Environmental Science & Technology

Article

pubs.acs.org/est

Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning

Tristan Cordier,[*,†] Philippe Esling,[‡] Franck Lejzerowicz,[†] Joana Visco,[§] Amine Ouadahi,[†] Catarina Martins,[∥] Tomas Cedhagen,[⊥] and Jan Pawlowski[†,§]

[†]Department of Genetics and Evolution, University of Geneva, Boulevard d'Yvoy 4, CH 1205 Geneva, Switzerland

[‡]IRCAM, UMR 9912, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

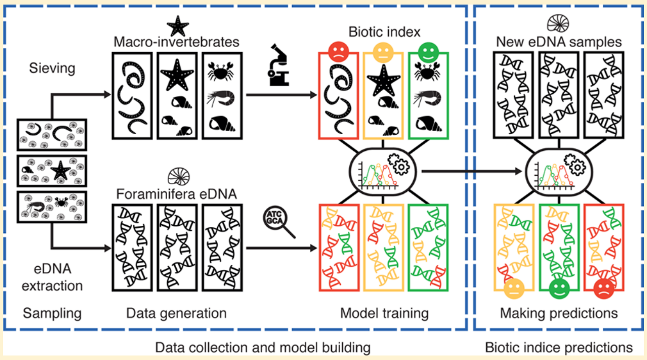
[§]ID-Gene ecodiagnostics, Ltd., chemin des Aulx 14, 1228 Plan-les-Ouates, Switzerland

[∥]Marine Harvest ASA, Sandviksboder 77AB, Bergen, 5035 Bergen, Norway

[⊥]Department of Bioscience, Section of Aquatic Biology, University of Aarhus, Building 1135, Ole Worms allé 1, DK-8000 Aarhus, Denmark

Supporting Information

ABSTRACT: Monitoring biodiversity is essential to assess the impacts of increasing anthropogenic activities in marine environments. Traditionally, marine biomonitoring involves the sorting and morphological identification of benthic macro-invertebrates, which is time-consuming and taxonomic-expertise demanding. High-throughput amplicon sequencing of environmental DNA (eDNA metabarcoding) represents a promising alternative for benthic monitoring. However, an important fraction of eDNA sequences remains unassigned or belong to taxa of unknown ecology, which prevent their use for assessing the ecological quality status. Here, we show that supervised machine learning (SML) can be used to build robust predictive models for benthic monitoring, regardless of the taxonomic assignment of eDNA sequences. We tested three SML approaches to assess the environmental impact of marine aquaculture using benthic foraminifera eDNA, a group of unicellular eukaryotes known to be good bioindicators, as features to infer macro-invertebrates based biotic indices. We found similar ecological status as obtained from macro-invertebrates inventories. We argue that SML approaches could overcome and even bypass the cost and time-demanding morpho-taxonomic approaches in future biomonitoring.



INTRODUCTION

Human activities are impacting the marine ecosystem functioning through climate change,[1] environmental pollution,[2] or human industry driven eutrophication,[3] and these pressures are likely to increase with the projected demographic expansion. Such impacts are traditionally monitored by surveying biodiversity, usually focusing on benthic macrofauna.[4,5] Biotic indices (BI) have been formalized to combine taxonomy and alpha-diversity measures (e.g., species richness, Shannon index) to reduce the data dimensions into single continuous values that are used to ascribe samples to an environmental quality status. Such indices include for instance the AZTI Marine Biotic Index[6] (AMBI) or more specific Indicator Species Index[7] (ISI), and the Norwegian Sensitivity and Quality Indices[8] (NSI and NQI1). The formulas of these indices include taxon-specific ecological weights or categories of tolerance to disturbance defined from empirical and experimental data.[6] Indeed, all indices currently applied in benthic monitoring are based on the sorting and identification of macro-invertebrate specimens, which is extremely time-consuming and taxonomic-expertise demanding. The need for faster and cost-effective ways to conduct biodiversity surveys is of prime importance to allow an effective management of marine resources.

High-throughput amplicon sequencing and the molecular identification of multiple species in environmental DNA (eDNA metabarcoding) offer a fast and cost-effective way to describe biological communities.[9] It gives the opportunity to overcome the limitations of morpho-taxonomy based biomonitoring.[10−12] Recently, several attempts have been made to use the eDNA metabarcoding for biomonitoring freshwater[13−18] and marine ecosystems.[19−23] Comparisons of BI inferred from eDNA metabarcoding and morphological data showed that similar values could be obtained in the case of freshwater diatoms[16,15] and marine invertebrates.[24,25] However,

Received: March 23, 2017
Revised: June 26, 2017
Accepted: June 30, 2017
Published: June 30, 2017

DOI: 10.1021/acs.est.7b01518
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

all these studies were dependent on reference sequence databases for taxonomic assignment, to retrieve the ecological weight associated with the assigned taxa (except for ref 18). This strongly limited the number of sequences included in the analysis, because a large proportion of sequences remained unassigned, or because they were assigned to taxa of unknown ecology.[20,24,26]

To overcome these limitations, we propose to use supervised machine learning (SML) for the development of predictive models to infer BI values from eDNA metabarcoding data without relying on taxonomic assignments. In SML approach, the predictive models are based on the knowledge extracted from a training data set, which typically consist of a set of features and associated labels (classification) or continuous values (regression). The aim of SML is to fit the training data to some function that can be used to predict a label or a continuous value to new input data[27] (e.g., new samples). In the recent years, there has been a growing interest to investigate the usefulness of SML for ecological,[28−30] genetic,[31] or microbiome analyses.[27,31−34] However, up to our knowledge, only one study has successfully used SML to predict pollution levels, as well as the values of 26 geochemical features based on a training data set composed of bacterial 16S eDNA metabarcoding data.[35]

In this study, we investigated the possibility of using SML to build predictive models for the inference of four biotic indices commonly used for the benthic monitoring of the fish farming industry. We tested this approach using benthic foraminifera, a group of ubiquitous unicellular eukaryotes that include sensitive bioindicators of pollution in marine environments.[36−39] Foraminifera are responsive to organic enrichment associated with fish farming activities, as shown by both morphological[40−43] and eDNA investigations.[22,23,44] This makes foraminifera particularly appropriate for SML approach, because of their small size, making them readily captured in eDNA samples, and because of the lack of well-defined indicator morphospecies.

MATERIAL AND METHODS

Sampling. A total of 144 sediment samples were collected in June and October 2015 at 24 stations distributed at the vicinity of five salmon farms in Norway (Table S1). Four farms were sampled at a rate of five stations and the remaining farm at only four stations because of pebbles at the fifth station. At each station, two van Veen grabs (1000 cm^2 model 12.211, KC-Denmark) were deployed. From each of the 48 grabs, three samples of ~5 mL were collected from the two first centimeters of the surface sediment using sterile spoons, transferred into 6 mL of LifeGuard Soil Preservation Solution (MO BIO, Laboratories Inc.) and frozen at −20 °C until DNA extraction. The rest of the sediment in each subsampled grab was sieved through 1 mm mesh and fixed in formalin for morpho-taxonomic inventory of macro-invertebrates. Macrofaunal data were used to calculate for each grab four BIs (AMBI, ISI, NSI, and NQI1) routinely used in benthic monitoring surveys in Norway.

DNA Extraction, PCR Amplification, and Sequencing. The 144 frozen sediments samples were thawed on ice, centrifuged at 2500 rpm for 5 min and the overlying solution was discarded. The total eDNA content was extracted using the PowerMax Soil DNA Isolation Kit (MO BIO Laboratories Inc.) following manufacturer instructions. The 37F hypervariable region of the SSU rRNA gene, commonly used as foraminiferal DNA barcode[45] was PCR amplified using the forward primer s14F1 (5′-AAGGGCACCACAAGAACGC-3′) and the reverse primer s15 (5′-CCACCTATCACAYAATCATG-3′), generating amplicons of about 200−250 bp as described previously.[22] The primers were tagged with 8-nt sequences appended at their 5′ ends to multiplex samples prior to sequencing library preparation. The tag sequences have been designed with a pairwise minimum edit distance of 2 and an edit distance of 8 to the corresponding positions of the conserved foraminiferal SSU rDNA sequence. We assigned the tag primer combinations to samples following a Latin Square Design (Table S2) to reduce mistagging events.[46] The PCR products were quantified by high-resolution capillary electrophoresis (QIAxcel System, Qiagen) and pooled in equimolar concentration. The pool was purified using the High Pure PCR Product Purification Kit (Roche), quantified using a fluorometric method (QuBit HS dsDNA kit, Invitrogen) and used for library preparation using the TruSeq DNA PCR-Free Library Prep Kit (Illumina). After quantification using the KAPA Library Quantification Kits (KAPA BIOSYSTEMS), the library was sequenced on an Illumina MiSeq System using MiSeq Reagent Kit v2 and a standard 14-tiles flow cell for 2*251 cycles. The raw data set is publicly available at the Sequence Read Archive under BioProject PRJNA376130.

Bioinformatics. The paired-end raw reads were quality filtered and assembled into full-length sequences with a pipeline written in C language for the fast processing of Illumina metabarcoding data (https://github.com/esling/illumina-pipeline). Filtering and assembly parameters are reported in Table S3. The pipeline included the demultiplexing of each sequence into its sample of origin by matching the combination of 8-nt tags present at the 5′-end of each sequence read. The tagged primers were trimmed from the sequences as well as the foraminifera-specific conserved region of ~70 nucleotides based on the detection of its GACAG motif after 60 positions. Every sequence lacking this motif was discarded. The data set was then filtered for potential chimeras using UCHIME,[47] version 4.2.40, implemented in the *identify_chimera_seq.py* function of the Qiime[48] 1.9.1 toolkit. We used the default parameters of the function, but the --split_by_sampleid option was used to restrict the de novo search by sample (i.e., by PCR). The filtered data set was then clustered into Operational Taxonomic Units (OTUs) using swarm[49] 2.1.8 with the default resolution ($d = 1$) and the fastidious option. This clustering also included homologous 37F sequences generated by previous fish farm sequencing surveys and obtained using identical bioinformatics processing, including sequences from Scotland,[22] from New Zealand,[23] and from Norway,[44] as well as unpublished data. The present data set was augmented with these sequences in order to increase the likelihood of the fastidious option implemented in swarm to form OTUs supported by the multiple occurrence of sequences specific to the fish farm environment. The representative sequences, that is, the most abundant Individual Sequence Unit (ISU) of each OTU, were used as input of the *assign_taxonomy.py* function of Qiime with default parameter for taxonomic assignment (uclust method), using a curated foraminifera database (http://forambarcoding.unige.ch). This database contains 1175 foraminiferal SSU rDNA sequences representing 125 morphospecies spread into 35 families, and including as well 106 environmental sequences for which morphospecies are not been yet identified but that are commonly found in eDNA samples. OTU-representative sequences that could not be assigned were compared by

BLAST[50] search against GenBank. The OTU-to-sample matrix was generated from the result of the clustering including all fish farms sequences with *make_otu_table.py* function of Qiime and the data corresponding to the samples analyzed in the present study was extracted from this table into the statistical R environment[51] for downstream statistical analysis.

Statistics. The relationship between diversity and distance from the cage as well as compositional variation of the studied communities were investigated based on normalized versions of the OTU-to-sample matrix. Because uneven sequencing depth across samples can introduce biases in the statistical analysis, samples with less than 10 000 reads were discarded. To investigate correlations between diversity and distance from the cage, 100 rarefied data sets were generated by aiming at 10 000 reads per sample using the *rrarefy* function of the R vegan v2.4-1 package.[52] Several alpha-diversity metrics, based on all OTUs (including unassigned ones) and their abundance, were then computed for each rarefied data set and averaged values were used to fit nonlinear polynomial models using the *lm* function in R. These diversity metrics included the OTU richness, the Shannon diversity,[53] the Pielou evenness,[54] the SN diversity used in the calculation of NQI1,[8] the expected OTU richness (ES100 and ES50 for respectively 100 and 50 reads) and the Chao richness estimator.[55] To investigate compositional variation, the read counts of the OTU-to-sample matrix were normalized according to the cumulative-sum scaling method (CSS) using the metagenomeSeq v1.16.0 R package[56] (see refs 57 and 58). The following beta-diversity analyses were performed using functions of the R vegan package. The Bray−Curtis dissimilarity index[59] was calculated using *vegdist* and the resulting pairwise dissimilarity matrix served for nonmetric multidimensional scaling (NMDS) analysis using *metaMDS* with default settings. The values of the BIs obtained from the morpho-taxonomic data and of distance to cages were fit to the NMDS ordination using *envfit*, and *ordisurf*, respectively.

Biotic indices values were predicted from the foraminiferal eDNA metabarcoding data using three supervised approaches. For reference, either the ecological weights associated with the taxa of the morpho-taxonomic inventories (correlation screening approach) or the BI values calculated from these inventories (supervised machine learning approaches) were considered. The BI values were predicted independently for the samples of each farm (testing data sets) based on the samples of the other four farms (training data sets) used for supervision in the three approaches. For the correlation screening approach, each OTU was compared to each morpho-taxon across the samples of the training set. The reads associated with each OTU in the PCR replicates of each grab were summed in order to measure Spearman rank correlations in terms of relative abundance across grabs. The ecological weight of the taxon that has the highest rho correlation above 0.7 was associated with the OTU. Biotic indices were then calculated for the testing data set using OTUs that were assigned ecological weights.

For the supervised machine learning approaches, either diversity metrics (diversity learning) or OTUs composition (composition learning) were used as features. Models were built for each training data set, and BI values were predicted for the samples composing each corresponding testing data set. For diversity learning, the predicted BI values were averaged for each sample over the 100 rarefied data sets. For composition learning, the effect of keeping rare OTUs was investigated by comparing the results obtained when the OTUs with less than 10 reads or less than 100 reads across the full data set were discarded before CSS normalization. Since only one reference BI value was available for each grab, the BI values predicted using both learning methods were averaged for each grab and the standard deviations were computed. For both supervised diversity and composition learning approach, we compared two different algorithms to predict BI values. We compared the Random Forest (RF) algorithm[60] implemented in the ranger v.0.6.0 R package[61] for multithreading and the Self-Organizing Map (SOM) algorithm implemented in the Kohonen v2.0.19 R package.[62] For the RF algorithm, we generated 300 trees and used the default "mtry" parameter for regression task, which is 1/3 of the features randomly picked to split the tree at each node, which usually give the best results.[63] RF models were measuring the importance of features in the prediction of BI values. For the SOM algorithm, the network was trained with 100 iterations with the default learning rate (linear decline from 0.05 to 0.01 over the 100 iterations). The tuning of the parameters (xdim, ydim, xweight, topo) were randomly searched over 100 parameters combinations (100 hyper-parameter search) with the *train* function of the caret v6.0-73 R package,[64] passing the "random" search option and a 10-fold cross validation repeated 10 times to the *trainControl* function. For each training data set (four farms), the set of parameters giving the lowest average Root Mean Squared Error (RMSE) on hold-out samples during cross-validation were used to build the model that predict BI values for the testing data set (the remaining farm).

To compare the accuracy of the supervised approaches, the relationships between the reference and predicted BI values were modeled using the *lm* function in R. These BI values were then converted into discrete ecological quality status, after averaging per grab in the case of the predicted values. Their agreement was tested using the *kappa2* function of the irr v0.84 R package,[65] with squared weight because the ecological status values are ordered from "very poor" to "very good". Agreement between the two classifications was considered as "poor agreement", for example, Kappa value ranging from 0.01 to 0.2 to "almost perfect agreement", for example, Kappa value ranging from 0.8 to 1.[66] For each BI, the best model was the one associated with the highest Kappa value. Its accuracy was investigated at the scale of the farm by testing for difference between predicted and reference values using the Mann−Whitney test with the *wilcox.test* in R.

■ RESULTS

Macro-Invertebrate Morpho-Taxonomic Inventories and Biotic Indices. The morpho-taxonomic inventories of 75,431 macro-invertebrate specimens comprised 432 morpho-species, of which 357 have been ascribed to an ecological category in at least one of the BIs (Table S4). All BIs showed that the five farms were impacted, with highest values (AMBI) and lowest values (ISI, NSI and NQI1) within 200 m from the fish cages (Figure S1). The impact decreased quickly with increasing distance from the cages. According to the results of nonlinear models, the effect of the distance from the cage on the values of every BI was highly significant (Table S5 and Figure S1).

Foraminifera eDNA Metabarcoding Data. PCR products were obtained for 123 out of 144 samples and sequencing yielded 11 257 700 paired-end reads. The quality filtering (i.e., base call quality and contig assembly) discarded 24.9% of the reads (Table S3) and de novo chimera filtering discarded a

DOI: 10.1021/acs.est.7b01518

Table 1. Accuracy of Biotic Indices Predictions Obtained from Correlation Screening, Diversity Learning, and Composition Learning, with the Random Forest and Self-Organizing Map Algorithms[a]

biotic index	approach	supervised learning algorithm	R^2	kappa
AMBI	correlation screening		0.568***	0.624***
	diversity learning	Random Forest	0.641***	0.69***
		Self-Organizing Map	0.492***	0.618***
	composition learning	Random Forest	0.662***	0.555***
		Self-Organizing Map	**0.669*****	**0.711*****
ISI	correlation screening		0.65***	0.53***
	diversity learning	Random Forest	0.505***	0.626***
		Self-Organizing Map	0.449***	0.61***
	composition learning	Random Forest	0.56***	0.631***
		Self-Organizing Map	**0.615*****	**0.774*****
NSI	correlation screening		0.508***	0.607***
	diversity learning	**Random Forest**	**0.83*****	**0.907*****
		Self-Organizing Map	0.83***	0.88***
	composition learning	Random Forest	0.827***	0.832***
		Self-Organizing Map	0.794***	0.871***
NQI1	correlation screening		0.76***	0.8***
	diversity learning	**Random Forest**	**0.834*****	**0.88*****
		Self-Organizing Map	0.805***	0.846***
	composition learning	Random Forest	0.81***	0.856***
		Self-Organizing Map	0.803***	0.873***

[a] Best predictive models are in bold (i.e., highest kappa statistics). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

further 449 sequences. The resulting 8 449 933 sequences were clustered along with 45 588 115 homologous sequences generated in previous studies and 69 716 OTUs were delineated. An OTU-to-sample matrix was formed from the 9235 OTUs that occurred in at least one of the 123 samples corresponding to the five studied fish farms. The sequencing depth of the samples was uneven. It ranged from 42 to 283 431 reads, with an average of 68 698 reads per sample. Seven samples represented by less than 10 000 reads were discarded. The final OTU-to-sample matrix was constituted of 116 sediment samples as rows, covering the 48 sediment grabs from which the macro-invertebrates were inventoried, 9170 OTUs as columns, and was filled with the 8 414 122 reads (Table S6). Two additional OTU-to-sample matrices were then composed from the OTUs represented by more than 10 reads and by more than 100 reads, and contained 3648 and 1579 columns, respectively.

Nonlinear models yielded significant correlations between alpha-diversity metrics and the distance from the cage, except Pielou's evenness (Figure S2). The NMDS analysis of the beta-diversity matrix showed that the distance from the cage is strongly structuring foraminifera communities, and that the reference BI values inferred from the macro-invertebrate data were strongly correlated with the ordination (Figure S3).

Taxonomic Composition. Among the 9 170 OTUs, 2,378 (representing 70% of the reads) were assigned to a given taxonomic rank based on the consensus of three maximum hits with a minimum of 90% similarity and coverage on reference sequences of our curated foraminifera SSU rDNA database. Among these assigned OTUs, the majority belong to orders Rotaliida and Textulariida and class "Monothalamea" (Table S7). Sixty-two OTUs (less than 0.1% of the reads) were assigned to one of the remaining orders (Miliolida, Globerinida, Spirillinida and Robertinida) and 190 OTUs (4.3% of the reads) matched uncultured foraminifera sequences of GenBank with more than 90% of similarity and coverage. The remaining 6,602 OTUs (25.7% of the reads) could not be assigned to any reference sequence in foraminiferal database and matched no GenBank sequence. The five most abundant taxa include 3 rotaliids: *Bulimina marginata* (37 OTUs, representing 10.2% of the reads), *Stainforthia fusiformis* (21 OTUs, representing 7.7% of the reads), and *Cibicidoides lobatulus* (39 OTUs, representing 7% of the reads), a monothalamid: *Bathysiphon argenteus* (18 OTUs, representing 7.4% of the reads) and a textularid genus: *Reophax sp.* (105 OTUs, representing 5.7% of the reads).

Predictions of BI Values from eDNA Metabarcoding Data. The three supervised approaches yielded accurate BI values predictions (Table 1). Linear models and Kappa tests showed significance for the correlation screening approach to predict the BI values from ecological weights assignments. Yet, R^2 and Kappa statistics were higher for the diversity and composition learning approaches than for the correlation screening approach, as measured for ISI, NSI and NQI1 (Table 1 and Figure 1). For AMBI, the correlation screening approach performed better than the diversity learning using SOM algorithm and the composition learning using the RF algorithm.

The diversity learning approach using the RF algorithm yielded predictions that had the highest Kappa for NSI and NQI1, with 33 and 37 correctly classified grabs and with 15 and 11 grabs classified within 1 category mismatch, respectively. The Kappa statistics were 0.907 for NSI and 0.88 for NQI1, indicating an almost perfect agreement between molecular and morpho-taxonomic data. Rarefying the OTU-to-sample data set had a small effect on the per-sample variation of the inferred values for both NSI (Figure S4) and NQI1 (Figure S5). The most important diversity metrics (as measured by the RF algorithm) to infer NSI and NQI1 were Chao and OTU richness, and to a lesser extent SN (Figures S6 and S7).

The composition learning approach using the SOM algorithm yielded predictions that had the highest Kappa for AMBI and ISI, with respectively 33 and 25 correctly classified grabs, 13 and 21 classified within 1 category mismatch, and 2 misclassified for both BIs. The Kappa statistics were 0.711 for

DOI: 10.1021/acs.est.7b01518

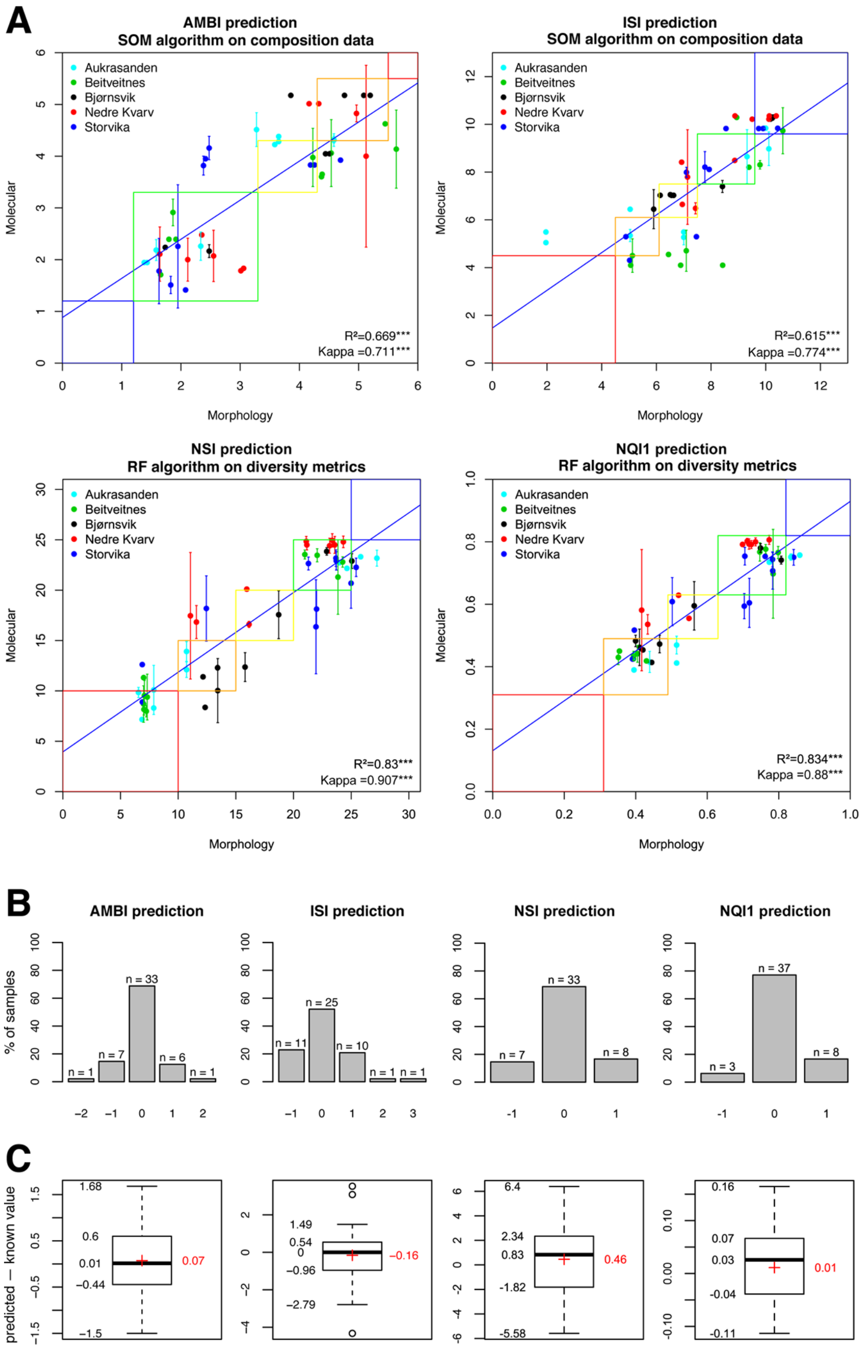


Figure 1. (A) Correlation between the BI values computed from the morpho-taxonomic inventories (morphologic data) and the one predicted from supervised machine learning from the foraminifera SSU 37F data (molecular data). The plots are obtained from the predictive models giving the best accuracy (i.e., highest Kappa statistic, see Table 1). AMBI and ISI predictions were done using the SOM algorithm on composition data. NSI and NQI1 predictions were done using the RF algorithm on diversity metrics. Each dot and error bars represent the average and standard deviation of the predicted BI values for the sediment sample of a grab. (B) Barplots indicate the amount (number over bars) and percentage (y axis) of correct classifications (0 on x axis) and misclassifications. (C) Boxplots indicates the median and quartile values (black numbers) and the mean value (red number) of the difference between the predicted and the reference BI values.

AMBI and 0.774 for ISI, indicating a substantial agreement between molecular and morpho-taxonomic data. The RF algorithm measurements of features importance showed that the OTUs that were most important for inferring AMBI and ISI were not necessarily the most abundant in terms of reads counts (Figure S8 and S9). The most useful OTU to infer AMBI values was unassigned and represented 0.7% of the reads (55 063 reads) and that to infer ISI values was an unidentified Monothalamid representing 0.2% of the reads (19 123 reads). Discarding rare OTUs represented by less than 10 or 100 reads did not significantly change the accuracy of BIs predictions using composition learning (Table S8).

Predicted values obtained with the best predictive models for each of the four BIs were neither over nor underestimated (Figure 1). Grabs were evenly distributed around the correct classification (Figure 1B). The median and the mean of the differences between the predicted and reference BI values were close to zero and with moderate variances, which means that our predictive models were not biased (Figure 1C). Mann−Whitney tests of difference between the predicted and reference BI values were not significant for all fish farm and BI combinations, which means that predictive models were accurately predicting BI values at the scale of the fish farm (Figure S10).

DISCUSSION

Our study demonstrates the usefulness of supervised machine learning (SML) approaches to infer biotic indices from eDNA metabarcoding data. To our knowledge, this is the first time that the SML approaches have been applied to eukaryotes eDNA-based biomonitoring surveys (but see35 for a bacteria-based survey using SML). Until now, only those groups of eukaryotes for which DNA barcodes are available in reference sequence database and for which an autecological value is known could be taken into account in eDNA-based biomonitoring studies.[24,25] Provided that an appropriate training data set is available, SML approaches offer a workaround to the dependency on reference databases and thus allow extending the range of potential genetic bioindicators to other taxonomic groups, especially to the small-sized inconspicuous taxa, which typically dominate the eDNA samples.[24] With the notable exception of diatoms,[67] most of these small-sized taxa remain poorly described in terms of autecology (but see68), which make them useless for computing biotic indices. Using SML approaches, there is no need for prior knowledge on the ecological signal conveyed by OTUs, because these signals are inferred during the statistical modeling.

Our results showed that accurate predictive models for BIs inference could be obtained with diversity metrics or composition data derived from eDNA metabarcoding data as features in SML approaches. These models led to similar bioassessments as the ones obtained using traditional morphology-based macrofaunal surveys. The possibility of using diversity metrics or composition data for making BIs predictions on new samples may give a practical flexibility for biomonitoring surveys. On the one hand, using the diversity metrics in SML reduces the dimensions of the data and therefore the computation time, and allows the prediction of BIs for samples coming from various geographical regions, where the taxonomic composition may be different. On the other hand, using composition data to predict BIs seems more appropriate if a significant proportion of OTUs are present in both the training and the testing data set. Composition data should capture species interactions[69,70] and give more importance to key taxa consistently responding to specific environmental variation,[71,72] which could yield better results for atypical samples or statistical outliers.

While analyzing eDNA metabarcoding data, various biological and technical issues need to be considered for the interpretation of alpha-diversity and beta-diversity. The presence of extracellular DNA necessarily blurs the ecological signal conveyed in eDNA data sets as living, dead and inactive cells cannot be readily distinguished.[73,74] In our study, samples were collected from surface sediments, where it is reasonable to expect that most of the DNA come from living or recently dead organisms. Intragenomic rRNA sequence variation[75] and rRNA gene copy-number variation[76] brings qualitative and quantitative bias in metabarcoding data. From a technical point of view, PCR, sequencing[77] and mistagging errors[46] could add further noise in the data. However, this noise can be assumed to be somewhat constant across samples, and therefore disentangled from the ecological signal by SML algorithms.

The occurrence of OTUs splitting, that is, morphospecies represented by several OTUs, could also be seen as a problem for the interpretation of eDNA data, because the biological meaning of OTUs is unclear[78−80] and because it increases the dimensionality of the data (see below). However, there is also some advantages related to the high number of OTUs. For instance, it has been shown that bacterial sequences diverging by a single nucleotide can display a different abundance profile across environmentally distinct samples,[81] pointing out that subspecies units can be ecologically informative. In our study, several foraminiferal morphospecies were represented by multiple OTUs. The detailed analysis of these OTUs in the case of the two dominant morphospecies: *Bulimina marginata* and *Cibicidoides lobatulus*, showed that in the first case the read abundance profiles of the two most abundant OTUs were highly similar across samples (spearman rho = 0.91, $p < 0.001$, Table S9), while in the second case the two most abundant OTUs displayed different profiles across samples (spearman rho = 0.09, $p = 0.32$, Table S9). This suggests that these OTUs could represent different populations of the same species that exhibit different ecological preference or that they belong to cryptic species complexes that our reference sequence database did not capture. The temptation to merge OTUs that are assigned to the same morphospecies, in an effort of matching the metabarcoding data with the observable morphospecies data, is questionable for two reasons. First, it means grouping sequences based on our current morpho-taxonomic knowledge, which is reflected in reference sequence databases, although morphospecies, from which DNA barcodes are generated, are not necessarily well-defined. Second, ecologically distinct subspecies units could be grouped together, losing the ecological signal conveyed by distinct OTUs into a single heterogeneous OTU. Because this signal is used for BI calculation, we think that SML approaches applied to biomonitoring would be more accurate in the case of OTUs splitting than in the case of OTUs merging. Furthermore, we think that SML approaches would as well benefit from the automation of a workflow that is entirety independent from reference taxonomic database, to be reproducible over time.

Building predictive models from eDNA composition data using SML approaches hold challenges too. Metabarcoding data are usually characterized by a much higher number of OTUs than samples,[82] and this can be further increased by the OTUs splitting, like in the present study. In SML, this data property makes the predictive models prone to be affected by "the curse of dimensionality".[31] Indeed, the probability to observe, by chance, a perfect correlation between the relative abundance of one OTU and the variation of a BI increases with the number of OTUs. This could lead to the overfitting of the model, thus decreasing its accuracy. Finding a trade-off between the OTU granularity (dimensionality) and the separation of ecologically relevant units appears as a major challenge for SML approaches applied to metabarcoding data. Although our data set was highly dimensional, our predictive models were likely not overfitted for two reasons. First, the overfitting is controlled through the growing of a "forest" of regression trees that are built on a random subset of the data in the RF algorithm,[60,30] and we used a 10-fold cross-validation step for model selection using the SOM algorithm, a common procedure to prevent such problem.[27,83] Second, we showed that our models still accurately predicted BI values without 80% of the OTUs

DOI: 10.1021/acs.est.7b01518

removed based in their low abundance, supporting that our models built on the full data set do not tend to overfit due to the presence of numerous rare OTUs.

Future efforts should investigate whether accurate SML predictions can be realized in broad routine applications. Our results were achieved with a relatively small training data set and increasing the number of sampled farms to cover a wider sampling area will likely further refine our predictive models. Given the ability to multiplex hundreds of samples and to perform analyses rapidly, cost-effective biomonitoring solutions could be proposed in matters of days instead of months.[84] Hence, an eDNA-based early warning system adapted for the pro-active management of marine industrial activities could be envisioned, as it has been already proposed for others ecosystems and type of bioindicators.[85,86]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.7b01518.

> Sampling site coordinates and distribution of collected samples, results of nonlinear models testing the effect of farming site, the cages and their interaction on four biotic indexes values, taxonomic composition of foraminiferal communities, accuracy of BI values predictions from composition data, nonlinear relationship between BI reference values and the distance to the cages, correlation between alpha-diversity metrics and the distance from the cages, NMDS analysis of the Bray−Curtis β-diversity matrix calculated from the CSS normalized foraminifera eDNA data set, effect of rarefying the OTU-to-sample matrix on the variation of inferred NSI values per sample, effect of rarefying the OTU-to-sample data set on the variation of inferred NQI1 values per sample, diversity metrics importance in the prediction of the NSI index from the Random Forest models, diversity metrics importance in the prediction of the NQI1 index from the Random Forest models, OTU importance in the prediction of the AMBI index from the Random Forest model, OTU importance in the prediction of the ISI index from the Random Forest model, and boxplot and Mann−Whitney tests for difference between predicted BI values with the best predictive models and reference BI values for each combination of farm and BI (PDF)
>
> Tag to sample reference table following a Latin Square Design and tagged primers sequences for successful PCR amplified samples, quality filtering and assembly parameters, morpho-taxonomic dataset, OTU-to-sample matrix with taxonomic assignments, and pair-wise spearman correlation matrix of OTUs represented by more than 1000 reads and assigned to the same species (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: tristan.cordier@gmail.com.

ORCID

Tristan Cordier: 0000-0001-7398-4790

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the Swiss Network for International Studies, by the Swiss National Science Foundation (grants 316030_150817 and 31003A_159709) and by the Norwegian Seafood Research Fund (FHF project number 901092). We thank Vegard Aambø Langvatn and Dagfinn Breivik Skomsø from Fiske Liv and Jeremiah Peder Ness from Aqua-kompetanse for help in collecting samples. We also thank Are Andreassen Moe and Arne Kvalvik from Marine Harvest for the organization of the sampling campaign. We also thank the four anonymous reviewers for their helpful comments.

■ REFERENCES

(1) Doney, S. C.; Ruckelshaus, M.; Emmett Duffy, J.; Barry, J. P.; Chan, F.; English, C. A.; Galindo, H. M.; Grebmeier, J. M.; Hollowed, A. B.; Knowlton, N.; et al. Climate Change Impacts on Marine Ecosystems. *Annu. Rev. Mar. Sci.* **2012**, *4* (1), 11−37.

(2) Peterson, C. H.; Rice, S. D.; Short, J. W.; Esler, D.; Bodkin, J. L.; Ballachey, B. E.; Irons, D. B. Long-Term Ecosystem Response to the Exxon Valdez Oil Spill. *Science* **2003**, *302*, 2082−2086.

(3) Smith, V. H.; Tilman, G. D.; Nekola, J. C. Eutrophication: Impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environ. Pollut.* **1999**, *100*, 179−196.

(4) Borja, A.; Ranasinghe, A.; Weisberg, S. B. Assessing ecological integrity in marine waters, using multiple indices and ecosystem components: Challenges for the future. *Mar. Pollut. Bull.* **2009**, *59* (1−3), 1−4.

(5) Tavakoly Sany, S. B.; Hashim, R.; Rezayi, M.; Salleh, A.; Safari, O. A review of strategies to monitor water and sediment quality for a sustainability assessment of marine environment. *Environ. Sci. Pollut. Res.* **2014**, *21* (813), 813.

(6) Borja, A.; Franco, J.; Pérez, V. A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. *Mar. Pollut. Bull.* **2000**, *40* (12), 1100−1114.

(7) Rygg, B. *Indicator Species Index for Assessing Benthic Ecological Quality in Marine Waters of Norway*, NIVA-rapport 4548; Norsk Institutt for Vannforskning, 2002.

(8) Rygg, B. *Developing Indices for Quality Status Classification of Marine Soft-Bottom Fauna in Norway*, NIVA-rapport 5208; Norsk Institutt for Vannforskning, 2006.

(9) Taberlet, P.; Coissac, E.; Pompanon, F.; Brochmann, C.; Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **2012**, *21* (8), 2045−2050.

(10) Woodward, G.; Gray, C.; Baird, D. J. Biomonitoring for the 21st Century: New perspectives in an age of globalisation and emerging environmental threats. *Limnetica* **2013**, *32* (2), 159−174.

(11) Aylagas, E.; Borja, Á.; Rodriguez-Ezpeleta, N. Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). *PLoS One* **2014**, *9* (3), e90529.

(12) Bohmann, K.; Evans, A.; Gilbert, M. T. P.; Carvalho, G. R.; Creer, S.; Knapp, M.; Yu, D. W.; de Bruyn, M. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution.* **2014**, *29*, 358−367.

(13) Kermarrec, L.; Franc, A.; Rimet, F.; Chaumeil, P.; Humbert, J. F.; Bouchez, A. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater diatoms. *Mol. Ecol. Resour.* **2013**, *13* (4), 607−619.

(14) Kermarrec, L.; Franc, A.; Rimet, F.; Chaumeil, P.; Frigerio, J.-M.; Humbert, J.-F.; Bouchez, A. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshw. Sci.* **2014**, *33* (1), 349−363.

(15) Visco, J. A.; Apothéloz-Perret-Gentil, L.; Cordonier, A.; Esling, P.; Pillet, L.; Pawlowski, J. Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. *Environ. Sci. Technol.* **2015**, *49* (13), 7597−7605.

(16) Zimmermann, J.; Abarca, N.; Enk, N.; Skibbe, O.; Kusber, W. H.; Jahn, R. Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. *PLoS One* **2014**, *9* (9), e108793.

(17) Zimmermann, J.; Glöckner, G.; Jahn, R.; Enke, N.; Gemeinholzer, B. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Mol. Ecol. Resour.* **2015**, *15* (3), 526−542.

(18) Apothéloz-Perret-Gentil, L.; Cordonier, A.; Straub, F.; Iseli, J.; Esling, P.; Pawlowski, J. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Mol. Ecol. Resour.* **2017**, DOI: 10.1111/1755-0998.12668.

(19) Chariton, A. A.; Court, L. N.; Hartley, D. M.; Colloff, M. J.; Hardy, C. M. Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Front. Ecol. Environ.* **2010**, *8* (5), 233−238.

(20) Chariton, A. A.; Stephenson, S.; Morgan, M. J.; Steven, A. D. L.; Colloff, M. J.; Court, L. N.; Hardy, C. M. Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environ. Pollut.* **2015**, *203*, 165−174.

(21) Bik, H. M.; Halanych, K. M.; Sharma, J.; Thomas, W. K. Dramatic shifts in benthic microbial eukaryote communities following the deepwater horizon oil spill. *PLoS One* **2012**, 7 (6), e38550.

(22) Pawlowski, J.; Esling, P.; Lejzerowicz, F.; Cedhagen, T.; Wilding, T. A. Environmental monitoring through protist next-generation sequencing metabarcoding: Assessing the impact of fish farming on benthic foraminifera communities. *Mol. Ecol. Resour.* **2014**, *14* (6), 1129−1140.

(23) Pochon, X.; Wood, S. A.; Keeley, N. B.; Lejzerowicz, F.; Esling, P.; Drew, J.; Pawlowski, J. Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Mar. Pollut. Bull.* **2015**, *100*, 370.

(24) Lejzerowicz, F.; Esling, P.; Pillet, L. L.; Wilding, T. a.; Black, K. D.; Pawlowski, J. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Sci. Rep.* **2015**, *5*, 13932.

(25) Aylagas, E.; Borja, Á.; Irigoien, X.; Rodríguez-Ezpeleta, N. Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Front. Mar. Sci.* **2016**, *3*, 96.

(26) Lanzén, A.; Lekang, K.; Jonassen, I.; Thompson, E. M.; Troedsson, C. High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Mol. Ecol.* **2016**, *25* (17), 4392−4406.

(27) Knights, D.; Costello, E. K.; Knight, R. Supervised classification of human microbiota. *FEMS Microbiology Reviews.* **2011**, *35*, 343−359.

(28) Olden, J. D.; Lawler, J. J.; Poff, N. L. Machine learning methods without tears: a primer for ecologists. *Q. Rev. Biol.* **2008**, *83* (2), 171−193.

(29) Philibert, A.; Desprez-Loustau, M. L.; Fabre, B.; Frey, P.; Halkett, F.; Husson, C.; Lung-Escarmant, B.; Marçais, B.; Robin, C.; Vacher, C.; et al. Predicting invasion success of forest pathogenic fungi from species traits. *J. Appl. Ecol.* **2011**, *48* (6), 1381−1390.

(30) Crisci, C.; Ghattas, B.; Perera, G. A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Modell.* **2012**, *240*, 113−122.

(31) Libbrecht, M. W.; Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16* (6), 321−332.

(32) Statnikov, A.; Henaff, M.; Narendra, V.; Konganti, K.; Li, Z.; Yang, L.; Pei, Z.; Blaser, M. J.; Aliferis, C. F.; Alekseyenko, A. V. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **2013**, *1* (1), 11.

(33) Beck, D.; Foster, J. A. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PLoS One* **2014**, *9* (2), e87830.

(34) Martínez-García, P. M.; López-Solanilla, E.; Ramos, C.; Rodríguez-Palenzuela, P. Prediction of bacterial associations with plants using a supervised machine-learning approach. *Environ. Microbiol.* **2016**, *18*, 4847.

(35) Smith, M. B.; Rocha, A. M.; Smillie, C. S.; Olesen, S. W.; Paradis, C.; Wu, L.; Campbell, J. H.; Fortney, J. L.; Mehlhorn, T. L.; Lowe, K. A.; et al. Natural bacterial communities serve as quantitative geochemical biosensors. *mBio* **2015**, *6* (3), e00326-15.

(36) Nigam, R.; Saraswat, R.; Panchang, R. Application of foraminifers in ecotoxicology: Retrospect, perspect and prospect. *Environ. Int.* **2006**, *32* (2), 273−283.

(37) Frontalini, F.; Coccioni, R. Benthic foraminifera as bioindicators of pollution: A review of Italian research over the last three decades. *Revue de Micropaleontologie.* **2011**, *54*, 115−127.

(38) Schönfeld, J.; Alve, E.; Geslin, E.; Jorissen, F.; Korsun, S.; Spezzaferri, S.; Members of the FOBMIO Group. The FOBIMO (FOraminiferal BIo-MOnitoring) initiative-Towards a standardised protocol for soft-bottom benthic foraminiferal monitoring studies. *Mar. Micropaleontol.* **2012**, *94*−*95*, 1−13.

(39) Alve, E.; Korsun, S.; Schönfeld, J.; Dijkstra, N.; Golikova, E.; Hess, S.; Husum, K.; Panieri, G. Foram-AMBI: A sensitivity index based on benthic foraminiferal faunas from North-East Atlantic and Arctic fjords, continental shelves and slopes. *Mar. Micropaleontol.* **2016**, *122*, 1−12.

(40) Scott, D. B.; Schafer, C. T.; Honig, C.; Younger, D. C. Temporal variations of benthic foraminiferal assemblages under or near aquaculture operations; documentation of impact history. *J. Foraminiferal Res.* **1995**, *25* (3), 224−235.

(41) Angel, D. L. Impact of a Net Cage Fish Farm on the Distribution of Benthic Foraminifera in the Northern Gulf of Eilat (Aqaba, Red Sea). *J. Foraminiferal Res.* **2000**, *30* (1), 54−65.

(42) Vidović, J.; Cosovic, V.; Jurasic, M.; Petricioli, D. Impact of fish farming on foraminiferal community, Drvenik Veliki Island, Adriatic Sea, Croatia. *Mar. Pollut. Bull.* **2009**, *58* (9), 1297−1309.

(43) Vidović, J.; Dolenec, M.; Dolenec, T.; Karamarko, V.; Žvab Rožič, P. Benthic foraminifera assemblages as elemental pollution bioindicator in marine sediments around fish farm (Vrgada Island, Central Adriatic, Croatia). *Mar. Pollut. Bull.* **2014**, *83* (1), 198−213.

(44) Pawlowski, J.; Esling, P.; Lejzerowicz, F.; Cordier, T.; Visco, J. A.; Martins, C. I. M.; Kvalvik, A.; Staven, K.; Cedhagen, T. Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. *Aquac. Environ. Interact.* **2016**, *8*, 371−386.

(45) Pawlowski, J.; Lejzerowicz, F.; Esling, P. Next-generation environmental diversity surveys of foraminifera: Preparing the future. *Biol. Bull.* **2014**, *227*, 93−106.

(46) Esling, P.; Lejzerowicz, F.; Pawlowski, J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res.* **2015**, *43*, 2513.

(47) Edgar, R. C.; Haas, B. J.; Clemente, J. C.; Quince, C.; Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **2011**, *27* (16), 2194−2200.

(48) Caporaso, J. G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F. D.; Costello, E. K.; Fierer, N.; Pena, A. G.; Goodrich, J. K.; Gordon, J. I.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7* (5), 335−336.

(49) Mahé, F.; Rognes, T.; Quince, C.; De Vargas, C.; Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **2015**, *3*, e1420.

(50) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(51) R Core Team. *R: A language and environment for statistical computing*; R Found. Stat. Comput.: Vienna, Austria, 2016; https://www.R-project.org/.

(52) Oksanen, J.; Blanchet, F. G.; Kindt, R.; Legendre, P.; Minchin, P. R.; O'Hara, R. B.; Simpson, G. L.; Solymos, P.; Stevens, M. H. H.; Wagner, H. *vegan: Community Ecology Package*, R package version 2.4-1; 2016.

(53) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Information*, 1949; Vol. 97.

(54) Pielou, E. C. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **1966**, *13*, 131−144.

(55) Chao, A.; Shen, T. J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **2003**, *10* (4), 429−443.

(56) Paulson, J. N.; Stine, O. C.; Bravo, H. C.; Pop, M. Differential abundance analysis for microbial marker gene surveys. *Nat. Methods* **2013**, *10* (12), 1200−1202.

(57) McMurdie, P. J.; Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* **2014**, *10* (4), e1003531.

(58) Weiss, S. J.; Xu, Z.; Amir, A.; Peddada, S.; Bittinger, K.; Gonzalez, A.; Lozupone, C.; Zaneveld, J. R.; Vazquez-Baeza, Y.; Birmingham, A.; et al. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ. Prepr.* **2015**, *3*, e1408.

(59) Bray, J. R.; Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **1957**, *27* (4), 325−349.

(60) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.

(61) Wright, M. N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. arXiv:1508.04409 [stat.ML]. arXiv.org e-Print archive. https://arxiv.org/abs/1508.04409.

(62) Wehrens, R.; Buydens, L. M. C. Self- and super-organizing maps in R: The kohonen package. *J. Stat. Softw.* **2007**, *21* (5), 1−19.

(63) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.

(64) Kuhn, M.; R Core Team. *caret: Classification and Regression Training*, R package version 6.0-73; 2016.

(65) Gamer, M.; Lemon, J.; Fellows, I.; Singh, P. Various Coefficients of Interrater Reliability and Agreement.

(66) Landis, J. R.; Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33* (1), 159−174.

(67) Stevenson, R. J.; Pan, Y.; VanDam, H. Assessing environmental conditions in rivers and streams with diatoms. *Earth Sci.* **2006**, 57−85.

(68) Grego, M.; De Troch, M.; Forte, J.; Malej, A. Main meiofauna taxa as an indicator for assessing the spatial and seasonal impact of fish farming. *Mar. Pollut. Bull.* **2009**, *58* (8), 1178−1186.

(69) Lima-Mendez, G.; Faust, K.; Henry, N.; Decelle, J.; Colin, S.; Carcillo, F.; Chaffron, S.; Ignacio-Espinosa, J. C.; Roux, S.; Vincent, F.; et al. Determinants of community structure in the global plankton interactome. *Science (Washington, DC, U. S.)* **2015**, *348* (6237), 1262073.

(70) Vacher, C.; Tamaddoni-Nezhad, A.; Kamenova, S.; Peyrard, N.; Moalic, Y.; Sabbadin, R.; Schwaller, L.; Chiquet, J.; Smith, M. A.; Vallance, J.; et al. Learning Ecological Networks from Next-Generation Sequencing Data. *Adv. Ecol. Res.* **2016**, *54*, 1−39.

(71) Fortunato, C. S.; Eiler, A.; Herfort, L.; Needoba, J. A.; Peterson, T. D.; Crump, B. C.; et al. Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *ISME J.* **2013**, *7* (10), 1899−1911.

(72) Dell'Anno, A.; Corinaldesi, C. Degradation and turnover of extracellular DNA in marine sediments: Ecological and methodological considerations. *Appl. Environ. Microbiol.* **2004**, *70* (7), 4384−4386.

(73) Corinaldesi, C.; Danovaro, R.; Dell'Anno, A. Simultaneous recovery of extracellular and intracellular DNA suitable for molecular studies from marine sediments. *Appl. Environ. Microbiol.* **2005**, *71* (1), 46−50.

(74) Pillet, L.; Fontaine, D.; Pawlowski, J. Intra-genomic ribosomal RNA polymorphism and morphological variation in elphidium macellum suggests inter-specific hybridization in foraminifera. *PLoS One* **2012**, *7* (2), e32373.

(75) Weber, A. A. T.; Pawlowski, J. Can Abundance of Protists Be Inferred from Sequence Data: A Case Study of Foraminifera. *PLoS One* **2013**, *8* (2), e56739.

(76) Schirmer, M.; Ijaz, U. Z.; D'Amore, R.; Hall, N.; Sloan, W. T.; Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **2015**, *43* (6), e37.

(77) Esling, P.; Lejzerowicz, F.; Pawlowski, J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res.* **2015**, *43* (5), 2513−2524.

(78) Schloss, P. D.; Westcott, S. L. Assessing and improving methods used in OTU-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* **2011**, *77* (10), 3219−3226.

(79) Chen, W.; Zhang, C. K.; Cheng, Y.; Zhang, S.; Zhao, H. A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLoS One* **2013**, *8* (8), e70837.

(80) He, Y.; Caporaso, J. G.; Jiang, X.-T.; Sheng, H.-F.; Huse, S. M.; Rideout, J. R.; Edgar, R. C.; Kopylova, E.; Walters, W. A.; Knight, R.; et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **2015**, *3* (1), 20.

(81) Eren, A. M.; Maignien, L.; Sul, W. J.; Murphy, L. G.; Grim, S. L.; Morrison, H. G.; Sogin, M. L. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **2013**, *4* (12), 1111−1119.

(82) Tsilimigras, M. C. B.; Fodor, A. A. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* **2016**, *26* (5), 330−335.

(83) Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55* (10), 78.

(84) Quinn, R. A.; Navas-molina, J. A.; Hyde, E. R.; Song, J.; Vázquez-baeza, Y.; Humphrey, G.; Gaffney, J.; Minich, J. J.; Melnik, A. V; Herschend, J. From sample to multi-omics conclusions in under 48 h. *mSystems* **2016**, DOI: 10.1128/mSystems.00038-16.

(85) Wepener, V.; van Vuren, J. H. J.; Chatiza, F. P.; Mbizi, Z.; Slabbert, L.; Masola, B. Active biomonitoring in freshwater environments: Early warning signals from biomarkers in assessing biological effects of diffuse sources of pollutants. *Phys. Chem. Earth* **2005**, *30* (11−16), 751−761.

(86) Maradona, A.; Marshall, G.; Mehrvar, M.; Pushchak, R.; Laursen, A. E.; McCarthy, L. H.; Bostan, V.; Gilbride, K. A. Utilization of multiple organisms in a proposed early-warning biomonitoring system for real-time detection of contaminants: Preliminary results and modeling. *J. Hazard. Mater.* **2012**, *219−220*, 95−102.