that are non-random and both biologically and ecologically informative.

The clustering results from *ulrb* were generally stronger for OTUs and mOTUs than for ASVs (based on Silhouette scores, Supplementary Figs. 4 and 6). ASVs may be harder to cluster into abundance classifications, because they are more prone to extreme values, which will affect the clustering result, for example, by creating a single cluster for outliers. This problem can be solved by removing outliers[31], but in this context abundant taxa are outliers that must be kept, because they represent real taxa. Therefore, we propose that taxa that are outliers relative to the remaining taxa should be considered abundant taxa, even if it means that very few taxa are defined as abundant. Another factor that might contribute for the difficulty of clustering ASVs by *ulrb* is that such datasets usually consist of diverse, highly similar phylogenetic units (e.g., sequences diverging in few nucleotides from one another), each of which possessing its own abundance score, but frequently representing one single microbial species (or subpopulations within one species)[44].

The *ulrb* method proved to be statistically robust for any variation of the main variables shaping an abundance table. Collectively, the classification of taxa into "rare" was usually of better statistical quality than the "undetermined" and "abundant" classifications. A reason for this is that the "rare" classification includes more taxa than the "undetermined" and "abundant" classification, which in turn gives the "rare" classification stronger clusters. Additional evidence for this assertion is that if we randomly select a certain number of taxa, the clustering quality becomes equivalent for all classifications. Another reason for the observation of stronger clusters in the rare biosphere is that the variability of abundance among rare taxa is much lower, thus contributing to better defined clusters. The robustness of *ulrb* was not affected in any way by the sequencing depth, which explains why *ulrb* was able to provide consistent results for microbial datasets derived from different sequencing methodologies. In fact, the reason why we cannot use the same relative abundance threshold for 16S rRNA gene metabarcoding (amplicon sequencing) and 16S rRNA gene data derived from shotgun metagenome sequencing is precisely the different order of magnitude of the datasets, which is an issue that is not solved by the compositional nature of the data. Furthermore, *ulrb* was also statistically robust independently of the number of samples, even though the Silhouette scores of the "undetermined" and "abundant" classifications varied more for datasets containing less than 30 samples. This was expected, because a low sample size (n < 30) may not be enough to characterize the mean value of a distribution of data[68]. Since *ulrb* is applied to a specific sample and its result is not impacted by the existence of other samples, a perfect result would be a horizontal line, representing no variation in the clustering quality. However,

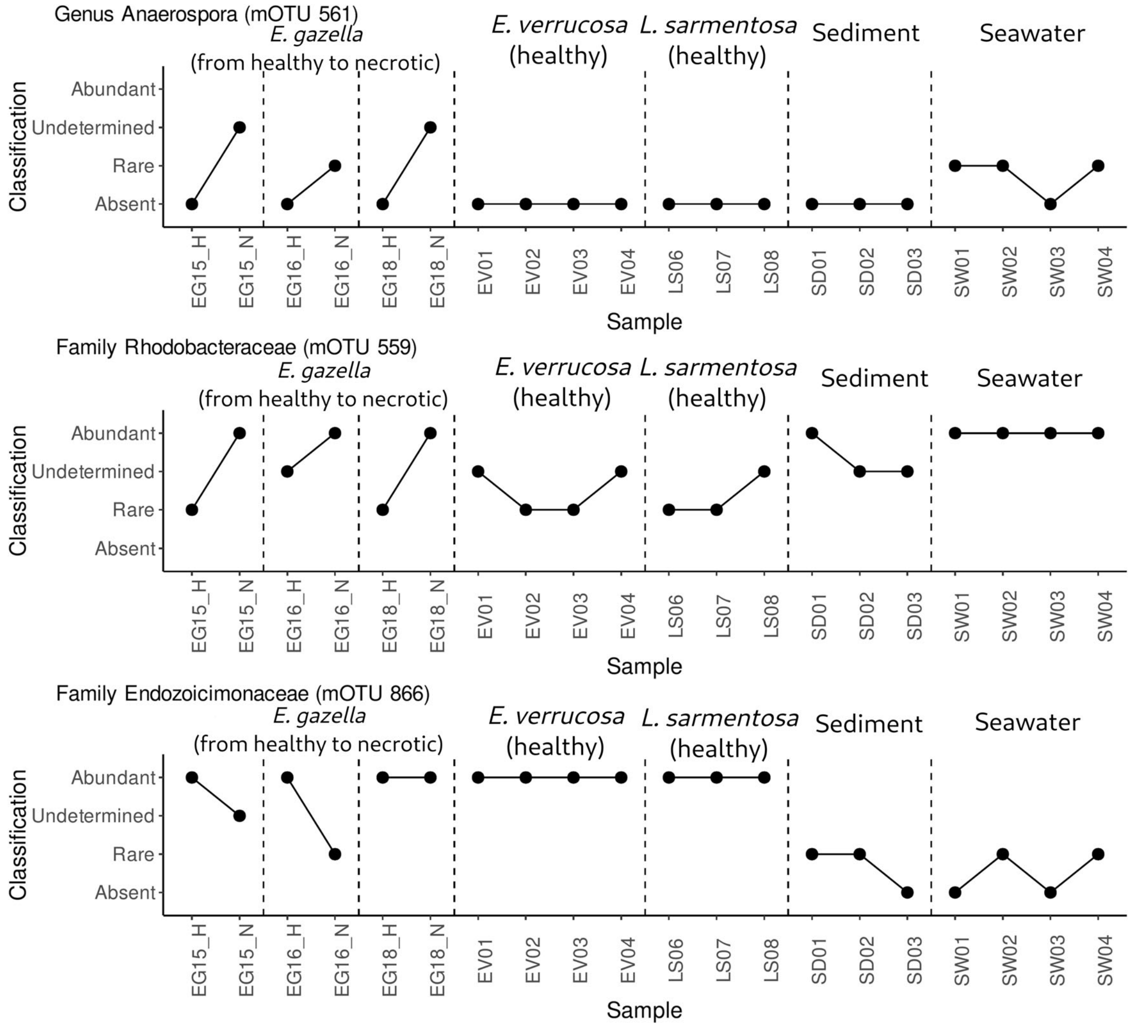

Fig. 9 | Monitoring of the abundance classifications of selected OTUs. Samples are represented on the x axis and the abundance classification on the y axis, with lines connecting samples that are related. For *Eunicella gazella* species, lines connect healthy to necrotic tissue (EG15_H, EG16_H, EG18_H and EG15_N, EG16_N, EG16_N), while *Eunicella verrucosa* (EV01, EV02, EV03, EV04) and *Leptogorgia sarmentosa* (LS06, LS07, LS08) only presented healthy tissue. Additionally, sediment (SD01, SD02, SD03) and seawater (SW01, SW02, SW03, SW04) samples were also represented in distinct groups. The different coral species, sediments and seawater were separated by vertically dashed lines. Relevant taxonomic information of the selected OTUs is indicated at the top of each plot. To indicate that an OTU was absent in a sample, we added the classification "absent", which is not obtained from *ulrb*.

because we are selecting random samples from the dataset, those samples will have random variation between them. As the sample size increases, the random variation decreases and approaches the true average.

ulrb was designed for handling large abundance tables derived from molecular analyses of microbial communities. Yet we showed that it can also be applied to non-microbial data, using the ants and plants datasets (Fig. 8). This was expected, because *ulrb* relies on the relative distance between the abundance scores of the taxa within a sample. Furthermore, the microbial and non-microbial abundance tables have the same underlying structure, with differences in the number of taxa and the abundance score of those taxa. Thus, since we showed that *ulrb* was consistent for any variation in taxa numbers, sequencing depth and number of samples (Fig. 7), *ulrb* was expected to work properly for non-microbial data.

Outside the scope of microbial ecology, a previous study has suggested the utilization of unsupervised learning to define rare and common taxa, using FuzzyQ[27]. FuzzyQ has an analogous framework to *ulrb*, because both are able to define rare taxa without the introduction of arbitrary thresholds, i.e., they provide automatic classifications. However, there are several differences between both methods, because of the number of features used (Supplementary Fig. 10), making it unreasonable to directly compare them. However, we applied FuzzyQ to similar data in parallel to provide perspective and identify potential advantages and disadvantages. Briefly, both methods can be used to define the rare biosphere for microbial and non-microbial data, but *ulrb* provides information at sample level, while FuzzyQ provides information at the whole dataset level. Thus, one disadvantage of FuzzyQ relative to *ulrb* is that it is not clear if a taxon is common/rare due to the frequency of occurrence or to its underlying abundance in the study. Consequently, it provides little information on the transition between rare and abundant states for a given taxon across samples. Some advantages of FuzzyQ include the commonality index, which indicates how common or rare a taxon is with a particular dataset. Additionally, the automatic inclusion of frequency of occurrence provides information on commonality, showing how often a taxon appears across samples. This information about the "rare" and "abundant" classification can be useful in certain experimental settings.

We showed that *ulrb* can be adapted to manually inspect the types of rarity within a given dataset, using data derived from a coral microbiome study[55]. Such an analysis supported the identification of likely mutualistic octocoral symbionts, such as members of the family *Endozoicomonadaceae*, which were abundant in healthy coral tissues, but rare or absent in most necrotic tissues, sediment and seawater. The same approach also allowed the identification of the genus *Anaerospora* as a seawater rare biosphere member with the ability to colonize necrotic corals, but not healthy ones. Such examples demonstrate that *ulrb* can be easily adapted to ascertain different types of rarity by monitoring selected taxa across relevant variables. However, *ulrb* is currently unable to automatically calculate types of rarity for any dataset, which means that the user must manually do such monitoring. We foresee the implementation of such capabilities in future versions of *ulrb*.

On the microbial side, this study focused specifically on prokaryotes, but *ulrb* should work equally well for other microbial groups (e.g., fungi and protists) obtained with high-throughput sequencing methods, because the data will have similar characteristics. Furthermore, any variation in such datasets will necessarily be within differences in number of taxa, samples and sequencing depth, which we showed did not have any impact on *ulrb* robustness and applicability.

The identification of types of rarity across a set of samples, the optimal number of abundance clusters to be used, and the eventual occurrence of clusters represented by outliers are all challenges that need to be met in current rare biosphere research. For each case, the present version of *ulrb* offers possible solutions, but also presents limitations. We show that types of rarity can be defined by manually inspecting target taxa, but we lack an automatic approach to do so in the current version; we suggest a standard number of clusters ($k = 3$), which might not be adequate for some experimental settings; and *ulrb* may also produce clusters composed of a single outlier taxon, which is explained by the extremely high abundance of such taxon relative to the remainder. Those limitations can be mitigated with tools available in the current version, but future work will attempt to solve those issues. In terms of computational power, since *ulrb* applies its calculations on a single dimension, it is quite fast. It is worth noting that if different studies select different numbers of clusters, then intercomparability across studies might be compromised.

Conclusion

This study presents the *ulrb* R package, with a methodology to define the rare biosphere across microbial communities. This R package is open-source and includes a dedicated website (https://pascoalf.github.io/ulrb/), with tutorials explaining how to use *ulrb* functions and extensive documentation.

We show that *ulrb* provides a more consistent interpretation of the microbial rare biosphere across different sequencing strategies and bioinformatic protocols than threshold-based methods, because it is statistically robust against variations in taxa counts, sequencing depth and number of samples. We demonstrate that *ulrb* can also be used for non-microbial data, because it depends only on the relative differences between the abundance scores of taxa within a community. Thus, *ulrb* is effectively independent of the methodology used to produce the abundance table.

Finally, we show that *ulrb* results can be used to manually monitor specific taxa and ascertain types of rarity[3,34]. However, future work is necessary to implement an automatic classification of types of rarity in the *ulrb* R package.

Owing to the features mentioned above, *ulrb* is readily applicable to discern rare from abundant organisms across various scenarios, showing great potential to standardize microbial rare biosphere analysis. *ulrb* can be used, but is not limited, to studying transitions from eubiosis to dysbiosis states in host-associated microbiomes, emerging microbial diseases because of climate change, biological invasions, community gradient analyses and landscape ecology data, to name a few possible applications.

Data availability

The FASTQ files from the N-ICE dataset are available at European Nucleotide Archive (ENA), with project ID PRJEB21950 (V4V5 16S rRNA gene amplicon sequencing) and PRJEB15043 (shotgun sequencing of metagenomes). The FASTQ files from the MOSJ dataset are all available in the projects PRJEB24517 (2016), PRJEB72025 (2017), PRJEB72030 (2018), PRJEB60815 (2019) and PRJEB72034 (2020). The Ants dataset used is available in the R package FuzzyQ[27]. The BCI dataset used was derived from original data made publicly available in DRYAD (https://doi.org/10.15146/5xcp-0d46)[58]. The octocoral microbiome dataset shotgun metagenomic sequencing data is available in ENA, under project PRJEB13222. Source data for all plots and analyses is available in the GitHub repository (https://doi.org/10.5281/zenodo.14922332)[64].

Code availability

The source code for *ulrb* R package is available at GitHub (https://github.com/pascoalf/ulrb) and CRAN. All the code to process raw reads and reproduce the figures and tables in this paper are available in a GitHub repository (https://doi.org/10.5281/zenodo.14922332)[64]. The code used for the current version of *ulrb* R Package (0.1.6) is also available in a repository (https://doi.org/10.5281/zenodo.14922442)[69].

Received: 8 March 2024; Accepted: 10 March 2025;
Published online: 02 April 2025

References

1. Pascoal, F., Costa, R. & Magalhães, C. The microbial rare biosphere: current concepts, methods and ecological principles. *FEMS Microbiol. Ecol.* **97**, 1–15 (2021).
2. Pedrós-Alió, C. The rare bacterial biosphere. *Annu. Rev. Mar. Sci.* **4**, 449–466 (2012).
3. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**, 217–229 (2015).
4. Jousset, A. et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J.* **11**, 853–862 (2017).
5. McGill, B. J. et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* **10**, 995–1015 (2007).
6. Darwin, C. *The Origin of Species* (Amsterdam University Press, 1859).
7. Gaston, K. & Fuller, R. Commonness, population depletion and conservation biology. *Trends Ecol. Evol.* **23**, 14–19 (2008).
8. Sogin, M. L. et al. Microbial diversity in the deep sea and the underexplored 'rare biosphere. *Proc. Natl Acad. Sci. USA* **103**, 12115–12120 (2006).
9. Taylor, M. W. et al. Sponge-specific' bacteria are widespread (but rare) in diverse marine environments. *ISME J.* **7**, 438–443 (2013).
10. Pascoal, F., Magalhães, C. & Costa, R. The link between the ecology of the prokaryotic rare biosphere and its biotechnological potential. *Front. Microbiol*. **11**, https://doi.org/10.3389/fmicb.2020.00231 (2020).
11. Ser-Giacomi, E. et al. Ubiquitous abundance distribution of non-dominant plankton across the global ocean. *Nat. Ecol. Evol.* **2**, 1243–1249 (2018).
12. Quero, G. M. & Luna, G. M. Diversity of rare and abundant bacteria in surface waters of the Southern Adriatic Sea. *Mar. Genom.* **17**, 9–15 (2014).
13. Fuentes, S., Barra, B., Caporaso, J. G. & Seeger, M. From rare to dominant: a fine-tuned soil bacterial bloom during petroleum hydrocarbon bioremediation. *Appl. Environ. Microbiol.* **82**, 888–896 (2016).
14. Idris, H., Goodfellow, M., Sanderson, R., Asenjo, J. A. & Bull, A. T. Actinobacterial rare biospheres and dark matter revealed in habitats of the Chilean Atacama Desert. *Sci. Rep.* **7**, 1–11 (2017).
15. Richa, K. et al. Distribution, community composition, and potential metabolic activity of bacterioplankton in an urbanized Mediterranean Sea Coastal Zone. *Appl. Environ. Microbiol.* **83**, 1–17 (2017).
16. Dawson, W., Hör, J., Egert, M., van Kleunen, M. & Pester, M. A small number of low-abundance bacteria dominate plant species-specific responses during rhizosphere colonization. *Front. Microbiol.* **8**, 1–13 (2017).
17. Wang, Y. et al. Quantifying the importance of the rare biosphere for microbial community response to organic pollutants in a freshwater ecosystem. *Appl. Environ. Microbiol.* **83**, e03321–16 (2017).

18. De Anda, V. et al. Understanding the mechanisms behind the response to environmental perturbation in microbial mats: a metagenomic-network based approach. *Front. Microbiol.* **9**, 1–24 (2018).
19. Gokul, J. K. et al. Illuminating the dynamic rare biosphere of the Greenland Ice Sheet's Dark Zone. *FEMS Microbiol. Ecol.* **95**, 1–17 (2019).
20. Pascoal, F., Costa, R., Assmy, P., Duarte, P. & Magalhães, C. Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies. *Microb. Ecol.* **84**, 59–72 (2022).
21. Tang, L. et al. Plant community associates with rare rather than abundant fungal Taxa in Alpine Grassland Soils. *Appl. Environ. Microbiol.* **89**, 1–15 (2023).
22. Pedrós-Alió, C. Marine microbial diversity: can it be determined? *Trends Microbiol.* **14**, 257–263 (2006).
23. Gobet, A., Quince, C. & Ramette, A. Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res.* **38**, e155–e155 (2010).
24. Jia, X., Dini-Andreote, F. & Falcão Salles, J. Community assembly processes of the microbial rare biosphere. *Trends Microbiol.* **26**, 738–747 (2018).
25. Jia, X., Dini-Andreote, F. & Salles, J. F. Unravelling the interplay of ecological processes structuring the bacterial rare biosphere. *ISME Commun.* **2**, 1–11 (2022).
26. Ramond, P., Siano, R., Sourisseau, M. & Logares, R. Assembly processes and functional diversity of marine protists and their rare biosphere. *Environ. Microbiome* **18**, 1–14 (2023).
27. Balbuena, J. A. et al. Fuzzy quantification of common and rare species in ecological communities (FuzzyQ). *Methods Ecol. Evol.* **12**, 1070–1079 (2021).
28. Kaufman, L. & Rousseeuw, P. J. Clustering by means of Medoids. *Stat. Data Anal. Based L1 Norm. Relat. Methods* **405**, 416 (1987).
29. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* **3**, 1–27 (1974).
30. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell*. **PAMI-1**, 224–227 (1979).
31. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
32. Vergin, K., Done, B., Carlson, C. & Giovannoni, S. Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. *Aquat. Microb. Ecol.* **71**, 1–13 (2013).
33. Baltar, F. et al. Response of rare, common and abundant bacterioplankton to anthropogenic perturbations in a Mediterranean coastal site. *FEMS Microbiol. Ecol.* **91**, 1–12 (2015).
34. Shade, A. et al. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* **5**, 1–9 (2014).
35. Jones, S. E. & Lennon, J. T. Dormancy contributes to the maintenance of microbial diversity. *Proc. Natl Acad. Sci. USA* **107**, 5881–5886 (2010).
36. Vinod, H. D. Integer programming and the theory of grouping. *J. Am. Stat. Assoc.* **64**, 506–519 (1969).
37. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. Cluster: cluster analysis basics and extensions. R package v2.1.6 (2023).
38. Wickham, H., Hester, J., Chang, W. & Bryan, J. devtools: Tools to Make Developing R Packages Easier. R package v2.4.5 (2022).
39. Kaufman, L. & Rousseuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. *Biometrics* https://doi.org/10.2307/2532178 (1990).
40. Walesiak, M. & Dudek, A. The Choice of Variable Normalization Method in Cluster Analysis. in *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges* (ed. Soliman, K. S.) 325–340 (International Business Information Management Association (IBIMA), 2020).
41. Granskog, M. et al. Arctic Research on thin ice: consequences of Arctic Sea Ice Loss. *Eos* **97**, https://doi.org/10.1029/2016EO044097 (2016).
42. de Sousa, A. G. G. et al. Diversity and composition of pelagic prokaryotic and protist communities in a thin Arctic Sea-Ice regime. *Microb. Ecol.* **78**, 388–408 (2019).
43. Renner, A. H. H., Dodd, P. A. & Fransson, A. An Assessment of MOSJ - The State of the Marine Environment around Svalbard and Jan Mayen. (Norwegian Polar Institute, Fram Centre, Tromsø, 2018).
44. Pascoal, F., Duarte, P., Assmy, P., Costa, R. & Magalhães, C. Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities. *Ann. Microbiol.* **74**, 1–12 (2024).
45. Kopf, A. et al. The ocean sampling day consortium. *GigaScience* **4**, 1–5 (2015).
46. Pascoal, F. et al. Inter-comparison of marine microbiome sampling protocols. *ISME Commun.* **3**, 1–16 (2023).
47. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108**, 4516–4522 (2011).
48. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
49. Apprill, A., Mcnally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **75**, 129–137 (2015).
50. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**, 1403–1414 (2016).
51. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
52. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
53. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
54. Yilmaz, P. et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
55. Keller-Costa, T. et al. Metagenomic insights into the taxonomy, function, and dysbiosis of prokaryotic communities in octocorals. *Microbiome* **9**, 1–21 (2021).
56. Arnan, X., Gaucherel, C. & Andersen, A. N. Dominance and species co-occurrence in highly diverse ant communities: a test of the interstitial hypothesis and discovery of a three-tiered competition cascade. *Oecologia* **166**, 783–794 (2011).
57. Calatayud, J. et al. Positive associations among rare species and their persistence in ecological assemblages. *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.9906092.v1.(2019)
58. Condit, R. et al. Complete data from the Barro Colorado 50-ha plot: 423617 trees, 35 years. Dryad. Dataset. https://doi.org/10.15146/5XCP-0D46 (2019).
59. Condit, R. et al. Beta-Diversity in Tropical Forest Trees. *Science* **295**, 666–669 (2002).
60. R. Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; (2023).
61. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York (2016).
62. Auguie, B. gridExtra: Miscellaneous Functions for 'Grid' Graphics. R package v2.3 (2017).
63. Oksanen, J. et al. Community Ecology Package. R Package Version 2.5-3 (2018).

64. Pascoal, F. pascoalf/Unsupervised-machine-learning-definition-of-the-microbial-rare-biosphere: v1.0.0; https://doi.org/10.5281/zenodo.14922332 (2025).
65. Thingstad, T. F., Vage, S., Storesund, J. E., Sandaa, R. A. & Giske, J. A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc. Natl Acad. Sci. USA* **111**, 7813–7818 (2014).
66. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).
67. Pester, M., Knorr, K. H., Friedrich, M. W., Wagner, M. & Loy, A. Sulfate-reducing microorganisms in wetlands - fameless actors in carbon cycling and climate change. *Front. Microbiol*. **3**, https://doi.org/10.3389/fmicb.2012.00072 (2012).
68. Krzywinski, M. & Altman, N. Visualizing samples with box plots. *Nat. Methods* **11**, 119–120 (2014).
69. Pascoal, F. pascoalf/ulrb. https://doi.org/10.5281/zenodo.14922442 (2025).

Acknowledgements

The Portuguese Science and Technology Foundation (FCT) funded this study through two grants to CM and FP (2022.02983.PTDC; PEX 2023.14123) and through a PhD grant to FP (2020.04453). This research has also been supported by FCT through the projects UIDB/04565/2020 and UIDP/04565/2020 of iBB, and UIDB/04423/2020 and UIDB/04565/2020 of CIIMAR, and the project LA/P/0140/2020 of i4HB. The work of Paula Branco was undertaken, in part, thanks to funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant. We acknowledge the contributions of Pedro Duarte and Philipp Assmy to ongoing support on the participation and organization of the expeditions for the MOSJ and N-ICE dataset. We also acknowledge logistic support from the Norwegian Polar Institute, regarding the MOSJ and N-ICE dataset. We also acknowledge the original authors of the publicly available datasets used: Richard Condit, Stephen Hubbell, et al (BCI); Xavier Arnan, Alan Anderson, et al. (Ants dataset); Tina Keller-Costa et al. (coral microbiome); and the Microbiome Ecology and Biogeochemistry team, from CIIMAR (MOSJ and N-ICE dataset).

Author contributions

Conceptualization (all authors); Data curation (F.P. and C.M.); Formal Analysis (F.P.); Funding acquisition (all authors); Methodology (F.P.); Resources (C.M.); Software (F.P., P.B., and L.T.); Supervision (L.T., P.B., R.C., and C.M.); Writing - original draft (F.P.); Writing - review and editing (all authors).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07912-4.

Correspondence and requests for materials should be addressed to Rodrigo Costa or Catarina Magalhães.

Peer review information *Communications Biology* thanks Maria Papadatou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Aylin Bircan, Christina Karlsson Rosenthal. [A peer review file is available.].