Vol. 88: 1–18, 2022
https://doi.org/10.3354/ame01979

AQUATIC MICROBIAL ECOLOGY
Aquat Microb Ecol

Published online January 13

Bacterioplankton composition as an indicator of environmental status: proof of principle using indicator value analysis of estuarine communities

Cecilia Alonso[1,*], Emiliano Pereira[1,2], Florencia Bertoglio[1], Miquel De Cáceres[3], Rudolf Amann[2]

[1]Microbial Ecology of Aquatic Systems Research Group, Universidad de la República, 27000 Rocha, Uruguay
[2]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany
[3]Joint Research Unit CTFC–AGROTECNIO, 25280 Solsona, Spain

ABSTRACT: Increasing awareness of environmental impacts caused by anthropogenic activities highlights the need to determine indicators of environmental status that can be routinely assessed at large spatial and temporal scales. Microbial communities comprise the greatest share of biological diversity on Earth and can rapidly reflect recent environmental changes while providing a record of past events. However, they have rarely been targeted in the search for ecological indicators of habitat types, environmental conditions, or environmental changes. Here, as a proof of principle, we analysed the bacterioplankton community composition of 4 estuaries in North and South America, Europe, and Asia, and looked for indicators of groups of samples defined using partition techniques, according to primary physicochemical variables typically monitored to infer water quality. Indicator value analysis (*IndVal*) was conducted to identify indicator operational taxonomic units (OTUs; analogous to species in other fields of ecology) in each group. These bacterioplankton-based indicators exhibited a high capacity to predict the group membership of the samples within each estuary and to correctly assign the samples to the appropriate estuary in a combined data set, employing different machine learning techniques. The indicators were composed of OTUs belonging to several bacterial phyla, which responded significantly and differentially to the environmental variables used to define the groups of samples. Moreover, the predictive values of these bacterial indicators were generally higher than those of other biological assemblages commonly used for environmental monitoring. Therefore, this approach appears to be a promising tool to complement existing strategies for monitoring and conservation of aquatic systems worldwide.

KEY WORDS: *IndVal* · Machine learning · Río de la Plata · Environmental management

1. INTRODUCTION

Monitoring the environmental impact of human activities is critical to preserve the existence, functioning, and ultimately the associated services provided by ecosystems. Thus, there is interest in finding indicators of environmental status that, while reflecting ecosystem complexity, remain relatively simple and easy to assess at large spatial and temporal monitoring scales (Dale & Beyeler 2001).

Microbial communities comprise the greatest share of biological diversity on Earth and have been shown to rapidly adjust their composition and/or functions in response to changing environments. In particular, significant correlations between aquatic microbial communities and environmental factors have been re-

[*]Corresponding author: calonso@cure.edu.uy

vealed when analysing natural gradients in key physicochemical variables such as salinity (Lozupone & Knight 2007), temperature (Fuhrman et al. 2008), pH (Fierer & Jackson 2006), inorganic nutrients and light availability (Schiaffino et al. 2011), and concentration and quality of dissolved organic matter (Amaral et al. 2016). Microbial communities have also been shown to respond significantly to specific disturbances like oil spills (Newton et al. 2013), contamination by heavy metals and polycyclic aromatic hydrocarbons (Sun et al. 2012), contamination by emergent contaminants (Subirats et al. 2018), and hydrological fragmentation (Fazi et al. 2013). Furthermore, bacterial communities have been used as biosensors to accurately distinguish unpolluted groundwater sites from those contaminated with uranium or nitrate across the watershed of an area contaminated with nuclear waste (Smith et al. 2015), to identify markers of urban impact (Fisher et al. 2015), and to construct indexes for assessing the ecological status of specific aquatic ecosystems (Lau et al. 2015, Aylagas et al. 2017, Li et al. 2018).

Thus, the accumulated evidence indicates that microbial communities are natural candidates for environmental monitoring. However, they have been out of the spotlight in this field, where animals and plants are the organisms typically chosen to act as environmental indicators (Siddig et al. 2016). The comparatively difficult methodology to approach microbial diversity and ecology undoubtedly has played a role discouraging their use by environmental practitioners. Even though microorganisms have long been used to assess faecal pollution using an array of culturable taxa (Clesceri et al. 1999, Meals et al. 2013), the use of molecular tools has greatly boosted the possibility of finding alternative indicators; e.g. suitable to discern sewage vs. animal sources, a fundamental advancement for the prevention and remediation of this kind of environmental impact (McLellan & Eren 2014, Roguet et al. 2018).

A key attribute of the widely employed high-throughput sequencing methodology is the opportunity to deeply evaluate microbial diversity in large numbers of samples. This allows for the application of microbial communities using similar approaches to the ones initially developed for macro-organisms. Among those approaches lie the tools for the identification of taxa indicators of environmental quality, which has been recently claimed by the scientific community in regards to the European Union Marine Strategy Framework Directive (Caruso et al. 2016, Pawlowski et al. 2018).

One of the indices most frequently employed to study the relationships between species and their habitats is the indicator value index (*IndVal*; Dufrene & Legendre 1997), which directly assesses the value of a species as a bioindicator of a specific type of habitat (De Cáceres & Legendre 2009). The application of *IndVal* to aquatic microbes has shown how bacterial community composition (BCC) changes with changing seasonal conditions across the Columbia River coastal margin (Fortunato et al. 2013), individual water column compartments in a meromictic lake (Gies et al. 2014), streams impacted by alkaline mine drainage (Bier et al. 2015), and different water masses of the eastern Mediterranean (Techtmann et al. 2015). This approach also identified indicator taxa of a threshold in oxygen concentration that leads to significant changes in community composition (Spietz et al. 2015). *IndVal* has also served to identify indicator archaeal lineages in a global survey of a wide range of aquatic habitats, highlighting their ecological importance and providing phylogeographical clues on the ecology and evolution of *Archaea* (Auguet et al. 2010).

Despite its proven suitability in conservation biology, *IndVal* remains relatively unexplored as a tool for environmental monitoring based on microbial communities. In particular, the application of *IndVal* to microbial communities has the potential for broader outcomes, like exploring their role as predictors of different habitat conditions, which has been rarely done (Xiong et al. 2014, Lanzén et al. 2021). To that end, effort is needed in terms of quantifying the performance (i.e. assessing the predictive capacity) of *IndVal*, analysing the practical challenges for its calculation while dealing with such diverse communities, and evaluating the suitability of its combination with common machine learning (ML) techniques for the classification of samples into environmental categories, using indicator species as predictors.

In this work, we analysed the bacterioplankton community composition of 4 different estuaries along with standard physicochemical metadata, either produced during this study or obtained from public repositories. Our goal was to search for combinations of bacterial operational taxonomic units (OTUs) as indicators of specific environmental conditions represented by groups of samples defined according to physicochemical variables typically measured while monitoring aquatic systems. These indicators should be able to assign each sample to its corresponding group, based only on the BCC data, acting as a proof of principle for the application of *IndVal* to bacterioplankton-based prediction of habitat categories. Additionally, we present and discuss methodological details for the application of *IndVal* to such diverse communities — including its combination with a suite of ML approaches — in order to further promote its utilization among microbial ecologists.

2. MATERIALS AND METHODS

2.1. Río de la Plata sample collection and environmental variables

Sub-superficial water samples (n = 30) were taken at 3 zones of the Río de la Plata (RdlP), located at the upper, middle, and lower parts of the estuary, during 5 sampling campaigns between March 2013 and March 2014 (Fig. 1). In each campaign, 2 samples were taken from each zone, differing in their distance to the coast (0.05–0.27 nautical miles, named 'coastal water' and 3.19–4.16 nautical miles, referred to as 'open water'). Exact site locations have been provided elsewhere (Martínez de la Escalera et al. 2017).

At each sampling point, conductivity (mS cm^{-1}), turbidity (nephelometric turbidity units [NTU]), and temperature (°C) were measured using a SeaBird 19 plus CTD profiler equipped with a turbidity sensor (SeaPoint Turbidity Meter). Dissolved oxygen (mg l^{-1}), salinity, total dissolved solids (TDS, g l^{-1}), density (sigma-t), and pH were measured with a Horiba multiparameter sensor. Total phosphorus (TP, µg l^{-1}), total nitrogen (TN, mg l^{-1}), phosphate (PO$_4$, µg l^{-1}), ammonium (NH$_4$, mg l^{-1}), reactive silica (Si, mg l^{-1}), and chlorophyll *a* (chl *a*, µg l^{-1}) were estimated following standard methods (Eaton et al. 1999, Martínez de la Escalera et al. 2017).

2.2. Bacterial DNA collection, extraction, and sequencing

Water samples were pre-filtered through a 23 µm sieve before collecting between 150 and 1000 ml sample^{-1} onto 0.22 µm mixed cellulose esters filters (Millipore). DNA extraction was carried according to a modified protocol of Zhou et al. (1996), previously applied in this system (Alonso et al. 2010). DNA was purified using QIAquick Gel Extraction Kit (QIAGEN), and its quality was checked using a standard PCR protocol for the 16S rRNA gene. The purified DNA was submitted to the company Research and Testing

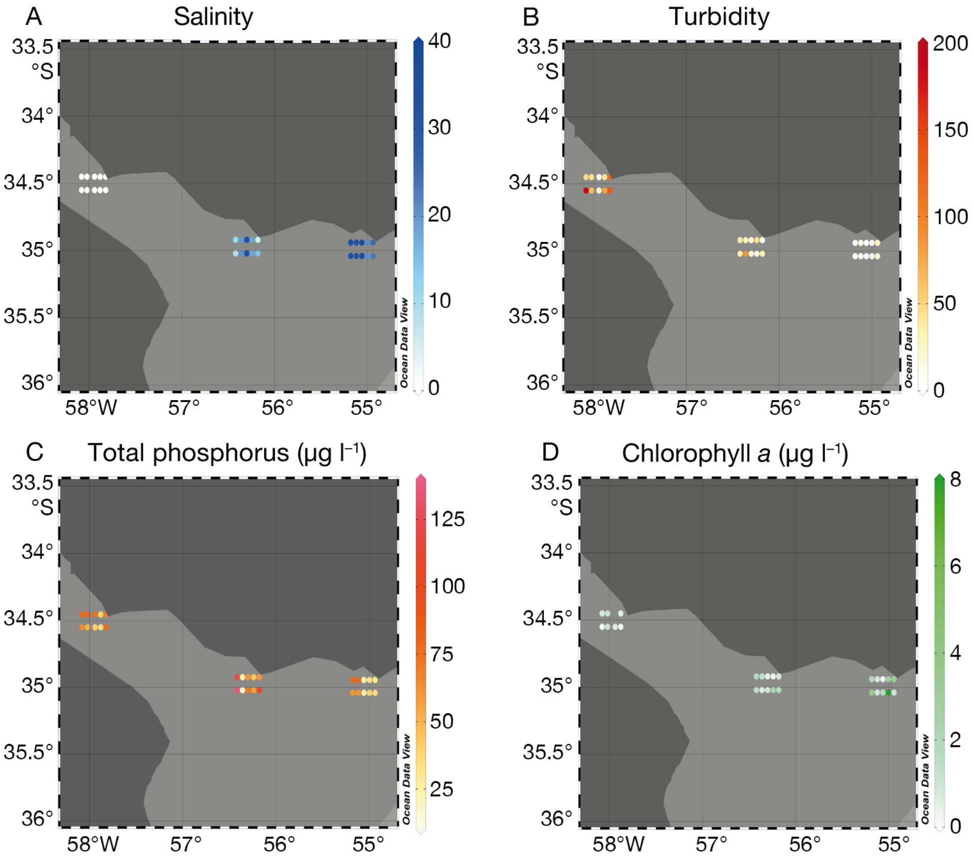


Fig. 1. The Río de la Plata estuary, showing sampling zones. Each zone was sampled 5 times, at 2 stations varying in distance to the coast. Different panels depict variation in some key environmental variables: (A) salinity, (B) turbidity, (C) total phosphorus, and (D) chlorophyll *a*. Data points in each zone are ordered according to the date of the campaign, from left to right

Laboratory for amplification using the primers 28F 5′-GAG TTT GAT CNT GGC TCA G-3′ and 519R 5′-GTN TTA CNG CGG CK-G CTG-3′ (V1–V3 region), and further sequencing with Illumina MiSeq. All samples, except for one (UppCW0613), were successfully sequenced. Sequences were deposited at the European Molecular Biology Laboratory (EMBL) (study accession code PRJEB29989).

2.3. BCC in other estuaries

A list of the main estuaries in South America, North America, Europe, Asia, Africa, and Oceania was compiled, and a search was performed using the terms 'bacterioplankton', '16S', and the name of each estuary. From the results retrieved, those studies having next generation sequencing (NGS)-derived sequences and accompanying physicochemical metadata were examined. Finally, a common data set was generated with studies encompassing the 16S rRNA gene V1–V3 region, accomplishing a sequence quality check (see Section 2.4.1) and sharing a minimum core of physicochemical data with our RdlP data set; those estuaries were the Delaware (Dlwr) (Campbell & Kirchman 2013, Kirchman et al. 2017), Krka (Korlević et al. 2016), and Pearl (Liu et al. 2015) (Table 1). For subsequent analyses, the database was curated, retaining the samples that had the complete set of physicochemical data and at least 1000 reads (see Table S1 in the Supplement at www.int-res.com/articles/suppl/a088p001_supp.pdf).

2.4. Bioinformatic analysis

2.4.1. Data pre-processing and integration

In this study, we integrated amplicon data of 16S rRNA genes generated with 2 different sequencing technologies: Illumina (RdlP data set) and Roche-454 (Dlwr, Krka, and Pearl data sets). Hence, the initial pre-processing step performed on the data differed accordingly. For the Illumina sequences, the pair-end reads were merged with the PEAR tool (Zhang et al. 2014), and for the 454 sequences, the adapter sequences from the single-end reads were removed with the 'cutadapt' tool (Martin 2011). Subsequently, in the 4 data sets, we quality trimmed to Q20 and filtered out reads shorter than 75 bp with the 'BBDuk' tool (http://jgi.doe.gov/data-and-tools/bb-tools), dereplicating the sequences and removing chimeras with the VSEARCH tool (Rognes et al. 2016).

Table 1. Main features of the data set, including basic information on the estuarine system along with the number of samples, number of sequences per sample, and accompanying physicochemical data. Physicochemical data retained for the definition of environmental groups within each estuary are highlighted in **bold**

Estuary name and code	Location	Drainage basin (km^2)	Number of samples	No. of sequences per sample (median)	Physicochemical data	Trophic state	References
Río de la Plata ('RdlP')	South America	3170000	30	14282	Depth, **Temperature, pH,** Conductivity, **Salinity,** Sigma T, **Turbidity, Oxygen, TDS, Chlorophyll a, TN, TP,** NO$_3$, **NH$_4$, PO$_4$,** SIO$_4$	Mesotrophic	This study
Delaware ('Dlwr')	North America	35066	128	2818	Depth, **Temperature, Salinity, Chlorophyll a, NO$_3$, NH$_4$, PO$_4$, SIO$_4$,** DOC	Eutrophic	Campbell & Kirchman (2013) Kirchman et al. (2017)
Krka ('Krka')	Europe	2088	28	6501	**Depth, Temperature, Salinity,** Sigma T, **Chlorophyll a, NO$_3$, NO$_2$, NH$_4$, PO$_4$,** SIO$_4$, **DOC**, POC	Oligotrophic	Korlević et al. (2016)
Pearl ('Pearl')	Asia	453700	17	3681	**Depth, Temperature**, pH, Salinity, **Turbidity, Oxygen**, Chlorophyll a, NO$_3$, **NO$_2$**, NH$_4$, PO$_4$	Eutrophic	Liu et al. (2015)

To integrate the data, the pre-processed sequences were aligned to the core reference alignment of GreenGenes (McDonald et al. 2012) using the PyNAST tool (Caporaso et al. 2010) to extract the V1–V3 region overlapping in all the data sets. PyNAST performs an alignment of the full amplicon sequences, allowing at the same time a straightforward visualization of the overlapping region in the multiple sequence alignment. To obtain the alignment coordinates for this task, we previously aligned the forward and reverse primer sequences used in the 4 data sets to the reference alignment. The 5' coordinate was then determined as the greatest end position of all forward primers, and the 3' coordinate, as the smallest start position of all reverse primers (i.e. 137 and 2227, respectively). All sequences that aligned with an identity lower than 70% were discarded.

2.4.2. Sequence clustering and taxonomic annotation

The sequences of all 4 pre-processed data sets were clustered into OTUs using a 97% identity threshold with the VSEARCH tool. We deliberately choose a relatively low identity threshold to counteract the different error profiles of Illumina and 454 sequencing technologies. However, to explore the possible outcomes using a more optimal identity threshold for defining bacterial species (Edgar 2018), the amplicon sequences generated in this study were also OTU-clustered with VSEARCH at 99% identity.

To taxonomically annotate the OTUs, we performed a local BLASTN search (Camacho et al. 2009) of the centroid sequences against the Silva 132 SSU Ref NR 99 database (Quast et al. 2013), using an e-value of 1e-4. To select the best hits, we used the same (empirically defined) criteria as applied in the SILVAngs pipeline (Quast et al. 2013), that is: (sequence identity + alignment coverage) / 2 ≥ 93. We formatted the final results as taxonomically annotated OTU abundance tables; from these tables, we filtered out the OTUs classified as mitochondria or chloroplasts and removed the singleton sequences.

All following analyses were conducted using R software (version 4.0.3) (R Core Team 2020).

2.5. Definition of groups of samples

For each estuary, a non-redundant set of standardized variables (*z*-score transformation) was determined based on their correlation coefficients and variance inflation factor calculation. The selected variables in each case are highlighted in bold characters in Table 1. Based on those variables, clustering and non-hierarchical partitioning of the samples was performed on each estuary data set using Euclidean distance and 2 different algorithms in each case (UPGMA and Ward for clustering, *K*-means and PAM for non-hierarchical partitioning). The optimal number of groups for each estuary was determined based on clustering consistency analyses (e.g. graph of fusion levels, silhouette analysis, Mantel analysis, calinski and ssi criteria; Legendre & Legendre 1998). A permutational multivariate analysis of variance (PERMANOVA) test was run to assess the statistical significance of each partition. The partition finally selected for each estuary was the one providing the highest number of significantly different groups, supported by at least 2 methods.

In addition, to test whether bacterial indicators could distinguish between estuaries, we created a common database, integrating the amplicon data of the bacterioplankton communities of the 4 estuaries, in which the samples were grouped according to the estuary of origin.

The R packages used for the variable standardization, clustering, partition, and PERMANOVA analyses were 'vegan' (Oksanen et al. 2012), 'cluster' (Oksanen et al. 2012), and 'pairwiseAdonis' (Martinez Arbizu 2020).

2.6. Search for indicator OTUs

The *IndVal* analysis was conducted to identify and select indicators for each sample group based on the OTUs abundance matrix (Dufrene & Legendre 1997), using the R package 'indicspecies' (De Cáceres & Legendre 2009). The *IndVal* index assesses the potential of a given species to act as an indicator for each target group. Its value varies between 0 and 1 and is the product of 2 components: component A (specificity or positive predictive value), which is maximum when the species is only present in the target group, and component B (sensitivity or fidelity), which is maximum when the species is present in all samples of the target group (Dufrene & Legendre 1997). Statistical significance of the *IndVal* statistic is tested using a permutation test (i.e. randomizing either the species vector or the group membership vector) (De Cáceres & Legendre 2009).

The *IndVal* framework has been extended to consider species combinations, in addition to single species, as potential indicators (De Cáceres et al. 2012). The large number of species in bacterial communities offers the opportunity to explore the indicator value of species combinations. Thus, we took the 500 or 1000 most abundant OTUs of each data set and, for each group to be indicated, we selected those OTUs with a frequency threshold value (B_t) of at least 0.5 (i.e. present in 50 % of samples in the target group) as candidates to form species combinations. Next, we used the function 'indicators' to assess the predictive value of combinations of up to 6 OTUs among those selected as candidates, as well as their statistical significance. The threshold for selecting the most abundant species as well as the maximal order of combinations (up to 6 OTUs) was chosen to keep the calculation time-effective, as there were many candidate species per group.

Based on this analysis, a list of potential indicators (either single OTUs or combinations thereof) was obtained for each group, from which the best indicator was selected as the one with the highest *IndVal*, utilising the 'pruneindicators' function in the R package 'indicspecies'. Whenever applicable, the 'pruning' function was also run, maximizing the positive predictive value (A) instead of maximizing the *IndVal* statistic. 'pruneindicators' selects indicators for which the lower bound of the 95 % confidence interval of *IndVal* (or one of its components) is equal to or greater than a user-defined threshold value. It keeps the set of non-nested indicators and selects the one with the highest *IndVal* (or one of its components) among those maximizing the coverage of the target group (De Cáceres 2013).

2.7. Prediction of sample membership based on bacterial indicators

The ability to predict the assignment of each sample to its corresponding group was tested using the 'predict' function of the 'indicspecies' package. This function takes into account the presence of the indicator in each sample in order to assign it to a given group. When the indicator is found in the sample, the probability of belonging to the target group is equal to its specificity (i.e. A, the positive predictive value of the indicator). For samples where 2 or more indicators are present, the probability of belonging to the target group is equal to the highest specificity value across all indicators found (De Cáceres 2013). Importantly, given a query sample, the 'predict' function independently estimates the probability of belonging to each group. Thus, the sum of probability values of a sample across all groups does not necessarily add up to 1.

For the purpose of this work, the 'predict' function was modified to include the leave-one-out cross-validation technique, in which the prediction of a given sample is done without taking into account the identification of indicators. This cross-validation tool is particularly useful for relatively small data sets (James et al. 2013), as is the case of the data sets of the different estuaries when analysed independently. Additionally, the function was also modified in order to use a training set and a test set. We applied this latter approach to evaluate the accuracy of the group predictions when using the combined data set, where each estuary was defined as a different group. Specifically, we used a random 80:20 partition to generate the training and test data sets (162 and 38 samples, respectively). Both modifications were developed in the context of this work and are now available in the current version (1.7.8) of the 'indicspecies' package.

Predictive performance of bacterial indicators for the combined data set was further explored using them as group predictors in diverse commonly used ML methods (C5.0 decision tree, linear discriminant analysis, neural network, random forest, support vector machine). Briefly, indicators exhibiting significant correlation coefficients >0.9 were removed, taking into account their importance in a preliminary classification test using the R package 'randomForest' (Liaw & Wiener 2002). Then, the random 80:20 data partition was taken, but instead of using the whole community as above, the matrices contained the abundance of every indicator in every sample. Each of the models was trained and the corresponding hyperparameters optimized, using the 'training set'. The performance of the final models was evaluated on the 'test set'. All analyses were run with the R Package 'caret' (Kuhn et al. 2016).

Finally, for the RdlP indicators, we assessed the importance of the different environmental variables in determining their abundance. Generalized linear models (GLMs) were run employing the negative binomial distribution to account for the overdispersion observed with the Poisson distribution (Faraway 2006). The environmental variables used were previously standardized in order to obtain comparable coefficients. The model choice was done using Akaike's information criterion (AIC), conducting a stepwise mixed selection procedure until the lowest possible AIC value was obtained. The R packages used for GLM were 'MASS' (Venables & Ripley 2002) and 'vegan' (Oksanen et al. 2012).

3. RESULTS

3.1. Río de la Plata estuary

3.1.1. BCC patterns

A total of 38 phyla were detected in the RdlP data set, with *Proteobacteria* dominating the community at all sites (Fig. 2A). *Acidobacteria*, *Actinobacteria*, and *Verrucomicrobia* were comparatively more abundant in the upper zone, whereas the contribution of *Cyanobacteria* and *Bacteroidetes* increased along the estuary (Fig. 2A). The 16 most abundant phyla (which accounted for >99% of the total abundance in all samples) were represented by 178 recognized families, of which the 22 most abundant represented 57% of the total abundance in all samples (Fig. 2B). Differential distribution of these families was also observed along the estuary, with SAR11 clade III (LD12) dominating in the

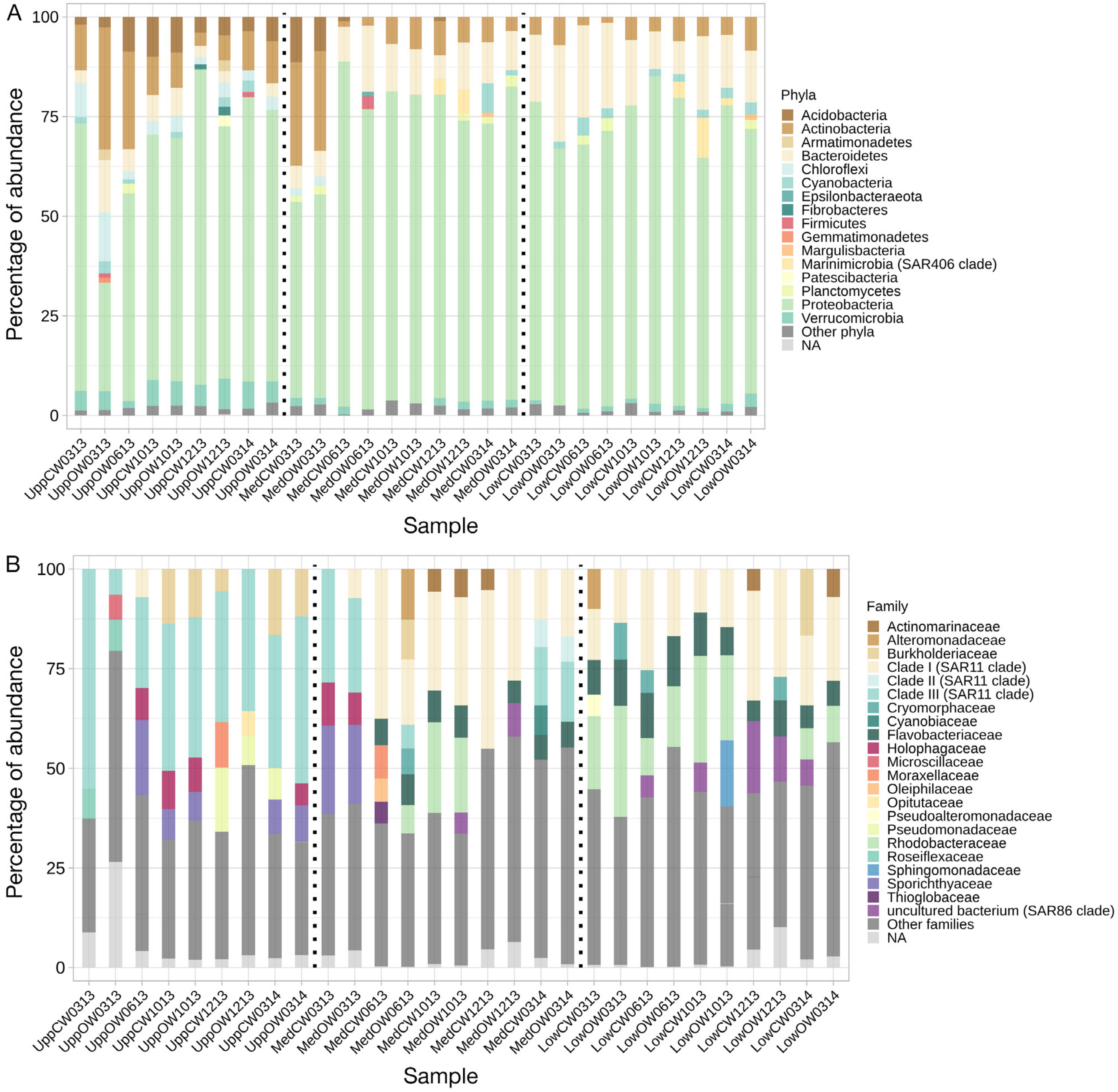


Fig. 2. Bacterial community composition of the different samples of the Río de la Plata estuary at the (A) phylum and (B) family level. Detailed proportions are given for the 16 most abundant bacterial phyla and the 22 most abundant bacterial families. The vertical dashed lines separate the samples taken in the 3 sections of the estuary (upper, medium and lower)

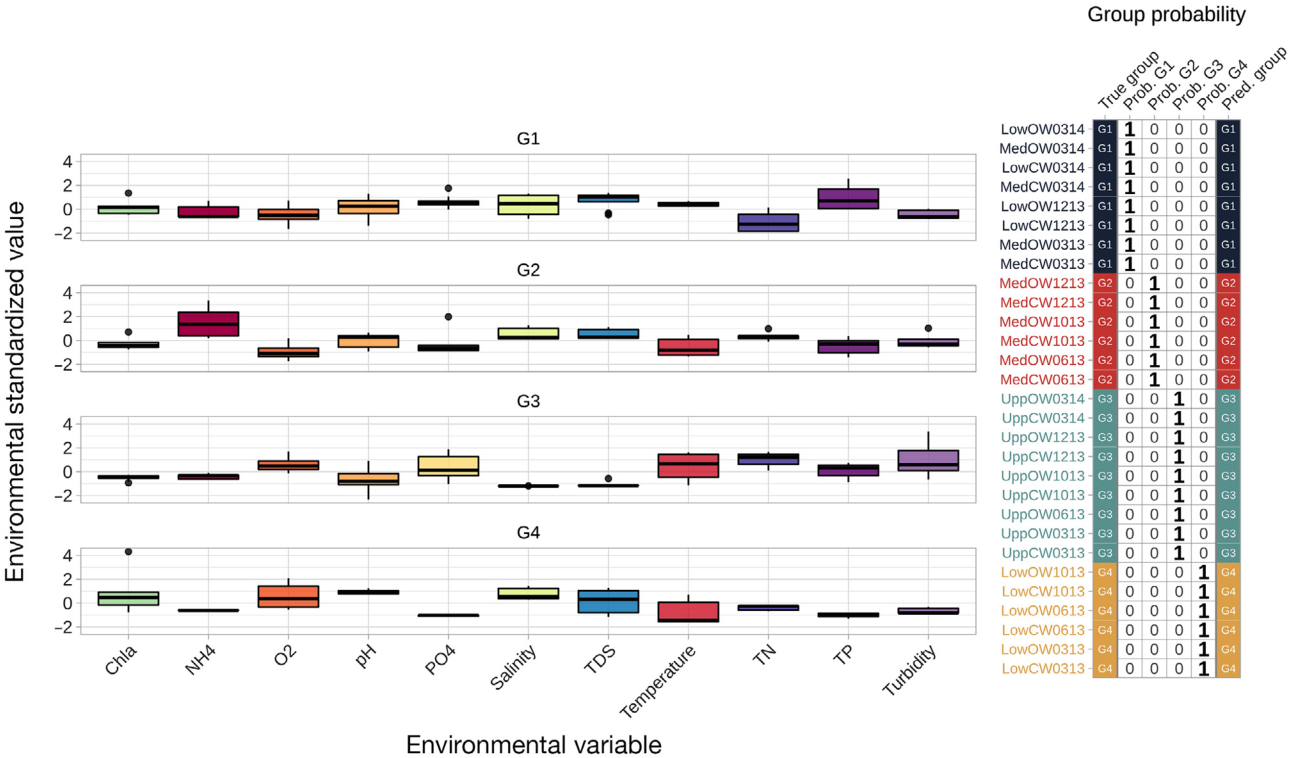
upper estuary and SAR11 clade I dominating the middle and lower estuary. Other families typically found in comparatively higher proportions at the freshwater end were *Burkholderiaceae*, *Holophagaceae*, *Sporichtychaceae*, and *Pseudomonadaceae* (Fig. 2B). Conversely, families relatively more abundant at the marine end were *Cyanobiaceae*, *Opitutaceae*, *Rhodobacteraceae*, *Flavobacteriaceae*, and SAR116 clade, while the proportion of *Thioglobaceae* members in the community increased in the middle and lower estuary (Fig. 2B).

3.1.2. Environmental clustering of samples

The grouping of samples reflected a strong spatial and (to a lesser extent) temporal variation (Fig. 3). G1 was the only group composed of samples from 2 different zones (medium and lower estuary), while the remaining groups were composed exclusively of samples from the medium (G2), upper (G3), or lower estuary (G4). Accordingly, group G3 was characterised by very low salinity values, high levels of TN and TP, and the lowest chl *a* concentrations. In contrast, G4 had the highest mean salinity values, the lowest mean levels of TN and TP, and the highest chl *a* values. Although G1 and G2 exhibited very similar mean salinity values, they mainly differed in turbidity and nutrient levels (Fig. 3).

3.1.3. Bacterial indicators

The indicators finally selected from the 97% similarity OTU clustering are summarized in Table 2. It was possible to find indicators for all groups exhibiting very high or even perfect *IndVal*. All indicators were based on the co-occurrence of 2 or 3 OTUs. In the case of G1 and G4, their indicators had 2 constituents, each of them given by the sum of 2–3 OTUs. Taxonomically, the indicators were composed of OTUs belonging to several of the most abundant phyla (*Actinobacteria*, *Bacteroidetes*, *Planctomycetes*, *Proteobacteria*, and *Verrucomicrobia*) and families (*Actinomarinaceae*, *Flavobacteriaceae*, *Alteromonadaceae*, *Burkholderiaceae*, *Halieaceae*, *Pseudohongiellaceae*, *Rhodobacteraceae*, SAR11 clades), although other less abundant families were also represented (Table 2).



	True group	Prob. G1	Prob. G2	Prob. G3	Prob. G4	Pred. group
LowOW0314	G1	1	0	0	0	G1
MedOW0314	G1	1	0	0	0	G1
LowCW0314	G1	1	0	0	0	G1
MedCW0314	G1	1	0	0	0	G1
LowOW1213	G1	1	0	0	0	G1
LowCW1213	G1	1	0	0	0	G1
MedOW0313	G1	1	0	0	0	G1
MedCW0313	G1	1	0	0	0	G1
MedOW1213	G2	0	1	0	0	G2
MedCW1213	G2	0	1	0	0	G2
MedOW1013	G2	0	1	0	0	G2
MedCW1013	G2	0	1	0	0	G2
MedOW0613	G2	0	1	0	0	G2
MedCW0613	G2	0	1	0	0	G2
UppOW0314	G3	0	0	1	0	G3
UppCW0314	G3	0	0	1	0	G3
UppOW1213	G3	0	0	1	0	G3
UppCW1213	G3	0	0	1	0	G3
UppOW1013	G3	0	0	1	0	G3
UppCW1013	G3	0	0	1	0	G3
UppOW0613	G3	0	0	1	0	G3
UppOW0313	G3	0	0	1	0	G3
UppCW0313	G3	0	0	1	0	G3
LowOW1013	G4	0	0	0	1	G4
LowCW1013	G4	0	0	0	1	G4
LowOW0613	G4	0	0	0	1	G4
LowCW0613	G4	0	0	0	1	G4
LowOW0313	G4	0	0	0	1	G4
LowCW0313	G4	0	0	0	1	G4

Fig. 3. Left panel: physicochemical variables characterising each of the 4 groups of the Río de la Plata estuary. Lower and upper hinges of the boxes: first and third quartiles; whiskers: minimum and maximum values without outliers; dots: outliers; horizontal lines: median values. Right panel: assignment to groups according to the bacterial indicators, using leave-one-out cross-validation

Table 2. Bacterial indicators of the 4 environmental groups defined for the Río de la Plata estuary. The candidate species are those who meet the criterion of being present in at least 50 % of the samples composing each group. A and B denote the *IndVal* components of specificity and sensitivity, respectively. The indicator operational taxonomic units (OTUs) are the combination of candidate OTUs that were selected as the best indicators of each group among all possible significant combinations ($p < 0.05$). All indicators achieved 100 % group coverage

Environmental group	No. of candidate species	*IndVal* A	*IndVal* B	Indicator OTUs and their taxonomic affiliation
G1	100	1	0.75	OTU1017 (*Methylomonaceae* Milano-WF1B-03) + OTU5016 (*Rubinisphaeraceae*)
		1	0.75	OTU12128 (*Ascidiaceihabitans*) + OTU3135 (NOR5 clade) + OTU5503 (*Pseudohongiella*)
G2	103	1	1	OTU1626 (SAR86) + OTU1236 (*Burkholderiaceae*, *Candidatus* Symbiobacter) + OTU10581 (*Candidatus* Actinomarina)
G3	128	1	1	OTU10228 (SAR11 clade III) + OTU3793 (*Verrucomicrobiae*)
G4	121	1	0.83	OTU12473 (*Rheinheimera*) + OTU1718 (NS4 marine group) + OTU11480 (SAR11 clade II)
		1	0.83	OTU13284 (*Rickettsiaceae*) + OTU10326 (SAR11 clade I)

Modelling the abundance of the indicators evidenced that they responded significantly to all of the environmental variables used for clustering the samples, although differing in the number and identity of variables to which each of them responded (Table 3). For all groups, at least one of the indicators was significantly influenced by salinity and temperature. TN, NH_4, and TDS were also key drivers for most of the indicators, whereas TP, oxygen, pH, and chl *a* were determinants for indicators of half of the groups.

Predictions of sample assignment to each one of the 4 groups based on these indicators are shown in Fig. 3. Bacterial indicators performed perfectly, assigning their target samples to the corresponding group using leave-one-out cross-validation. The same was true when using the indicators obtained with the OTUs defined at 99 % similarity (Tables S2 & S3).

Table 3. Summary of the generalized linear model coefficients of each environmental variable and their significance in explaining the abundance of the indicators of the different groups of sites. TDS: total dissolved solids; TP: total phosphorus; TN: total nitrogen. $^{***}p < 0.0001$; $^{**}p < 0.001$; $^{*}p < 0.01$; (.) $p < 0.05$; ns: not significant ($p > 0.05$)

	indG1a	indG1b	indG2	indG3	indG4a	indG4b
Salinity	-0.85^{***}	ns	1.51^{***}	-1.17^{**}	ns	3.78^{***}
Temperature	0.58^{*}	ns	-0.72^{***}	0.92^{***}	-1.29^{***}	ns
pH	ns	0.71^{**}	ns	-0.53^{*}	ns	ns
Turbidity	ns	ns	0.50^{*}	ns	ns	ns
Oxygen	ns	ns	-0.53^{**}	ns	-0.65^{**}	0.68^{*}
TDS	ns	-0.86^{**}	ns	0.53^{*}	ns	-1.06^{*}
NH_4	ns	0.40.	ns	-0.78^{**}	ns	-0.47^{*}
PO_4	0.47^{*}	ns	ns	ns	ns	ns
TP	ns	ns	0.37^{*}	ns	ns	-1.49^{*}
TN	ns	-1.38^{***}	ns	1.09^{**}	-0.68^{**}	ns
Chl *a*	0.70^{***}	ns	ns	−0.43.	ns	ns

3.2. Krka and Pearl estuaries

According to the physicochemical data, Krka estuary samples were divided into 3 significantly different groups (G1–G3). G1 included all deep samples, characterised by high salinity and low nutrients and chl *a* (Fig. S1). G2 exhibited the lowest mean depth along with the highest means for salinity, nutrients, and chl *a*. Finally, G3 was intermediate between the former groups but closer to G1 in terms of mean salinity and nutrient values.

Pearl estuary samples were also divided into 3 significantly different groups (G1–G3) based on the physicochemical variables. G1 was composed of samples taken at intermediate depths, characterised by very low oxygen concentrations while exhibiting the highest nutrient and turbidity values (Fig. S2). G2 included only subsurface samples along with intermediate nutrients and the highest oxygen concentration. Finally, G3 was composed of most of the samples taken at greater depths, exhibiting intermediate oxygen concentrations and lower nutrient values (Fig. S2).

Table 4 summarizes the indicators found for the Krka and Pearl estuaries. In both cases, it was possible to find bacterial indicators exhibiting the highest possible *IndVal* that were able to assign almost all samples to the correct group using leave-one-out cross-validation (Figs. S1 & S2). Here, representatives of widespread families were also the constituents of the different indicators (*Balneolaceae*, *Burkholderiaceae*, *Cyanobiaceae*, *Microbacte-*

Table 4. Bacterial indicators of the environmental groups defined for the Krka and Pearl estuaries. The candidate species are those who meet the criterion of being present in at least 50 % of the samples composing each group. A and B denote the *IndVal* components of specificity and sensitivity, respectively. The indicator operational taxonomic units (OTUs) are the combination of candidate OTUs that were selected as the best indicators of each group among all possible significant combinations ($p < 0.05$). All indicators achieved 100 % group coverage

Environmental group	No. of candidate species	*IndVal*		Indicator OTUs and their taxonomic affiliation
		A	B	
Krka				
G1	259	1	1	OTU7716 (SAR86) + OTU27236 (SAR11 clade Ia) + OTU26668 (*Rhodobacteraceae*)
G2	265	1	1	OTU12036 (*Candidatus* Aquiluna) + OTU2691 (*Burkholderiaceae* RS62 marine group) + OTU13774 (*Pseudohongiella*)
G3	410	0.96	1	OTU2076 (*Balneola*) + OTU3691 (*Oligoflexaceae*) + OTU5482 (SAR86)
Pearl				
G1	221	1	1	OTU_12048 (hgcI) + OTU_2127 (*Burkholderiaceae*)
G2	243	1	1	OTU_23055 (*Cyanobium* PCC-6307) + OTU_26114 (*Cyanobium* PCC-6307)
G3	200	1	1	OTU_23055 (*Cyanobium* PCC-6307) + OTU_13159 (SAR202)

riaceae, *Oligoflexaceae*, *Pseudohongiellaceae*, *Rhodobacteraceae*, *Sporichthyaceae*, SAR11 clades, and SAR86).

3.3. Delaware estuary

Delaware estuary samples split into 4 significantly different groups (G1–G4) according to strong spatial and temporal variations in the physicochemical variables. G1 and G3 were both composed of low-salinity samples, while G1 exhibited the highest mean temperature value and G3 had a very low average temperature. Furthermore, G1 exhibited higher means in chl *a* and PO_4 levels while G3 displayed higher average Si and NO_3 concentrations (Fig. S3). G2 was composed of samples with the highest salinity and the lowest nutrient concentrations. Finally, G4 was composed of samples with intermediate salinity and nutrient values and the lowest average temperature (Fig. S3).

Contrary to the other estuaries, no indicators exhibiting perfect *IndVal* were found for the Dlwr samples. Still, 92 % of the samples were correctly assigned to a group using leave-one-out cross-validation (Fig. S3). The notable exception was G3, for which no indicator was found with 100 % coverage of the group (Fig. S3, Table 5). Most of the indicator's components were members of different SAR11 clades (Table 5).

3.4. Indicators to identify samples in a combined data set

Initially, we attempted to search for indicators in one estuary and try to predict the group membership of the samples from another estuary. To perform such analysis, it is necessary to have environmental groups composed of samples from different estuaries. This way, it would be possible to assess whether the indicators obtained for the samples from a given estuary can be used to classify the samples from other estuaries. However, this exercise was not possible to perform given the characteristics of the data sets: while Dlwr and Pearl were the only pair not significantly distinguishable based on the physicochemical variables measured (Table S4), they harboured completely different communities (Fig. 4). On the contrary, while Dlwr and RdlP bore similar communities (Fig. 4), they were significantly distinct based on the physicochemical variables measured (Table S4).

Thus, the samples from the 4 estuaries were combined in a single data set, in which each estuary was defined as a different group, to test whether it was possible to identify bacterial indicators that would recognize their estuary of origin. Indicators for each group were identified and the assignment of the samples was performed, as before, using leave-one-out cross-validation and the 80:20 partition of samples, using the smaller set for testing. The indicators characteristic of each estuary are presented in Table 6.

Using leave-one-out cross-validation, we were able to assign 93 % of the samples to the correct estuary, except for a few Dlwr samples which were either not assigned to any group (9 samples) or wrongly assigned to RdlP estuary (5 samples) (Table S5). When the prediction was performed on the 'test set', the percentage of correctly assigned samples was 92–95 %, depending on whether the predictors were the indicators pruned in *IndVal* or its A component.

Table 5. Bacterial indicators of the 4 environmental groups defined for the Delaware estuary. The candidate species are those who meet the criterion of being present in at least 50% of the samples composing each group. A and B denote the *IndVal* components of specificity and sensitivity, respectively. The indicator operational taxonomic units (OTUs) are the combination of candidate OTUs that were selected as the best indicators of each group among all possible significant combinations ($p < 0.05$). All indicators achieved 100% group coverage, with exception of the G3 indicator which had 95.8% coverage

Environmental group	No. of candidate species	*IndVal* A	*IndVal* B	Indicator OTUs and their taxonomic affiliation
G1	87	0.99	0.54	OTU9092 (SAR11 clade II) + OTU6107 (hgcI)
		0.99	0.49	OTU10327 (SAR11 clade III) + OTU9372 (Cyanobium PCC-6307)
		0.97	0.49	OTU9241 (SAR11 clade III) + OTU12191 (SAR11 clade I) + OTU5103 (*Rubinisphaeraceae*)
		0.98	0.42	OTU10304 (*Cyanobium* PCC-6307) + OTU800 (*Nitrosomonadaceae*; IS-44) + OTU10580 (SAR11 clade II)
G2	145	0.90	0.82	OTU8635 (SAR11 clade I) + OTU11046 (*Ascidiaceihabitans*)
		0.87	0.36	OTU933 (MB11C04 marine group) + OTU10773 (*Synechococcus* CC9902) + OTU9765 (SAR11 clade I)
G3	110	0.84	0.71	OTU6051 (CL500-29) + OTU4920 (hgcI) + OTU4868 (hgcI)
		0.80	0.71	OTU9092 (SAR11 clade II) + OTU9241 (SAR11 clade III) + OTU6291 (hgcI)
		0.79	0.42	OTU10327 (SAR11 clade III) + OTU6265 (SAR324) + OTU5151 (*Pseudohongiella*)
G4	154	0.93	0.72	OTU9052 (SAR11 clade I) + OTU10059 (SAR11 clade Ia) + OTU9352 (SAR11 clade Ia) +
		0.94	0.68	OTU11164 (SAR116) + OTU1593 (NS3a marine group)
		0.94	0.68	OTU9241 (SAR11 clade III) + OTU13500 (SAR11 clade II) + OTU12434 (SAR11 clade III) + OTU9903 (SAR11 clade II) + OTU5657 (SAR324) + OTU875 (*Pseudohongiella*)
				OTU9052 (SAR11 clade I) + OTU10059 (SAR11 clade Ia) + OTU12763 (SAR11 clade III) + OTU12434 (SAR11 clade III) + OTU1687 (SAR86) + OTU9761 (SAR11 clade III)

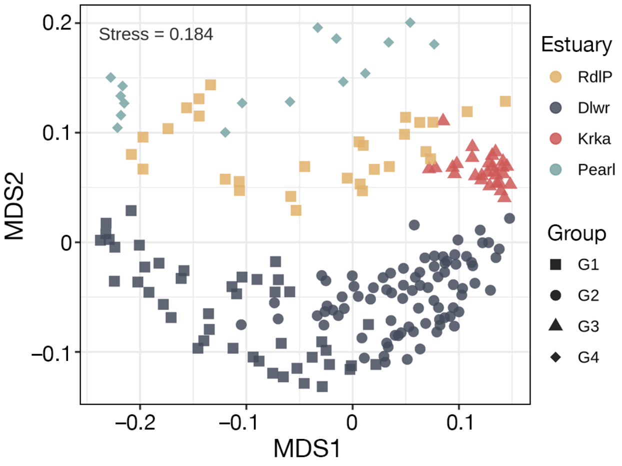


Fig. 4. Non-metric multidimensional scaling with groups given by Ward clustering superimposed, both based on the Bray-Curtis distance among samples, according to their bacterial community composition. The shapes of the symbols indicate the Ward clusters; symbol colours indicate the estuary to which each sample belongs (RdlP: Río de la Plata; Dlwr: Delaware)

This combined data set was further used to train and test different ML models using the indicators' abundance as predictor variables. Fig. 5 shows the accuracy and kappa values obtained by each model in the training data set, with neural networks displaying significantly larger values of both parameters. Nevertheless, for the 'test set', the accuracy of all 5 models was the same (95%), whereas the misclassified cases depended on the model used (Table 7). Table S6 shows the detailed performance of each model by group.

4. DISCUSSION

Estuaries are particularly appealing for evaluating the relationships between bacterial communities and their habitat as, in a relatively small spatial scale, they cover an ample range of physicochemical conditions found in aquatic systems. In the 4 estuaries analysed in this work, bacterial OTUs differed significantly in their distribution along the spatial gradient, due to environmental and geographical factors (Alonso et al. 2010, Campbell & Kirchman 2013, Liu et al. 2015, Korlević et al. 2016). Accordingly, it was possible to find bacterial OTUs to be used as indicators of each estuary in a combined database and of the different environmental groups defined within each estuary.

The development and use of ecological indicators relies on 2 key premises: (1) organisms effectively integrate different environmental variables of interest and (2) costs are lower than directly measuring the target variables (Niemi & McDonald 2004). In this

Table 6. Bacterial indicators of each estuary in the combined data set containing the bacterial community composition of all estuaries. The candidate species are those who meet the criterion of being present in at least 50% of the samples composing each estuary. A and B denote the *IndVal* components of specificity and sensitivity, respectively. The indicator operational taxonomic units (OTUs) are the combination of candidate OTUs that were selected as the best indicators of each group among all possible significant combinations ($p < 0.05$). The selection was done either choosing those OTUs combination with the highest *IndVal* (pruning in *IndVal*) or with the highest predictive value, that still guaranteed the maximal % of coverage (pruning in A). All indicators achieved 100% coverage, except for indicators for Delaware estuary which maximally achieved 96.1% of coverage

Estuary	Indicators (pruning in *IndVal*)	Indicators (pruning in A)
Río de la Plata	OTU_27433 (SAR11 clade III) + OTU_33567 (SAR11 clade III) OTU_21881 (SAR11 clade Ia) + OTU_26239 (SAR11 clade Ia) OTU_11399 (*Candidatus* Actinomarina) + OTU_40726 (SAR11 clade Ia)	OTU_27433 (SAR11 clade III) + OTU_33567 (SAR11 clade III) OTU_21881 (SAR11 clade Ia) + OTU_40726 (SAR11 clade Ia) + OTU_42727 (SAR11 clade Ia) OTU_21881 (SAR11 clade Ia) + OTU_5186 (SAR86) + OTU_40726 (SAR11 clade Ia)
Delaware	OTU_23224 (SAR11 clade Ia) + OTU_21881 (SAR11 clade Ia) OTU_23277 (SAR11 clade II) + OTU_24561 (SAR11 clade III) OTU_23224 (SAR11 clade Ia) + OTU_33237 (SAR11 clade Ia) OTU_24561 (SAR11 clade III) + OTU_27507 (SAR11 clade III)	OTU_21881 (SAR11 clade Ia) + OTU_33237 (SAR11 clade Ia) + OTU_32798 (SAR11 clade Ia) OTU_21881 (SAR11 clade Ia) + OTU_23277 (SAR11 clade II) + OTU_26847 (SAR11 clade Ia) + OTU_31962 (SAR11 clade Ia) OTU_23827 (SAR11 clade III) + OTU_30521 (SAR11 clade II)
Krka	OTU_23489 (*Synechococcus* CC9902) + OTU_32443 (SAR11 clade Ia) + OTU_6261 (SAR86)	OTU_23489 (Synechococcus CC9902) + OTU_32443 (SAR11 clade Ia) + OTU_35155 (SAR11 clade Ia) + OTU_6261 (SAR86) OTU_23489 (*Synechococcus* CC9902) + OTU_32443 (SAR11 clade Ia) + OTU_28045 (*Ascidiaceihabitans*) + OTU_6261 (SAR86)
Pearl	OTU_12048 (hgcI) + OTU_2950 (*Polynucleobacter*) OTU_21113 (SAR11 clade Ia) + OTU_11495 (*Candidatus* Actinomarina)	OTU_12048 (hgcI) + OTU_2950 (*Polynucleobacter* OTU_21113 (SAR11 clade Ia) + OTU_11495 (*Candidatus* Actinomarina)

work, we demonstrated that bacterial indicators responded significantly to the set of environmental variables used to define the groups of sites, with each indicator displaying a unique pattern of response (Table 3). While the environmental variables measured here are certainly cheaper to monitor than the bacterial indicators, this work aims to act as a proof of principle and a guide for further studies where the variables of interest are costly or cumbersome to measure. A potential application could be the identification of hotspots of the presence and concentration of certain pollutants. In particular cases, aquatic bacterial communities have been successfully used to categorize samples according to how they are impacted by different pollutants such as uranium, nitrate, hydrocarbons, heavy metals, or organic enrichment (Smith et al. 2015, Lanzén et al. 2021).

4.1. On the definition of groups of samples and *IndVal* calculation

Although it might seem obvious, it is worth stressing the importance of the variables chosen to define the groups, as different sets led to different groupings. This study aimed to evaluate the performance of bacterial indicators for categorizing samples into groups defined based on variables external to the BCC (i.e. all non-redundant environmental variables). Certainly, there are other possible partitions, depending on which variables are used for the definition of the groups.

Therefore, in order to correctly apply this approach, it is critical to (1) decide which is the focus of the grouping for later sample assignment to select the variables which are most relevant for the grouping and (2) assure a non-redundant set of variables to allow for optimal performance of the ordination/clustering methods chosen to define the groups (Legendre & Legendre 1998). The bacterial response to the variables of interest might be masked by the influence of main drivers of BCC, such as salinity or temperature (Fuhrman et al. 2008, Barberán & Casamayor 2010).

Although it is possible to define the groups of samples based on the BCC itself (Fortunato et al. 2013), when looking for indicators of environmental conditions it is advisable to

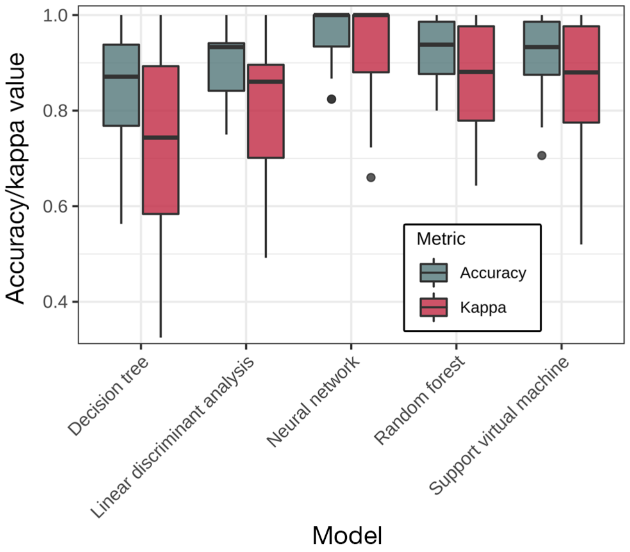


Fig. 5. Accuracy and Kappa values obtained by each model in the training data set. Boxplot parameters: lower and upper hinges: first and third quartiles; dots: outliers; horizontal lines: median values

Table 7. Confusion matrices for each machine learning method, showing the assignation of the samples in the test data set composed of 5 samples from Río de la Plata estuary (RdlP), 25 samples from Delaware estuary (Dlwr), 5 samples from Krka estuary, and 3 samples from Pearl estuary. RF: Random Forest; LDA: Linear Discriminant Analysis; NNet: Neural Network; SVM: Support Vector Machine; C5.0: Decision Tree

	Reference			
	RdlP	Dlwr	Krka	Pearl
RF				
RdlP	4	1	0	0
Dlwr	1	24	0	0
Krka	0	0	5	0
Pearl	0	0	0	3
LDA				
RdlP	5	0	0	0
Dlwr	0	25	1	1
Krka	0	0	4	0
Pearl	0	0	0	2
NNet				
RdlP	4	0	0	0
Dlwr	1	25	1	0
Krka	0	0	4	0
Pearl	0	0	0	3
SVM				
RdlP	5	1	0	0
Dlwr	0	24	1	0
Krka	0	0	4	0
Pearl	0	0	0	3
C5.0				
RdlP	4	0	0	0
Dlwr	1	25	0	1
Krka	0	0	5	0
Pearl	0	0	0	2

maintain statistical independence between the variables used for the definition of groups and the community composition (De Cáceres & Legendre 2009), as exemplified here.

IndVal has mainly been used for macro-organisms, which form communities that are typically far less diverse than microbial communities and for which the identification methods yield similar numbers of individuals examined per sample. This has practical consequences for the *IndVal* calculation using 'indicspecies' while dealing with microbial communities. To start with, the number of individuals sampled across sites can vary by orders of magnitude, not due to true sampling effort, but to the method used to obtain the identity of the microorganisms. Thus, although much information can potentially be lost, the samples need to be rarefied to the minimum sampling effort to obtain comparable abundances, at least within each group of sites. Rarefaction is a long-standing technique to render different sampling efforts comparable and remains a very robust tool compared to alternative approaches more recently developed for dealing specifically with bacterial communities (Weiss et al. 2017).

Another very distinctive issue of microbial communities is that there could be orders of magnitude difference in the abundance of the dominant and rare species. A remarkable advantage of the *IndVal* approach as a strategy for detecting differential abundance is that the indicator value of each species is evaluated independently of the others, and thus indicators are identified regardless of their abundance (Dufrene & Legendre 1997). This feature has the drawback that rare species can be selected as indicators, posing concern for whether their differential distribution is due to environmental filtering or under-sampling. However, this disadvantage can be easily overcome due to the very high diversity of microbial communities. This high diversity allows hundreds of candidate species to act as indicators of a given group. Although this is a notable advantage, it poses a computational challenge since a large number of possible combinations of OTUs will have to be evaluated. Thus, the number of candidate OTUs to consider per group needs to be reduced by applying different strategies. In particular, while using the 'indicspecies' package this can be achieved by setting a relatively stringent threshold to the B value and selecting a convenient maximum order of possible OTUs combinations (De Cáceres 2013).

A final methodological consideration concerns prediction; in this work, the 'indicspecies' package was modified in order to introduce 2 validation strategies.

We showed that the bacterial indicators were very effective in predicting the true group membership of virtually all samples of the 4 estuaries using leave-one-out cross-validation. The indicators also performed very well when utilised as predictors of a test set, generated by a random partition of samples, using the combined database.

As mentioned above, the 'predict' function of the 'indicspecies' package relies on the presence of the indicator — not on its abundance — restricting its performance. However, this work shows that provided the number of samples is sufficiently large, robust indicators can be selected using the *IndVal* approach and then their abundance utilised as predictors by employing a suite of ML techniques, overcoming that limitation. ML has formerly been used for classifying samples using the entire community composition (Smith et al. 2015, Lanzén et al. 2021) or known indicator taxa as predictors, i.e. faecal bacteria (Roguet et al. 2018). Microbial composition data sets are typically characterized by exhibiting a much larger number of OTUs compared to the number of samples (known as a high-dimensional problem). *IndVal* might be conveniently applied as a feature-selection methodology, allowing dimensionality reduction prior to the application of any ML technique. Feature selection allows for better control of collinearity, helping to deal with model overfitting (Hastie et al. 2009) and substantially decreasing computational time.

In this work, the predictions performed using the 'indicspecies' package already exhibited a very high degree of accuracy and, although the combination of *IndVal* + ML performed in the upper range of accuracy, there was not much room for improvement. However, while dealing with very large data sets and/or when the predictive capacity of individual indicators is not as powerful, the combination of preselecting the predictive variables using *IndVal* and then testing different ML models could be of particular utility. Thus, depending on the data set, it would be advisable to evaluate and compare the performance of both approaches in combination.

4.2. How good are bacterial indicators?

The *IndVal* for our bacterial indicators were relatively high (median value: 0.74; Tables 2, 4, & 5), in the upper range of values when compared to other biological communities to which this approach has been traditionally applied. Examples include plant indicators of peatland restoration with *IndVal* between 0.25 and 0.61 (González et al. 2013), spiders as indicators of heathland restoration (*IndVal*: 0.27–0.92) (Cristofoli et al. 2010), or multi-taxa (plants, moths, and songbirds) indicators of ecosystem recovery after reduction in deer density (*IndVal*: 0.50–0.97) (Bachand et al. 2014). Even though the environmental assessment of those works mainly evaluated restoration, which is broader than evaluating the response to a restricted set of physicochemical variables, the high *IndVal* shown by the microbial communities are noteworthy.

The high *IndVal* obtained here were mainly due to the specificity component (A), which was generally high, but they also exhibited high fidelity (meaning presence in all sites of their target group), especially compared to plant indicators (Bachand et al. 2014, Vieira et al. 2015). This high specificity might be a characteristic of microbial communities, due to their large population sizes and potential for long-distance dispersal (Finlay & Clarke 1999, Cohan & Koeppel 2008) combined with their high responsiveness to environmental conditions (de Wit & Bouvier 2006). Although more comparative studies are needed, the results obtained so far indicate that biofilm and benthic microbial indicators revealed by massive rRNA gene sequencing are equivalent to, or even outperform, metazoan indicators in the prediction of ecological status (Lau et al. 2015, Cordier et al. 2018).

4.3. Who are the bacterial indicators?

Indicator OTUs belonged to a handful of phyla (*Proteobacteria*, *Actinobacteria*, *Cyanobacteria*, *Bacteroidetes*, *Planctomycetes*, *Verrucomicrobia*, and *Chloroflexi*), probably reflecting the numerical dominance of these groups in aquatic systems (Barberán & Casamayor 2010). Also, within each phylum, indicator OTUs were mostly affiliated with certain taxa. Thus, *Alphaproteobacteria*, in particular SAR11, was the most common taxonomic affiliation of indicator OTUs, while SAR86, *Burkholderiaceae*, and *Pseudohongiella* stood out among the *Gammaproteobacteria*. Within the *Actinobacteria*, hgcI and *Actinomarinaceae* were the clades with indicator OTUs, while all cyanobacterial indicator OTUs belonged to 2 genera: *Cyanobium* and *Synechococcus*. This could merely derive from their numerical abundance in estuarine environments, and it could also be explained by a particularly strong ecological diversification of these groups. Aquatic *Alphaproteobacteria* (mainly SAR11) have been shown to exhibit very high levels of micro-diversity (Acinas et al. 2004, Ettema & Andersson 2009) attributable to environmental forcing (Logares et al. 2009). There is emerging evidence that this could be also the case for SAR86 (Hoarfrost et al. 2020) and hgcI (Neuenschwander et al. 2018).

The species composition of the indicators was in concordance with the known ecology of the respective clades. Salinity was clearly an important factor, easily linkable to the eco-physiological characteristics of several taxa. For example, members of the actinobacterial cluster hgcI, typically found in freshwater habitats (Warnecke et al. 2004), were frequently components of indicators for groups of samples characterized by low salinity, whereas high-salinity groups had among their indicators common members of marine communities, e.g. the proteobacterial clades SAR11 and SAR86, the NS4 *Flavobacteriaceae* cluster, or the *Chloroflexi* SAR 202 clade (Morris et al. 2002, Alonso et al. 2007, Schattenhofer et al. 2009, Mehrshad et al. 2018). The pattern of response to salinity was also evident in clusters from intermediate salinity where common freshwater representatives (hgcI and *Burkholderiaceae*) were indicators along with typical marine members (SAR11, SAR86, *Actinomarinaceae*, or the flavobacterial NS3a marine group).

Other eco-physiological features, although more subtle, were also evident: indicators for the groups exhibiting lower chl *a* values frequently included members known to thrive under less-rich conditions, such as SAR11, SAR202, SAR324, SAR116 (Schattenhofer et al. 2009, Choi et al. 2015, Cao et al. 2016, Mehrshad et al. 2018), while the opposite was true for indicators of the groups with high chl *a* values, which included members of *Rheinheimera*, NS4, *Cyanobium*, and *Planctomycetales*, which have frequently been found associated with phytoplankton blooms or high chl *a* (Brettar et al. 2006, Pizzetti et al. 2011, Díez-Vives et al. 2019, Li et al. 2020).

Thus, *IndVal* also appears to be a powerful tool to gain further insight into the ecology of different bacterial taxa. This aspect is remarkable, as we have only recently been able to evidence eco-physiological patterns of aquatic bacteria at the species level while dealing with whole community data sets (Osterholz et al. 2016, Chafee et al. 2018).

4.4. How do the identified bacterial indicators relate to water quality?

A meta-analysis of databases containing both community composition and appropriate environmental parameters would yield information on how general or specific the bacterial indicators are for the different types of aquatic systems, as well as the degree of impact to which they are subjected. This was one of the objectives of this work. However, it was surprising to see how little information is available in public repositories regarding bacterioplankton composition evaluated by NGS that is also accompanied by environmental metadata for estuaries worldwide. Moreover, a comparative analysis is further impeded by the utilisation of different regions of the 16S rRNA gene as sequencing targets. Thus, after an exhaustive search, only a handful of data sets could be included and the question of universality could not be addressed, as there was no true overlap among the environmental conditions measured in different estuaries (Fig. 4, Table S4). In particular, the apparent environmental similarity of Dlwr and Pearl suggested by the statistical tests probably resulted from the difference between the set of physicochemical variables measured in each study. Notably, while dissolved oxygen was the main driver of the BCC in the Pearl estuary (Liu et al. 2015), it was not reported in the Dlwr data set (Table 1).

Despite its limitations, this data set contained representatives from the full range of trophic states (Table 1), and the bacterial indicators were able to identify their estuary in the combined data set (Tables S5 & 7). Although the amount of data is not sufficient to establish a correlation among indicator taxa and trophic condition, the prevalence of different SAR11-related OTUs among the indicator taxa is outstanding, suggesting that the high niche differentiation within this clade could be a useful trait in the search for bioindicators. In this context, the selection of taxa associated with human activity is also remarkable; i.e. *Polynucleobacter* (Hosen et al. 2017) as an indicator of the Pearl estuary, which is the most impacted system in the data set.

The search for novel microbial indicators appears to be a promising tool for environmental management, complementing long-standing legislation standards typically restricted to bacterial indicators of faecal contamination with more integrative indicators of environmental quality. This could include further development of quality indexes based on BCC developed for specific habitats and/or anthropogenic impacts (Lau et al. 2015, Aylagas et al. 2017, Li et al. 2018) or the identification of non-culturable target taxa to be monitored through qPCR, in a similar approach to the detection of certain waterborne microbial pathogens (Clark et al. 2011, Eschbach et al. 2017). Moreover, the *IndVal* approach is also applicable beyond rRNA genes (Paula et al. 2014), opening up the possibility to search for indicators of a given process/pathway related to specific environmental conditions.

5. CONCLUSIONS

IndVal is an index that can be compared across very different biological communities as a measure of their suitability to reflect different environmental conditions. The results of this proof of principle study indicate that bacterioplankton species, as reported in 16S rRNA-based OTUs, appear to be promising candidates to consider in the context of monitoring and conservation of aquatic systems. The possibility of generating large databases harbouring environmental and functional information along with the microbial community composition data for several habitats is a good basis for developing and improving the environmental quality indicator potential of these diverse communities.

Acknowledgements. This work was developed in the framework of the MPI-UdelaR Partner group headed by C.A., in collaboration with R.A. Sequencing services were paid by a grant from CSIC provided to F.B., who held a postgraduate scholarship from ANII during the period of the project. Bioinformatic analyses were carried out during E.P.'s postdoctoral fellowship provided by ANII. Carla Kruk and Claudia Piccini are thanked for providing physicochemical data and assistance during the sampling campaigns. Marino Korlević is thanked for providing the environmental metadata of Korlević et al. (2016) in an open format. Dave Kirchman is thanked for sharing data of the Delaware estuary and for his critical and generous revision of the manuscript. The insightful comments of 2 anonymous reviewers were fundamental for the improvement of an earlier version of the manuscript.

LITERATURE CITED

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430: 551–554

Alonso C, Warnecke F, Amann R, Pernthaler J (2007) High local and global diversity of *Flavobacteria* in marine plankton. Environ Microbiol 9:1253–1266

Alonso C, Gómez-Pereira P, Ramette A, Ortega L, Fuchs BM, Amann R (2010) Multilevel analysis of the bacterial diversity along the environmental gradient Río de La Plata–South Atlantic Ocean. Aquat Microb Ecol 61: 57–72

Amaral V, Graeber D, Calliari D, Alonso C (2016) Strong linkages between DOM optical properties and main clades of aquatic bacteria. Limnol Oceanogr 61:906–918

Auguet JC, Barberan A, Casamayor EO (2010) Global ecological patterns in uncultured archaea. ISME J 4:182–190

Aylagas E, Borja Á, Tangherlini M, Dell'Anno A and others (2017) A bacterial community-based index to assess the ecological status of estuarine and coastal environments. Mar Pollut Bull 114:679–688

Bachand M, Pellerina S, Côtéa SD, Morettie M and others (2014) Species indicators of ecosystem recovery after reducing large herbivore density: comparing taxa and testing species combinations. Ecol Indic 38:12–19

Barberán A, Casamayor EO (2010) Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. Aquat Microb Ecol 59:1–10

Bier RL, Voss KA, Bernhardt ES (2015) Bacterial community responses to a gradient of alkaline mountaintop mine drainage in central Appalachian streams. ISME J 9: 1378–1390

Brettar I, Christen R, Höfle MG (2006) *Rheinheimera perlucida* sp. nov., a marine bacterium of the *Gammaproteobacteria* isolated from surface water of the central Baltic Sea. Int J Syst Evol Microbiol 56:2177–2183

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Campbell BJ, Kirchman DL (2013) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. ISME J 7:210–220

Cao H, Dong C, Bougouffa S, Li J and others (2016) Deltaproteobacterial SAR324 group in hydrothermal plumes on the south Mid-Atlantic Ridge. Sci Rep 6:22842

Caporaso JG, Bittinger K, Bushman TZ, DeSantis FD, Andersen GL, Knight R (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26:266–267

Caruso G, La Ferla R, Azzaro M, Zoppini A and others (2016) Microbial assemblages for environmental quality assessment: knowledge, gaps and usefulness in the European Marine Strategy Framework Directive. Crit Rev Microbiol 42:883–904

Chafee M, Fernàndez-Guerra A, Buttigieg L, Gerdts G, Eren AM, Teeling H, Amann RI (2018) Recurrent patterns of microdiversity in a temperate coastal marine environment. ISME J 12:237–252

Choi DH, Park KT, An SM, Lee K and others (2015) Pyrosequencing revealed Sar116 clade as Dominant *dddP*-containing bacteria in oligotrophic NW Pacific Ocean. PLOS ONE 10:e0116271

Clark ST, Gilbride KA, Mehrvar M, Laursen AE, Bostan V, Pushchak R, McCarthy LH (2011) Evaluation of low-copy genetic targets for waterborne bacterial pathogen detection via qPCR. Water Res 45:3378–3388

Clesceri LS, Greenberg AE, Eaton AD (1999) Standard methods for the examination of water and wastewater. American Public Health Association, Washington, DC

Cohan FM, Koeppel AF (2008) The origins of ecological diversity in prokaryotes. Curr Biol 18:R1024–R1034

Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J (2018) Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Mol Ecol Resour 18: 1381–1391

Cristofoli S, Mahy G, Kekenbosch R, Lambeets K (2010) Spider communities as evaluation tools for wet heathland restoration. Ecol Indic 10:773–780

Dale VH, Beyeler S (2001) Challenges in the development and use of ecological indicators. Ecol Indic 1:3–10

De Cáceres M (2013) How to use the indicspecies package (ver. 1.7.1). R Project 29. https://github.com/cran/indicspecies/blob/master/vignettes/indicspeciesTutorial.Rnw

De Cáceres M, Legendre P (2009) Associations between species and groups of sites: indices and statistical inference. Ecology 90:3566–3574

De Cáceres M, Legendre P, Wiser SK, Brotons L (2012) Using species combinations in indicator value analyses. Methods Ecol Evol 3:973–982

de Wit R, Bouvier T (2006) '*Everything is everywhere*, but, *the environment selects*'; What did Baas Becking and Beijerinck really say? Environ Microbiol 8:755–758

Díez-Vives C, Nielsen S, Sánchez P, Palenzuela O and others (2019) Delineation of ecologically distinct units of marine bacteroidetes in the Northwestern Mediterranean Sea. Mol Ecol 28:2846–2859

Dufrene M, Legendre P (1997) Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol Monogr 67:345–366

Edgar RC (2018) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375

Eschbach E, Martin A, Huhn J, Seidel C and others (2017) Detection of enteropathogenic *Vibrio parahaemolyticus*, *Vibrio cholerae* and *Vibrio vulnificus*: performance of real-time PCR kits in an interlaboratory study. Eur Food Res Technol 243:1335–1342

Ettema TJG, Andersson SGE (2009) The α-proteobacteria: the Darwin finches of the bacterial world. Biol Lett 5: 429–432

Faraway JJ (2006) Extending the linear model with R. In: Carlin BP, Chatfield C, Tanner M, Zidek J (eds) Texts in statistical science series. Chapman & Hall/CRC, Boca Raton, FL

Fazi S, Vázquez E, Casamayor EO, Amalfitano S, Butturini A (2013) Stream hydrological fragmentation drives bacterioplankton community composition. PLOS ONE 8:e64109

Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci USA 103:626–631

Finlay BJ, Clarke KJ (1999) Ubiquitous dispersal of microbial species. Nature 400:828

Fisher JC, Newton RJ, Dila DK, McLellan SL (2015) Urban microbial ecology of a freshwater estuary of Lake Michigan. Elementa 3:000064

Fortunato CS, Eiler A, Herfort L, Needoba JA, Peterson TD, Crump BC (2013) Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. ISME J 7:1899–1911

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH (2008) A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci USA 105:7774–7778

Gies EA, Konwar KM, Beatty JT, Hallam SJ (2014) Illuminating microbial dark matter in meromictic Sakinaw Lake. Appl Environ Microbiol 80:6807–6818

González E, Rochefort L, Boudreau S, Hugron S, Poulin M (2013) Can indicator species predict restoration outcomes early in the monitoring process? A case study with peatlands. Ecol Indic 32:232–238

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY

Hoarfrost A, Nayfach S, Ladau J, Yooseph S, Arnosti C, Dupont CL, Pollard KS (2020) Global ecotypes in the ubiquitous marine clade SAR86. ISME J 14:178–188

Hosen JD, Febria CM, Crump BC, Palmer MA (2017) Watershed urbanization linked to differences in stream bacterial community composition. Front Microbiol 8:1452

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer, New York, NY

Kirchman DL, Cottrell MT, DiTullio GR (2017) Shaping of bacterial community composition and diversity by phytoplankton and salinity in the Delaware Estuary, USA. Aquat Microb Ecol 78:93–106

Korlević M, Šupraha L, Ljubešić Z, Henderiks J, Ciglenečki I, Dautović J, Orlić S (2016) Bacterial diversity across a highly stratified ecosystem: a salt-wedge Mediterranean estuary. Syst Appl Microbiol 39:398–408

Kuhn M, Wing J, Weston S, Williams A and others (2016) Classification and regression training. R package version 6.0-71. https://github.com/topepo/caret/

Lanzén A, Mendibil I, Borja A, Alonso-Sáez L (2021) A microbial mandala for environmental monitoring: predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. Mol Ecol 30:2969–2987

Lau KEM, Washington VJ, Fan V, Neale MW, Lear G, Curran J, Lewis GD (2015) A novel bacterial community index to assess stream ecological health. Freshw Biol 60: 1988–2002

Legendre P, Legendre L (1998) Numerical ecology. Elsevier, Amsterdam

Li Y, Yang N, Qian B, Yang Z, Liu D, Niu L, Zhang W (2018) Development of a bacteria-based index of biotic integrity (Ba-IBI) for assessing ecological health of the Three Gorges Reservoir in different operation periods. Sci Total Environ 640-641:255–263

Li H, Barber M, Lu J, Goel R (2020) Microbial community successions and their dynamic functions during harmful cyanobacterial blooms in a freshwater lake. Water Res 185:116292

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

Liu J, Fu B, Yang H, Zhao M, He B, Zhang XH (2015) Phylogenetic shifts of bacterioplankton community composition along the Pearl Estuary: the potential impact of hypoxia and nutrients. Front Microbiol 6:64

Logares R, Braten J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K (2009) Infrequent marine–freshwater transitions in the microbial world. Trends Microbiol 17:414–422

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. Proc Natl Acad Sci USA 104:11436–11440

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17: 10–12

Martinez Arbizu P (2020) pairwiseadonis: pairwise multilevel comparison using adonis. R package version 0.4. https://github.com/pmartinezarbizu/pairwiseAdonis

Martínez de la Escalera G, Kruk C, Segura AM, Nogueira L, Alcántara I, Piccini C (2017) Dynamics of toxic genotypes of *Microcystis aeruginosa* complex (MAC) through a wide freshwater to marine environmental gradient. Harmful Algae 62:73–83

McDonald D, Price MN, Goodrich J, Nawrocki EP and others (2012) An improved GreenGenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618

McLellan SL, Eren AM (2014) Discovering new indicators of fecal pollution. Trends Microbiol 22:697–706

Meals DW, Harcum JB, Dressing SA (2013) Monitoring for microbial pathogens and indicators. Tech Notes 9. https://www.epa.gov/nps/nonpoint-source-monitoring-technotes

Mehrshad M, Rodriguez-Valera F, Amoozegar MA, López-García P, Ghai R (2018) The enigmatic SAR202 cluster up close: shedding light on a globally distributed dark ocean lineage involved in sulfur cycling. ISME J 12: 655–668

Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420:806–810

Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM (2018) Microdiversification in genome-streamlined ubiquitous freshwater actinobacteria. ISME J 12:185–198

Newton RJ, Huse SM, Morrison HG, Peake CS, Sogin ML, McLellan SL (2013) Shifts in the microbial community composition of Gulf Coast beaches following beach oiling. PLOS ONE 8:e74265

Niemi GJ, McDonald ME (2004) Application of ecological indicators. Annu Rev Ecol Evol Syst 35:89–111

Oksanen J, Blanchet FG, Kindt R, Legendre P and others (2012) vegan: community ecology package. R package version 2.0-2. https://github.com/vegandevs/vegan

Osterholz H, Singer G, Wemheuer B, Daniel R, Simon M, Niggemann J, Dittmar T (2016) Deciphering associations between dissolved organic molecules and bacterial communities in a pelagic marine system. ISME J 10: 1717–1730

Paula FS, Rodrigues JLM, Zhou J, Liyou W and others (2014) Land use change alters functional gene diversity, composition and abundance in Amazon forest soil microbial communities. Mol Ecol 23:2988–2999

Pawlowski J, Kelly-Quinn M, Altermatt F, Apothéloz-Perret-Gentil L and others (2018) The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Sci Total Environ 637-638:1295–1310

Pizzetti I, Fuchs BM, Gerdts G, Wichels A, Wiltshire KH, Amann R (2011) Temporal variability of coastal Planctomycetes Clades at Kabeltonne Station, North Sea. Appl Environ Microbiol 77:5009–5017

Quast C, Pruesse E, Yilmaz P, Gerken J and others (2013) The SILVA Ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584

Roguet A, Eren AM, Newton RJ, McLellan SL (2018) Fecal source identification using random forest. Microbiome 6: 185

Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J (2009) Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. Environ Microbiol 11:2078–2093

Schiaffino MR, Unrein F, Gasol JM, Massana R, Balagué V, Izaguirre I (2011) Bacterial community structure in a latitudinal gradient of lakes: the roles of spatial versus environmental factors. Freshw Biol 56:1973–1991

Siddig AAH, Ellison AM, Ochs A, Villar-Leeman C, Lau MK (2016) How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in Ecological Indicators. Ecol Indic 60:223–230

Smith MB, Rocha AM, Smillie CS, Olesen SW and others (2015) Natural bacterial communities serve as quantitative geochemical biosensors. MBio 6:e00326-15

Spietz RL, Williams CM, Rocap G, Horner-Devine MC (2015) A dissolved oxygen threshold for shifts in bacterial community structure in a seasonally hypoxic estuary. PLOS ONE 10:e0135731

Subirats J, Timoner X, Sánchez-Melsió A, Balcázar JL, Acunha V, Sabater S, Borrego C (2018) Emerging contaminants and nutrients synergistically affect the spread of class 1 integron-integrase (*intI1*) and *sul1* genes within stable streambed bacterial communities. Water Res 138: 77–85

Sun MY, Dafforn KA, Brown MV, Johnston EL (2012) Bacterial communities are sensitive indicators of contaminant stress. Mar Pollut Bull 64:1029–1038

Techtmann SM, Fortney JL, Ayers KA, Joyner DC, Linley TD, Pfiffner SM, Hazen TC (2015) The unique chemistry of eastern Mediterranean water masses selects for distinct microbial communities by depth. PLOS ONE 10: e0120605

Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer, New York, NY

Vieira LTA, Polisel RT, Ivanauskas NM, Shepherd GJ, Waechter JL, Yamamoto K, Martins FR (2015) Geographical patterns of terrestrial herbs: A new component in planning the conservation of the Brazilian Atlantic forest. Biodivers Conserv 24:2181–2198

Warnecke F, Amann R, Pernthaler J (2004) Actinobacterial 16S RRNA genes from freshwater habitats cluster in four distinct lineages. Environ Microbiol 6:242–253

Weiss S, Xu ZZ, Peddada S, Amir A and others (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27

Xiong J, Zhu J, Zhang D (2014) The application of bacterial indicator phylotypes to predict shrimp health status. Appl Microbiol Biotechnol 98:8291–8299

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. Appl Environ Microbiol 62: 316–322

Editorial responsibility: Ruben Sommaruga,
Innsbruck, Austria
Reviewed by: A. Lanzén and 1 anonymous referee

Submitted: December 14, 2020
Accepted: September 20. 2021
Proofs received from author(s): December 8, 2021