






## RESEARCH ARTICLE

# A spatial matrix factorization method to characterize ecological assemblages as a mixture of unobserved sources: An application to fish eDNA surveys

Letizia Lamperti<sup>1,2,3</sup>  | Olivier François<sup>4</sup> | David Mouillot<sup>5,6</sup> | Laëtizia Mathon<sup>1</sup>  |  
Théophile Sanchez<sup>2,3</sup>  | Camille Albouy<sup>2,3</sup> | Loïc Pellissier<sup>2,3</sup>  | Stéphanie Manel<sup>1,6</sup> 

<sup>1</sup>CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France; <sup>2</sup>Landscape Ecology, Institute of Terrestrial Ecosystems, ETH Zürich, Zürich, Switzerland; <sup>3</sup>Swiss Federal Research Institute WSL, Birmensdorf, Switzerland; <sup>4</sup>TIMC, CNRS UMR 5525, Université Grenoble-Alpes, Grenoble-INP, Grenoble, France; <sup>5</sup>MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France and <sup>6</sup>Institut Universitaire de France, Paris, France

**Correspondence**

Letizia Lamperti and Stéphanie Manel

Email: [le.lamperti@gmail.com](mailto:le.lamperti@gmail.com) and [stephanie.manel@umontpellier.fr](mailto:stephanie.manel@umontpellier.fr)**Funding information**

Université de Montpellier

**Handling Editor:** Chloe Robinson**Abstract**

1. Understanding how ecological assemblages vary in space and time is essential for advancing our knowledge of biodiversity dynamics and ecosystem functioning. Metabarcoding of environmental DNA (eDNA) is an efficient method for documenting biodiversity changes in both marine and terrestrial ecosystems. However, current methods fail to detect and display the biodiversity structure within and between eDNA samples limiting ecological and biogeographical interpretations.
2. We present a spatial matrix factorization method that identifies optimal eDNA sample assemblages—called pools—assuming that taxonomic unit composition is based on a fixed number of unknown sources. These sources, in turn, represent taxonomic units sharing similar habitat properties or characteristics. The method aims to reduce the multi-taxa composition structure into a low number of dimensions defined by these sources. This method is inspired by admixture analysis in population genetics. Using a marine fish eDNA survey on 263 sampling stations detecting 2888 molecular operational taxonomic units (MOTUs), we apply this method to analyse the biogeography and mixing patterns of fish assemblages at regional and large scales.
3. At large scale, our analysis reveals six primary pools of fish samples characterized by distinct biogeographic patterns, with some mixtures between these pools. We identify pools composed of unique sources, corresponding to distinct and more isolated regions such as the Mediterranean and Scotia Seas. We also identify pools composed of a greater mix of sources, corresponding to geographically connected areas, such as tropical regions. Additionally, we identify the taxa underpinning the formation of each pool. In the regional analysis of Mediterranean

Loïc Pellissier and Stéphanie Manel—Co-senior authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

eDNA samples, our method successfully identifies different pools, allowing the detection of not only geographic gradients but also human-induced gradients corresponding to protection levels.

4. Spatial matrix factorization adds a new method in community ecology, where each sample is considered as a mixture of  $K$  unobserved sources, to assess the dissimilarity of ecological assemblages revealing environmental and human-induced gradients. Beyond the study of fish eDNA samples, this method has the potential to shed new light on any biodiversity survey and provide new bioindicators of global change.

#### KEYWORDS

biogeography, environmental DNA (eDNA), fish communities, matrix factorization, metabarcoding

## 1 | INTRODUCTION

Species assemblages exhibit spatiotemporal variations influenced by a multitude of factors, such as environmental disturbance, biotic interactions and stochasticity (Blowes et al., 2024; Ji et al., 2024; Leibold et al., 2022). Assessing the structure of species assemblages in different spatial contexts is essential for advancing our understanding of biodiversity distribution and ecosystem functioning (Barwell et al., 2015; Mori et al., 2018). Since no single species assemblage can support all ecological functions and contributions to people (Mayor et al., 2024), distinct local species assemblages across heterogeneous environments must be maintained (Hillebrand & Matthiessen, 2009; Loiseau et al., 2021).

In community ecology, variation in species assemblages across space is typically quantified using  $\beta$  diversity, which quantifies differences in species identity and abundance between local assemblages—an essential component of biodiversity (Mori et al., 2018; Whittaker, 1972). Yet, it does not explicitly reveal the composition of species assemblages or species composition at specific locations and its high dimensionality can make these multiple pairwise distances difficult to interpret.

Environmental DNA (eDNA) metabarcoding has opened new possibilities for studying the biogeography of marine species assemblages across scales (Leray & Knowlton, 2015; Mathon et al., 2022). The analysis of eDNA samples involves the extraction, amplification and sequencing of the genetic material collected (Deiner et al., 2017). Next-generation sequencing and bioinformatic analysis can identify specific eDNA sequences associated with different organisms, often referred to as taxonomic units (Marques et al., 2020). These units serve as proxies that help identify species or taxonomic groups in a given environment, similar to the molecular operational taxonomic units (MOTUs) described by Marques et al. (2020). For example, Mathon et al. (2023) observed a strong correlation between fish MOTUs  $\beta$  diversity and environmental factors in coastal ecosystems worldwide. While  $\beta$  diversity is a valuable metric, it does not reveal the main structure of

assemblages nor it does offer a clear visualization of assemblages in terms of taxonomic unit compositions (Podani & Schmera, 2011). To reduce the hyper-dimensionality of pairwise  $\beta$  diversity values, ordination methods such as principal coordinate analysis (PCoA; Gower, 1966), non-metric multidimensional scaling (NMDS; Kruskal, 1964) or canonical correspondence analysis (CCA; ter Braak, 1986) are classically used to represent the grouping of assemblages in a lower number of dimensions (Legendre & Legendre, 1998). However, these methods are limited by their inability to simultaneously represent the spatial and compositional relationships of assemblages at individual sites. As a result, our understanding of how species composition varies across geographic locations is limited. Alternatively, probabilistic models can provide a more complete understanding of assemblage structure within and across eDNA surveys. For example, occupancy models are used in eDNA studies to estimate the probability of species occurrence while accounting for imperfect detection (McClenaghan et al., 2020; Uthicke et al., 2022). Sommeria-Klein et al. (2021) used a probabilistic model of species occurrence based on eDNA data from 70 major eukaryotic plankton groups collected during the Tara Ocean Project in different ocean basins. They reveal an alignment of plankton assemblages with biogeographic patterns, particularly in the most diverse groups. Although their method analysed species assemblages rather than  $\beta$  diversity patterns, their use of a Bayesian model like Latent Dirichlet Allocation (LDA; Blei et al., 2003; Valle et al., 2014) has some limitations. LDA was originally developed to describe a corpus of text documents as composed of unknown topics. The statistical analogy between the presence/absence of a word in a document and a species in an ecological community allows the inference of community structure as the composition of species from a specified number of sources. The same analogy and statistical model were used in population genetics to evaluate genetic admixture in population samples (Pritchard et al., 2000). However, these Bayesian approaches are computationally intensive and cannot account for the spatial nature of observations, making them inefficient for handling and understanding large-scale datasets (Caye et al., 2016).

With over 30,000 taxa—half of them being marine species—fishes are the most diverse vertebrate group (Costello et al., 2015). Conducting biogeography studies on this group in open and connected spaces such as oceans and seas is challenging due to the vastness and heterogeneity of these environments (Benestan et al., 2021). In addition to their substantial taxonomic richness, marine fishes exhibit a wide range of life history characteristics, behaviour and diet that contribute to key ecological processes in marine ecosystems (Villéger et al., 2017). They inhabit diverse marine environments from the equator to the poles and from coastal regions to the abyss (Nelson, 2016). However, marine fish populations face significant threats, including industrial fishing practices and pollution, that threaten their existence (Johnston et al., 2021; Jouffray et al., 2020). Therefore, understanding the structure and resilience of fish assemblages is essential for their conservation (Makiola et al., 2020).

Here, we present a novel approach to describe and infer patterns of taxonomic unit assemblage composition within their geographic and environmental context. Like LDA, our model assumes that observed taxonomic unit frequencies in sampled assemblages come from a mixture of  $K$  unobserved sources. However, our model integrates spatial information and is computationally more efficient. Our approach estimates (i) the sample pools from our spatial eDNA surveys as mixtures of  $K$  different sources and (ii) the frequency of the taxonomic unit assemblages, that is the frequency of each taxonomic unit from the  $K$  different sources. These mixtures exhibit continuous distributions across geographic and environmental gradients, with parameters derived from genetic and spatial data. We conduct analyses of taxonomic units and assemblage compositions on a large scale as well as regionally, focusing on data from the Mediterranean Sea where some samples come from no-take marine reserves.

## 2 | MATERIALS AND METHODS

### 2.1 | Geospatial constraints in taxonomic unit assemblage mixture estimation

The community matrix, denoted as  $X$ , consists of  $n$  samples and  $m$  taxonomic units. The matrix  $X$  can represent various data types, such as the presence of each taxonomic unit in each sample or the abundance or detection probability of each unit. Our model assumes that the frequencies of taxonomic units detected in a given assemblage are drawn from a mixture of  $K$  unobserved sources, where  $K$  is unknown. These sources represent taxonomic units that share similar habitat characteristics or traits.

The algorithm estimates two matrices: the assemblage sample matrix  $S$  (with dimension  $n \times K$ ), representing the proportions of each sample in each source, and the matrix  $M$  (with dimensions  $K \times m$ ), representing for each source the frequencies of each taxonomic unit. In this context, the coefficient  $s_{ik}$  represents the fraction of sample  $i$  belonging to source  $k$ , while the coefficient  $m_{kj}$  represents the frequency of the taxonomic unit  $j$  in source  $k$ .

We assume that the frequency of the specific taxonomic unit  $j$  in sample  $i$  is determined by the law of total probability:

$$x_{ij} = \text{frequency of taxonomic unit } j \text{ in sample } i = \sum_{k=1}^K s_{ik} m_{kj}.$$

According to this formula, each sample is formed by a mixture of sources and by the taxonomic unit frequencies in each source. In matrix terms, solving the above equation is equivalent to finding two matrices,  $S$  and  $M$ , with non-negative coefficients and subject to some probabilistic constraints, such that:

$$X = SM.$$

Thus, the estimation of the  $S$  and  $M$  matrices can be formalized as a non-negative matrix factorization problem (Lee & Seung, 1999). To account for the probabilistic constraints on  $S$  and  $M$ , we perform a matrix factorization using an alternate least-squares minimization algorithm, as implemented in the method estimating individual ancestry coefficients from population genetic samples (Caye et al., 2016; Frichot et al., 2014). To account for spatial autocorrelation among samples, additional constraints are introduced into the minimization problem to ensure that geographically proximate samples are more likely to have the same taxonomic unit composition than distant samples (Caye et al., 2018).

The model allows us to estimate the mixture of samples ( $S$ ) and the respective pools, and the taxonomic unit frequencies in the sources ( $M$ ) in the community under study. A given taxonomic unit could be present in multiple sources, and variable proportions of sources define each sample.

We determine the number of sources,  $K$ , by evaluating a cross-entropy criterion for each  $K$  (Caye et al., 2018). The choice of  $K$  is based on a cross-validation method that partitions the input matrix entries into a training and a testing dataset. The cross-entropy criterion compares the predicted taxonomic unit frequencies from the training set with those calculated from the testing set in each sample. Smaller values of the criterion typically indicate better values for  $K$ .

To identify the taxonomic units that show the greatest differences between the  $K$  sources, we use the  $M$  matrix to calculate taxonomic unit frequencies within the  $K$  sources, and we compute an ANOVA statistic for each taxonomic unit. The ANOVA  $F$  statistic is used as a measure of differentiation between sources, and null hypotheses (no difference between sources) were tested using Fisher thresholds with  $K-1$  and  $n-K$  degrees of freedom (François et al., 2016).

### 2.2 | Case study: eDNA Surveys of coastal marine fishes

#### 2.2.1 | eDNA collection and sample processing

We used eDNA samples of seawater collected at 263 stations, in 11 marine regions covering the global ocean from pole to pole (Mathon et al., 2023). Between 1 and 4 replicates were sampled

at each station, and in this study, we considered only samples filtered between 0 and 40 m deep. Four different sampling methods were used: on-point water collection with sterile containers, from the surface or close to the substrate, and filtration along a 2-km long transect, below the surface or close to the bottom. Details on which sampling method was used in each region are provided in Mathon et al. (2023). For samples collected with sterile containers, the seawater was filtered with sterile Sterivex filter capsules (Merck® Millipore; pore size 0.22 µm) using disposable sterile syringes. Immediately after, the filter units were filled with CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored in 50 mL screw-cap tubes at room temperature. eDNA filtrations along transects were performed with an Athena® peristaltic pump (Proactive Environmental Products LLC, Bradenton, Florida, USA; nominal flow of 1.0 L/min ± 15%), a VigiDNA® 0.2 µm cross-flow filtration capsule (SPYGEN, le Bourget du Lac, France) and disposable sterile tubing for each filtration capsule. At the end of each filtration, the water inside the capsules was emptied, and the capsules were filled with 80 mL of CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored at room temperature. For each sampling campaign, a strict contamination control protocol was followed in both field and laboratory stages (Valentini et al., 2016), and each water sample processing included the use of disposable gloves and single-use filtration equipment. Negative field controls were performed in multiple sites across all sampling locations and revealed no contamination from the boat or samplers.

### 2.2.2 | eDNA extraction, amplification and sequencing

DNA extraction was performed in a dedicated DNA laboratory (SPYGEN, [www.spygen.com](http://www.spygen.com)) equipped with positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulations. eDNA extractions were performed following the protocols described by Pont et al. (2018) for SPYGEN capsules, and by Juhel et al. (2020) for the sterivex filters. A teleost-specific 12S mitochondrial rRNA gene primer pair (teleo, forward primer—ACACCGCCCGTCACTCT, reverse primer—CTCCGGTACTTACCATG; Valentini et al., 2016) was used for the amplification of metabarcode sequences. As we analysed our data using MOTUs as a proxy for species to overcome genetic database limitations, we chose to amplify only one marker. Twelve DNA amplifications PCR per sample (i.e. replicates) were performed in a final volume of 25 µL, using 3 µL of DNA extract as the template. The amplification was performed following the protocol of Pont et al. (2018). The purified PCR products were pooled in equal volumes, to achieve a theoretical sequencing depth of 1,000,000 reads per sample. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). A total of 45 libraries were prepared using the MetaFast protocol for Illumina sequencing platforms. A paired-end

sequencing (2 × 125 bp) was carried out using an Illumina HiSeq 2500 sequencer with the HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) or a MiSeq (2 × 125 bp, Illumina, San Diego, CA, USA) using the MiSeq Flow Cell Kit v3 (Illumina, San Diego, CA, USA) or a NextSeq sequencer (2 × 125 bp, Illumina, San Diego, CA, USA) with the NextSeq Mid kit following the manufacturer's instructions. This generated an average of 624,468 sequence reads (paired-end Illumina or Ion Torrent) per sample. Many extraction and amplification negative controls were performed for each sample.

### 2.2.3 | Bioinformatic analysis

Following sequencing, reads were processed using clustering and post-clustering cleaning to remove errors and estimate the number of species using MOTUs (Marques et al., 2020). First, reads were assembled using vsearch (Rognes, 2016), then demultiplexed and trimmed using Cutadapt (Martin, 2011) and clustering was performed using Swarm v.2 (Mahé et al., 2014) with a minimum distance of 1 mismatch between clusters. Taxonomic assignment of MOTUs was carried out using the lower common ancestor (LCA) algorithm ecotag implemented in the ObiTools toolkit (Boyer et al., 2016) and the European Nucleotide Archive (ENA; Leinonen et al., 2011) as a reference database (release 143, March 2020), supplemented by our custom reference database, containing approximately 800 sequences. To avoid spurious MOTUs originating from a PCR error, we applied quality filters, and we discarded all samples with less than 10 reads and present in only one PCR replicate. Then, errors generated by index-hopping (MacConaill et al., 2018) were filtered using a threshold empirically determined per sequencing batch using experimental blanks (combinations of tags not present in the libraries; Taberlet et al., 2018). Tag-jump (Schnell et al., 2015) was corrected using a threshold of 0.001 occurrence for a given MOTU within a library. At the species level, taxonomic assignments were accepted, as putative species, if the percentage of similarity with the reference sequence was 100%; at the genus level, if the similarity was between 90% and 99%, and at the family level if the similarity was >85%. If these criteria were not met, the MOTU was left unassigned. The post-LCA algorithm correction threshold of 85% similarity for the family assignment was chosen to include a maximum of the correct family assignment while minimizing the risk of adding wrong family assignments in the family detections. The number of reads, MOTUs and species after each cleaning step are available in the Supplementary material of Mathon et al. (2023).

### 2.2.4 | Dataset and PCR proportion

The final large-scale dataset consists of 522 samples and a total of 2888 detected MOTUs. The regional dataset, which includes only samples from the Mediterranean Sea, comprises 108 samples and

249 MOTUs. Each geo-localized sample records the frequency of detection for each MOTU across 12 PCR replicates. Subsequently, the dataset was scaled by dividing the PCR detection number of each MOTU per sample by the total number of replicates conducted. The frequency of detection for each MOTU across 12 PCR replicates was chosen to account for the variability in DNA extraction from samples and potential errors in the amplification and sequencing processes of eDNA data. This approach provides a more reliable estimate of MOTU presence compared with binary presence/absence data.

## 2.3 | Algorithm and computing pipeline

### 2.3.1 | Assemblage analysis

To estimate the number of  $K$  sources and the two matrices  $S$  and  $M$ , we used the *tess3* function from the 'tess3r' R-package to implement the matrix factorization method. Originally designed for the analysis of large georeferenced genotype datasets (Caye et al., 2018; François, 2016), we customized the algorithm for eDNA metabarcoding data and adapted it to the context of fish biogeography. Details on the new matrix factorization algorithm are provided in Supporting Information (Appendix S1). The method and the analysis were performed using R (version 4.1.3; R Core Team, 2022).

The number of sources,  $K$ , was determined by varying  $K$  in the range of 3–9. The geographic coordinates of each sample were provided as latitude and longitude. To select the optimal number of sources,  $K$ , we computed the cross-entropy score using the 'tess3r' package. A principal component analysis (PCA) was also performed to examine the explanatory variance of the axes. To perform the PCA, we used the *prcomp* function in the R package 'stats'. We considered that the optimal value of  $K$  to reveal the biogeographical delimitation of our samples was at the edge of the plateau when representing the cross-validation score performed by the model.

### 2.3.2 | Identification of the most differentiating MOTUs at a large scale

In addition to the spatial assemblage analysis, we explored the correlation between specific MOTUs and their respective sources. Using the matrix of MOTU frequencies across different sources,  $M$ , we performed an ANOVA to detect significant differences between sources. The null hypothesis was that there were no discernible differences between the sources, providing a baseline for our analyses.

To identify the MOTUs that most strongly differentiate between sources, we focused on those showing the highest statistical significance in the ANOVA results. We calculated the  $p$ -values for each MOTU and then applied a logarithmic transformation to

these values, representing them as  $-\log_{10}(p\text{-values})$  to enhance visualization.

To emphasize species that are highly specific to a particular source, we set an expected false discovery rate (FDR) of  $q=10\text{--}30$ . This stringent threshold helped us isolate a small number of MOTUs that are particularly indicative of specific sources, highlighting their unique contributions to the overall assemblage composition.

The results of these tests were presented visually using a plot of the  $-\log_{10}(p\text{-values})$ , allowing us to clearly identify the most differentiating MOTUs at a large scale.

To test the species composition characterizing each pool and the mixture of species frequencies, we extracted geographic ranges from the global species distribution map provided by Duhamet et al. (2023). We intersected the species detected in our dataset and the global species distribution map obtained for 730 species. We transformed the detection frequency of each species in our model to presence–absence. Furthermore, we considered that the frequency of a species within a pool is detected if its value exceeds the third quartile, to ensure significant detection concerning the value of frequencies.

### 2.3.3 | Comparison with other methods

We compared our method to one of the most commonly used methods for studying species assemblages: PCA followed by  $k$ -means clustering on the first six most explainable axes (cumulative explained variance 0.42). The 'stats' package in R was used to execute the PCA using the *prcomp* function. The first six most explainable PCA scores, representing the projections of the data onto the principal components, were then subjected to  $k$ -means clustering using the *kmeans* function from the same package, specifying 6 as the number of clusters, corresponding to the optimal number of pools found by our method.

### 2.3.4 | Test of the protection effect

To test the capability of the method on a smaller spatial scale, we selected a subset of data specific to the Mediterranean Sea, resulting in 108 samples (34 inside and 74 outside no-take marine reserves) and 249 MOTUs, and analysed it with our matrix factorization approach, using the same parameters as for the whole dataset (Appendix S1). To test the correlation between the probability matrix  $S$ , and the protection level of the samples, we transformed the probability matrix  $S$  using an isometric log-ratio (ILR) transformation implemented by the *ILR* function of the 'compositions' R package. We then applied a generalized linear model (GLM) to the transformed matrix  $S$ , where the transformed  $S$  follows a binomial model with the level of protection (inside vs. outside reserves) as the explanatory variable. To assess the significance of the model, we performed a chi-squared test. Furthermore, we calculated the

pseudo- $R$ -squares of McFadden using the  $pR2$  function from the 'pscl' R package.

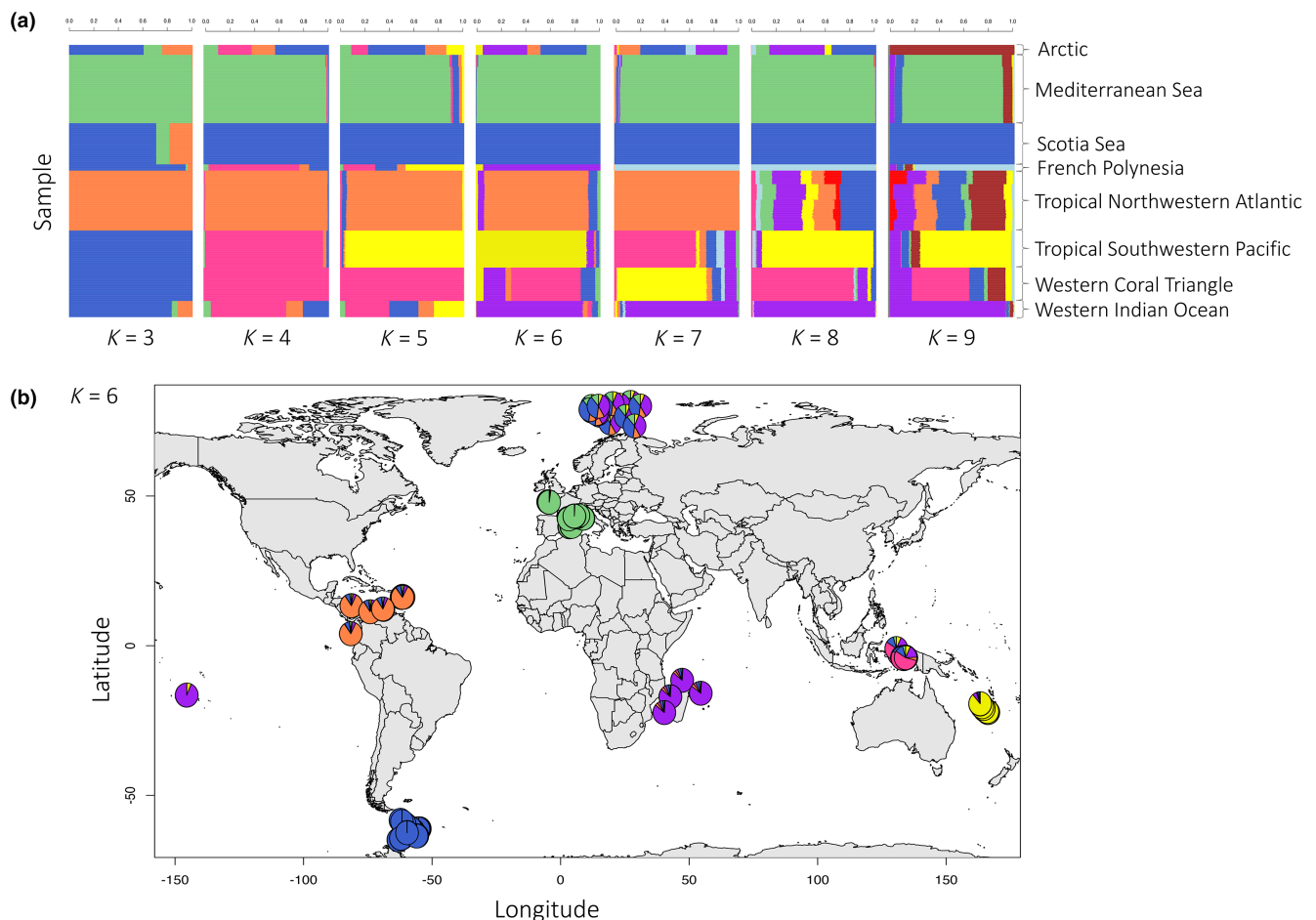
### 3 | RESULTS

#### 3.1 | Assemblage analysis at large scale

We explored a range of values for the number of sources  $K$ , varying from 3 to 9, and for each  $K$  value, we derived the matrix  $S$ , representing the proportion of each sample in each source. This allowed us to study how sample proportions and pool composition changed with the number of sources. We presented the results in bar plots illustrating the proportion of samples within each sample pool for  $K$  values ranging from 3 to 9 (Figure 1a). Specifically, for  $K=3$ , the algorithm discriminated the Mediterranean Sea, the tropical northwestern Atlantic Ocean, and a third pool which grouped all other samples in blue. For  $K=4$ , the algorithm revealed the emergence of a tropical Indo-Pacific-French Polynesia pool, shown in pink. Moving

to  $K=5$ , the algorithm detected the presence of the New Caledonian pool, identified as the tropical Southwest Pacific in yellow. At  $K=6$ , the algorithm provided clear distinctions for the Indian Ocean pool and French Polynesia, characterized by purple. For  $K=7$ , a pool associated with French Polynesia emerged, represented by the light blue. For  $K=8$ , the algorithm identified a new split within the tropical northwestern Atlantic with varying proportions of different sources due to the presence of samples from the Pacific coast of Colombia. For  $K=9$ , the algorithm identified an additional pool associated with the Arctic, indicated in brown. Notably, due to the limited data available from Arctic samples, this assemblage was only detected at  $K=9$ .

The optimal value of  $K$  was reached for  $K=6$  (Figure S1a). In addition, a scree plot resulting from the PCA performed on the raw data matrix, representing the probability of MOTU detection in PCR replicates for each sample, showed an elbow at PCA axis 3 and a decrease at PCA axis 6 (Figure S1b). We examined the sensitivity of our method's biogeographic delimitation of samples to  $K=6$  and observed that the six sample pools correspond to well-identified biogeographical regions (Figure 1b).



**FIGURE 1** Environmental DNA sample mixture mapping at large scale. (a) Bar plots representing fish samples (horizontal lines) for varying numbers of sources ( $K$ ) from 3 to 9. Each source is depicted with a different colour. (b) World map with a pie plot illustrating the sample mixture for  $K=6$ . Samples cluster into six pools corresponding to distinct biogeographic regions: The Tropical southwestern Pacific (yellow), the Western Coral Triangle (pink), the Mediterranean Sea (green), the Tropical northwestern Atlantic (orange), the western Indian Ocean (purple) and the Scotia Sea (blue).

### 3.2 | MOTU proportions at large scale

For  $K=6$ , we examined the M matrix, which represents the frequencies of MOTUs within each source (Figure 2a). To facilitate the interpretation, we streamlined the results by selecting the most abundant MOTUs. Specifically, we selected those MOTUs for which cumulative frequencies exceeded a predefined threshold, fixed at 30%, selecting 56 MOTUs out of 2888 (Figure 2b). This threshold was arbitrary and served as a filter, emphasizing MOTUs with substantial representation in the large-scale dataset (2% of all MOTUs).

Our analysis distinguished 'pure' and 'mixed' MOTUs, based on whether MOTUs' frequencies originated from only one source or multiple sources, respectively. Pure MOTUs were prevalent in regions characterized by isolation and endemism, such as the Mediterranean and Scotia Seas (depicted in green and blue, respectively in Figure 2a,b). In contrast, mixed MOTUs were indicative of regions featuring a mixture of sources, such as MOTUs characteristic of the Indo-Pacific tropical regions (illustrated in yellow, pink and purple; Figure 2a,b). This analysis revealed a direct correspondence between the frequency of mixtures within MOTUs and the mixture of sample proportions (Figures 1 and 2). This alignment is closely related to the biogeographic regions delineated by the six distinct pools, providing valuable insights into how MOTU compositions in samples relate to biogeographical regions. In addition, by reporting the taxonomic assignments of the various MOTUs to infer species identity, when possible, we were able to associate the proportion of each taxon in the pool set (Figure 2b).

Our analysis revealed that MOTUs with the highest global differentiation played a critical role in delineating distinct pools. After assigning taxon to MOTUs, our observations indicated that the most differentiated MOTUs corresponded to species with localized distributions, as shown in the  $p$ -value plot derived from the ANOVA test performed on the M matrix (Figure 3a). Indeed, the most differentiated species according to the test are the Antarctic cod, that is *Nototothenia coriiceps*, the yellowback fusilier, that is *Pterocaesio tessellata* and the dreamfish, that is *Sarpa Salpa*. These species correspond to assemblages sampled from the Scotia Sea, the tropical Indo-Pacific and the Mediterranean Sea, respectively (Figure 3b, ref. FishBase/NCBI). In addition, the list of MOTUs exceeding the threshold defined by the FDR includes taxa specific to the defined pools (Figure 3c).

To evaluate the species composition characterizing each pool and the mixture of species frequencies, we used geographic ranges extracted from the global species distribution map provided by Duhamet et al. (2023). We selected the geographic range of each species within our study area (Figure S2) and intersected these ranges with the MOTUs for which we had taxonomic information at the species level. A linear correlation between the global species distribution and the transformed M matrix produced a Pearson correlation of  $r=0.62$  ( $p$ -value  $<0.001$ ). This analysis establishes a clear link between endemic species (species unique to specific regions) and the biogeographic regions defined by the identified assemblages

(Figure 2b). This exploration helps us better understand how species are distributed in specific regions and how their assemblage composition differs between regions (Figures 1b and 2b).

### 3.3 | Comparison with other methods

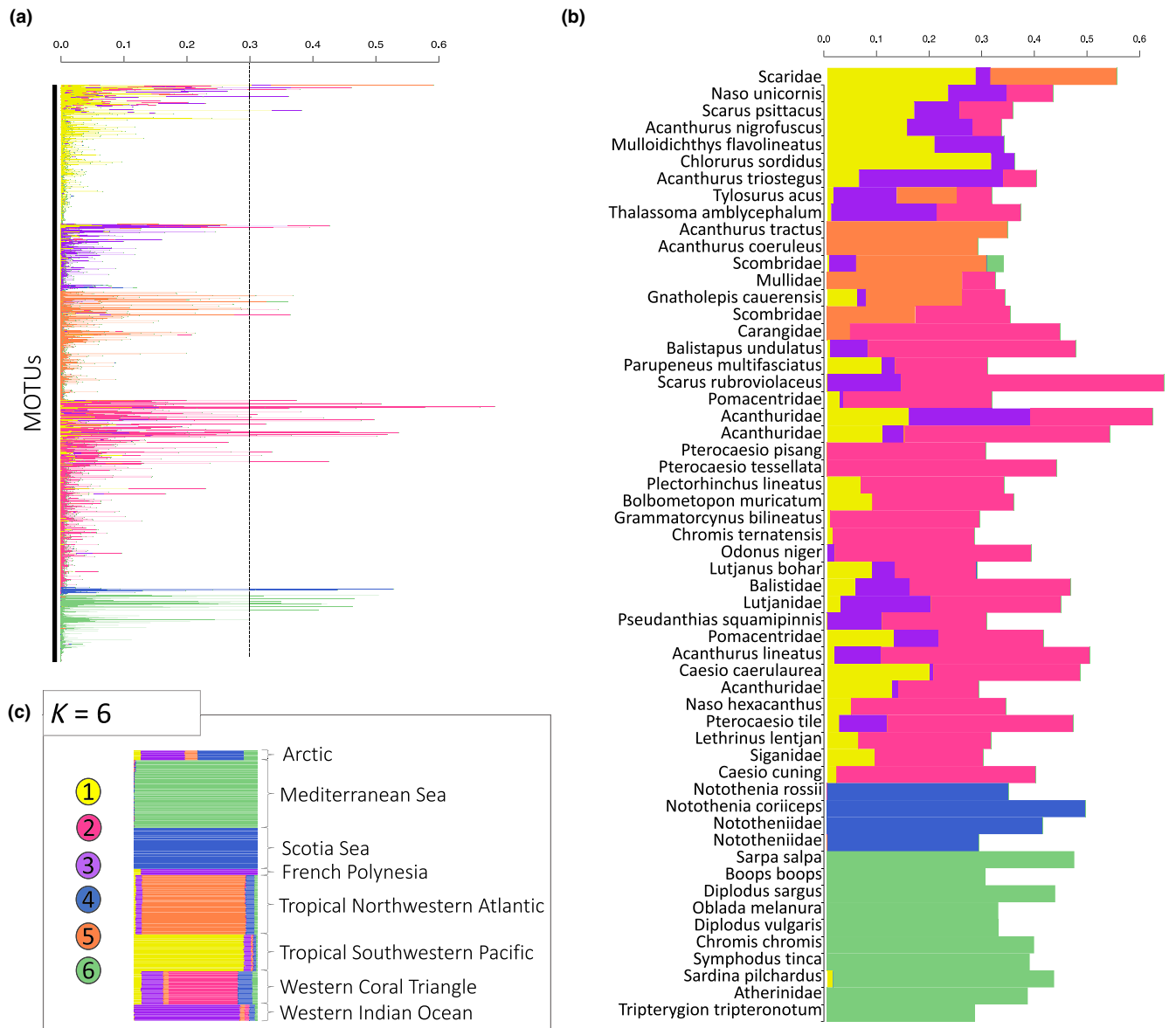
To benchmark our method on the representations obtained from the large-scale dataset, we performed a PCA followed by a  $k$ -means clustering on the most explanatory axes, with the number of clusters set to  $K=6$ . In the representation plane using the first two PCA axes (Figure S3a), we observed that the samples were primarily distributed along three directions, corresponding to the Mediterranean Sea, the northwestern tropical Atlantic region, and the Indo-Pacific region. Applying the  $k$ -means method to the first six PCA axes (cumulative explained variance 0.42) with  $K=6$ , we obtained distinct clusters corresponding to different biogeographic regions (Figure S3b). We compared the clustering results of the PCA and  $k$ -means method to those obtained with our new method. Both methods identified clusters that correspond to well-defined biogeographic regions (Figure 1; Figure S3). However, our new method provided a more nuanced and detailed delimitation of these regions, capturing finer-scale patterns of taxon distribution that were not as clearly distinguished by the PCA and  $k$ -means clustering alone.

### 3.4 | Assemblage analysis in the Mediterranean Sea

To explore whether our matrix factorization approach could be used at a regional scale, we applied our algorithm to a subset of the large dataset focusing on the Mediterranean Sea. By analysing the bar plots generated with  $K$  values spanning from 3 to 7 (Figure 4a), the algorithm revealed distinct pools that initially aligned with the geographical distribution of samples.

For  $K=3$ , the algorithm identified distinct geographical areas within the Mediterranean, including notable regions like the Balearic Islands, Corsica, and various coastal zones. Progressing to  $K=4$  and  $K=5$ , the algorithm provided finer characterizations, pinpointing different coastal areas such as Banyuls, Carry le Rouet, Riou, Porquerolles and Cap Roux in France (Figure 4b). The optimal value of  $K$  was reached for  $K=5$  (Figure S4a). In addition, a scree plot resulting from PCA performed on the raw data matrix, representing the probability of MOTU detection in PCR replicates for each sample, showed an elbow at the level of the PCA axes 4 and 6 (Figure S4b). Subsequently, at  $K=6$  and  $K=7$ , the algorithm detected variations in the taxa composition of proportions within each geographical zone (Figure 4c). On a regional scale, such as the Mediterranean Sea, the analysis of the M matrix for  $K=5$  did not reveal strong compositional MOTU characterizations concerning the sources as in the large-scale case (Figure S4c).

To test for a protection effect, that is the difference between samples from a reserve versus outside, we fitted GLM to examine



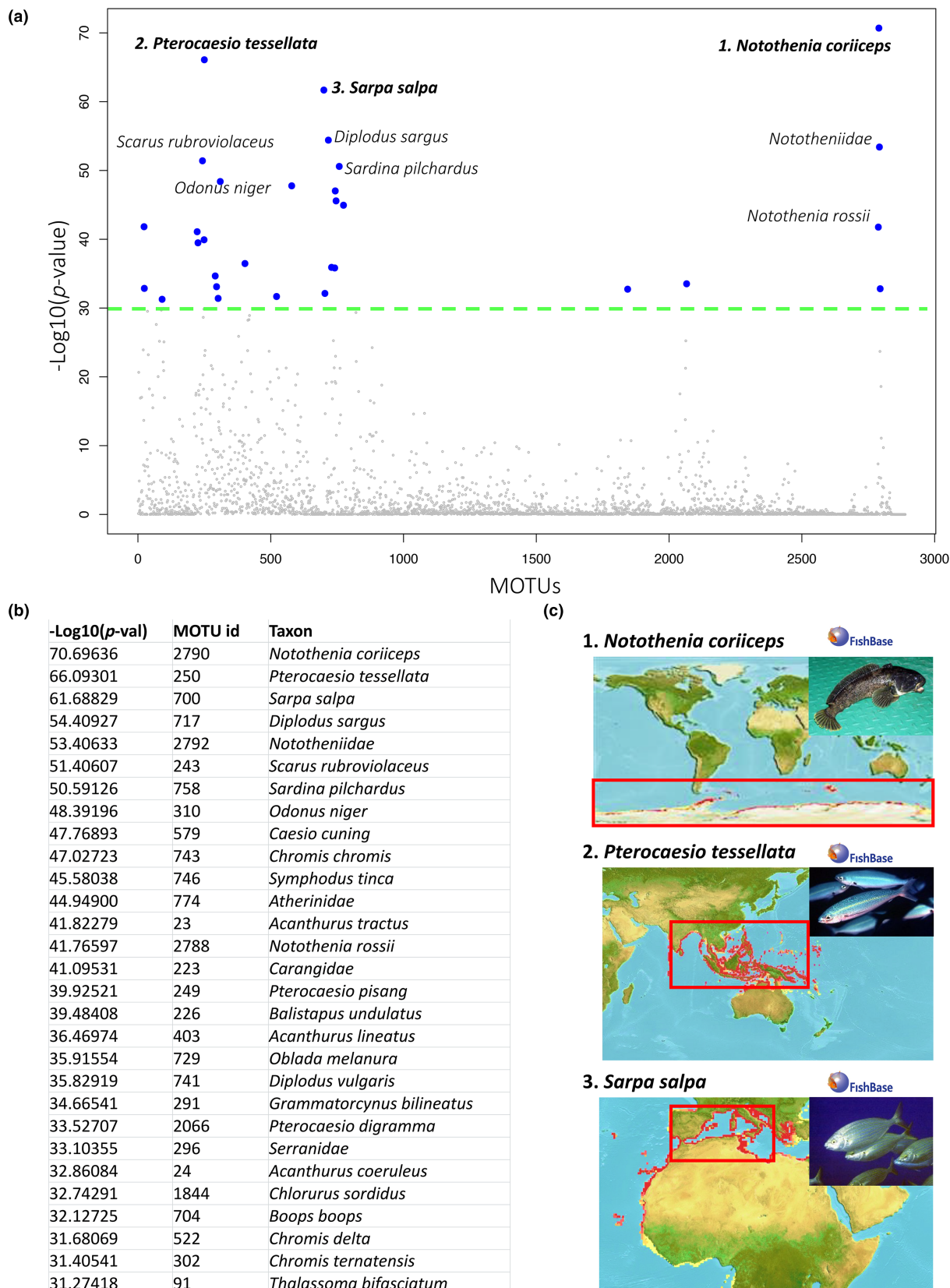
**FIGURE 2** Frequencies of molecular operational taxonomic units from environmental DNA samples in each source for the large-scale dataset: (a) Bar plot representing the frequencies of 2888 fish molecular operational taxonomic units (MOTUs) in each source for  $K=6$ . (b) Bar plot for a selection of MOTUs (56 out of 2888) with a cumulative frequency exceeding 30% in (a). For these selected MOTUs, we provide the best taxonomic assignments. (c) The colour association with sources is derived from the assemblage results presented in Figure 1, for which we report the bar plot for  $K=6$ .

the effect of protection level on the probability of belonging to one of the seven sources. The chi-squared test of GLM between the two levels of protection (reserve—non-reserve) and the combined probabilities of the seven sources by an ILR transformation had a significance value ( $p$ -value  $< 0.001$ ). The pseudo values of the GLM resulted in a McFadden pseudo  $R^2$  of 0.29 and a Cragg and Uhler's pseudo  $R^2$  of 0.41. By calculating the correlation between the MOTU richness in each sample and the probabilities combination by the ILR transformation of the  $S$  matrix for  $K=7$ , we obtained a value  $R^2$  of 0.72 ( $p$ -value  $< 0.001$ ). This finding confirmed that for  $K=7$ , the algorithm recognized the human gradient of protection among samples and the distribution of MOTU richness.

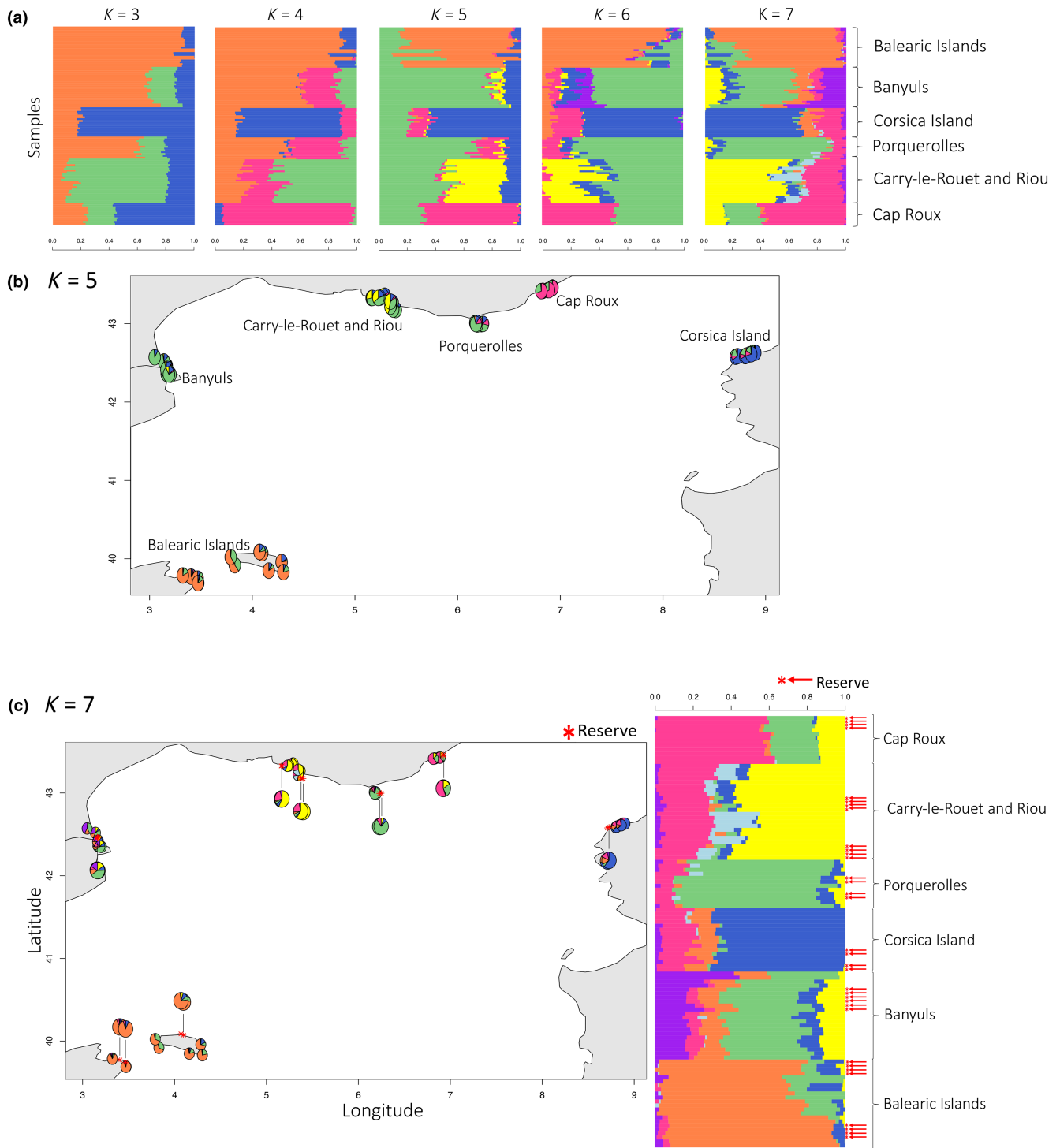
## 4 | DISCUSSION

Ecological systems are complex and characterized by numerous interactions between species and environmental factors (Riva et al., 2023). To better understand the structure of assemblages underpinning biodiversity distribution, effective methods are needed both in terms of data acquisition and data analysis. In this study, we introduced a new method for studying taxonomic unit assemblage composition, emphasizing the mixture that can exist between different pools of samples and taxa. This method is inspired by admixture analysis in population genetics and latent Dirichlet allocation (Pritchard et al., 2000; Valle et al., 2014). We applied the method to reanalyse a fish eDNA metabarcoding dataset at both large and





**FIGURE 3** Most differentiating molecular operational taxonomic units (MOTUs) detected in the large-scale environmental DNA dataset: (a) Plot of  $-\log_{10}(p\text{-values})$  for the large-scale dataset. The horizontal line represents an expected false discovery rate of  $q=10-30$ . The taxonomy of the nine most differentiated MOTUs is reported. (b) Table reporting the 29 MOTUs above the  $q$  threshold and their corresponding taxonomic identification. (c) Representation of the top three  $-\log_{10}(p\text{-values})$  and the corresponding fish distribution based on FishBase and Aquamaps. The geographic areas corresponding to our samples are highlighted in red rectangles.



**FIGURE 4** Mixture mapping and molecular operational taxonomic unit (MOTU) composition at a local scale: The Mediterranean Sea (a) Barplots illustrating the composition of fish samples for varying numbers of sources ( $K$ ) from 3 to 7 in the Mediterranean Sea. As  $K$  increases, the different pools are associated with the continent-island gradient ( $K=3$ ) and the latitudinal and longitudinal gradients ( $K=4-6$ ). (b) At  $K=5$ , distinct regions are identified: Balearic Islands (orange), Banyuls (purple), Carry-le-Rouet and Riou (yellow), Porquerolles (green), Cap Roux (pink) and Corsica (blue). (c) At  $K=7$ , the effect of protection within each specific geographic region is observed, with local variation in composition between samples from no-take reserves and those from fished areas within each region. On the left, in the geographic map, reserve samples are represented by larger circles, and their coordinates are indicated by asterisks. On the right, a bar plot for  $K=7$  shows samples sorted by latitude. Bars corresponding to samples from reserves are indicated by an arrow and an asterisk.

regional scales (Mathon et al., 2023). At large scale, the method was able to detect biogeographic sample pools and to associate a probabilistic MOTU composition with the biogeography of the respective

pool (Figure 1). We observed mixing between pools, revealing the complex dynamics of marine ecosystems and the fact that a MOTU can belong to several bioregions (Figure 2). By examining MOTU

compositions and their respective abundance in the sources, the model linked the taxa corresponding to the MOTUs to their respective geographic ranges (Figures 2 and 3). In the regional analysis of Mediterranean samples, our algorithm successfully identified biogeographic gradients such as the latitudinal gradient and the continent-island gradient. It also identified clusters of MOTUs corresponding to the level of protection for each sample (Figure 4).

The delineation of fish assemblages in different oceanic regions is influenced by a complex interaction of environmental, historical and ecological factors (Costello & Chaudhary, 2017; Deutsch et al., 2020; Stuart-Smith et al., 2017). In the eastern Pacific and tropical Atlantic, ocean currents, temperature gradients and geographic barriers form distinct communities adapted to local conditions (Bender et al., 2017; Pellissier et al., 2014). Similarly, Antarctic fish communities reflect the influence of polar currents and unique habitats like ice shelves (Crame, 2018; Eastman, 2005). The Mediterranean biodiversity stems from its geographic isolation and habitat diversity (Coll et al., 2010). In the Indo-Pacific, subtle differences in assemblages are due to the weak physical barriers to dispersal (Trembl et al., 2015), although areas of endemism persist (Kulbicki et al., 2013). Our method identified increasingly distinct sample pools with higher parameter values. On the large scale, regions like the Mediterranean and Tropical Northwestern Atlantic can be discriminated while at higher resolution the Indo-Pacific was split. This highlights the Indo-Pacific's vastness and heterogeneity, driving species evolution, diversification and adaptation (Cowman & Bellwood, 2013).

By dissecting MOTU frequencies within sources and classifying them into 'pure' and 'mixed' types, we identified relationships between MOTU composition, geographical regions and biodiversity (Figure 2). The taxonomic identification of MOTUs revealed how pure taxa are characteristic of areas such as the Mediterranean Sea (e.g., *Sparidae* like *Salpa Salpa*, *Boops boops* and *Diplodus sargus*) and Scotia Sea (e.g., *Nototheniidae*, like *Notothenia rossii* and *Notothenia corliceps*), and how mixed MOTUs are characteristic of those taxa from the tropical belt, mainly mixed taxa from the Indo-Pacific zone. (e.g., *Scaridae*, *Acanthuridae* and *Lutjanidae*) (Figure 3).

The comparison with PCA followed by *k*-means shows that our mixed representation is more effective in clustering the various samples. In the PCA (Figure S3), samples are stretched along PC axes, and some samples are assigned by the *k*-means to a cluster even if they are geometrically distant from the centre of their cluster. This behaviour is typical of mixed compositions, such as fish communities, where spatial connections, vastness and heterogeneity of environments promote mixing (van Denderen et al., 2015). Moreover, although the clustering largely matches the biogeographic regions, some samples within each biogeographic region of the large-scale dataset remain grouped in the cluster corresponding to the poles (Cluster 1 Figure 3). These samples correspond to near-zero coordinates in the PC1-PC2 plane, which have a very low contribution on the axes, resulting in their failure to be clustered into their respective biogeographic groups. The representation provided by PCA limits the identification of community data mixing and fails to adequately handle samples with near-zero coordinates. Our method

appears more appropriate for the study of community assemblages, allowing more effective analysis of both spatial and compositional relationships at individual sites.

To conserve biodiversity and understand the effects of conservation efforts on ecological systems, the study of species assemblages within specific regions is of great importance, particularly species responses to different levels of protection (Loiseau et al., 2021). Our investigation, focusing on a subset of the Mediterranean Sea, showed a significant association between protection levels and the distribution of MOTUs (McFadden pseudo  $R^2=0.29$ ,  $p$ -value  $<0.001$ ; Cragg and Uhler's pseudo  $R^2=0.41$ ,  $p$ -value  $<0.001$ ), illustrating the algorithm's capacity to discern human-induced protection gradients and their impact on MOTU composition ( $R^2=0.72$ ,  $p$ -value  $<0.001$ ). This result confirms several previous studies on similar data (Boulanger et al., 2021; Dalongeville et al., 2022; Lamperti et al., 2023) but has the merit to show the mixture characterizing each area and local conditions (Figure 4c). Our method thus proved to be an effective tool for understanding changes in species assemblages from eDNA metabarcoding data, allowing us to accurately track both environmental and human changes, aspects that are crucial for understanding the current impact of global change (Blowes et al., 2019) but also monitor restoration strategies across coastal habitats (Vozzo et al., 2023).

The method was then able to link different samples by mixing assemblage compositions from different sources. It provides a more accurate view of species assemblages in space than  $\beta$  diversity (Figure S5). The latter is able to describe pairwise dissimilarity between different samples (Loiseau et al., 2021) and, when combined with a reduction method such as PCoA, to cluster samples along gradients that are often difficult to interpret. In contrast, our method provides a clear multidimensional representation of individual samples described by the proportion of different sources. It also provides mixed frequencies on MOTUs, thereby offering a simultaneous visualization of the geographic and compositional relationships of the assemblages at each site. While  $\beta$  diversity is a measure of pairwise dissimilarity between samples, spatial matrix factorization is a sample-based assessment of similarity to all other samples, considering common sources.

The method proves to be an additional relevant tool for studying the composition of species assemblages. However, the selection of parameters can influence its outcomes, and identifying scattered samples can pose challenges. At large scale, Arctic samples often elude clear identification as a distinct pool unless the algorithm is adjusted to accommodate a larger number of sources, thereby introducing instability in the optimization process (Figure S1a). This phenomenon arises from the sparse nature of the samples, characterized by a high prevalence of null values and a limited number of samples and MOTUs.

In conclusion, our comprehensive analysis from large to regional scales provides original insights into the structuring of marine fish assemblages, the interaction between assemblages and taxonomic units, and the impact of human factors on biogeography. These results demonstrate that the model is more suitable for studying community assemblages, clearly highlighting the mixing that can

occur within an ecosystem and effectively visualizing the biodiversity structure both within and between samples. The importance of accurate assemblage identification could also have significant implications for ecological conservation efforts and biodiversity management. The method has the potential to derive relevant bioindicators based on the composition of taxonomic units to detect early signals of ecosystem shifts. With the arrival of massive eDNA data worldwide (Duarte et al., 2023; Mathon et al., 2023), future improvements should exploit the information contained in nucleotide sequences (Lamperti et al., 2023). Although the spatial matrix factorization method was tested on a large-scale fish eDNA dataset, it has broad applications in the study of biogeography and community composition with other types of data, such as abundance or detection data, species inventories, OTUs and ASVs.

### AUTHOR CONTRIBUTIONS

Letizia Lamperti, Olivier François, and Stéphanie Manel conceived the ideas and designed the methodology. Letizia Lamperti analysed the data. Letizia Lamperti wrote a first draft of the manuscript. Stéphanie Manel and Olivier François contributed substantially to the writing. All authors led the writing of the manuscript and critically to the drafts and gave final approval for publication.

### ACKNOWLEDGEMENTS

This work is part of the Horizon 2020-Marie Sklodowska-Curie Actions-COFUND project Artificial Intelligence for the Sciences (AI4theSciences) and the IA-Biodiv ANR project FISH-PREDICT (ANR-21-AAFI-0001-01). This project was partly funded by the WSL internal grant 'eDNA Proof' and the Swiss Data Science Centre 'DNai'. We would also like to thank Alice Valentini and Spygen for performing the data extraction analysis and contributing to the funding of this project.

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

### PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14430>.

### DATA AVAILABILITY STATEMENT

All eDNA data (except New Caledonia) are available in open access in Zenodo: <https://doi.org/10.5281/zenodo.7805935> (Mathon, 2023). New-Caledonia eDNA data are available from Zenodo upon request at: <https://doi.org/10.5281/zenodo.6381130> (Vigliola et al., 2022). Codes and scripts for reproducing the analyses in this manuscript are available at <https://doi.org/10.5281/zenodo.13709657> (Lamperti, 2024).

### ORCID

Letizia Lamperti  <https://orcid.org/0000-0001-8059-1354>

Laëtitia Mathon  <https://orcid.org/0000-0001-8147-8177>

Théophile Sanchez  <https://orcid.org/0000-0001-8571-0578>

Loïc Pellissier  <https://orcid.org/0000-0002-2289-8259>

Stéphanie Manel  <https://orcid.org/0000-0001-8902-6052>

### REFERENCES

- Barwell, L. J., Isaac, N. J. B., & Kunin, W. E. (2015). Measuring  $\beta$ -diversity with species abundance data. *Journal of Animal Ecology*, 84(4), 1112–1122. <https://doi.org/10.1111/1365-2656.12362>
- Bender, M. G., Leprieux, F., Mouillot, D., Kulbicki, M., Parravicini, V., Pie, M. R., Barneche, D. R., Oliveira-Santos, L. G. R., & Floeter, S. R. (2017). Isolation drives taxonomic and functional nestedness in tropical reef fish faunas. *Ecography*, 40(3), 425–435. <https://doi.org/10.1111/ecog.02293>
- Benestan, L., Fietz, K., Loiseau, N., Guerin, P. E., Trofimenko, E., Rühls, S., Schmidt, C., Rath, W., Biastoch, A., Pérez-Ruzafa, A., Baixauli, P., Forcada, A., Arcas, E., Lenfant, P., Mallol, S., Goñi, R., Velez, L., Höppner, M., Kininmonth, S., ... Manel, S. (2021). Restricted dispersal in a sea of gene flow. *Proceedings of the Royal Society B: Biological Sciences*, 288(1951), 20210458. <https://doi.org/10.1098/rspb.2021.0458>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Blowes, S. A., McGill, B., Brambilla, V., Chow, C. F. Y., Engel, T., Fontrodona-Eslava, A., Martins, I. S., McGlenn, D., Moyes, F., Sagouis, A., Shimadzu, H., van Klink, R., Xu, W. B., Gotelli, N. J., Magurran, A., Dornelas, M., & Chase, J. M. (2024). Synthesis reveals approximately balanced biotic differentiation and homogenization. *Science Advances*, 10(8), eadj9395. <https://doi.org/10.1126/sciadv.adj9395>
- Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., Moyes, F., Magurran, A., McGill, B., Myers-Smith, I. H., Winter, M., Bjorkman, A. D., Bowler, D. E., Byrnes, J. E. K., Gonzalez, A., Hines, J., Isbell, F., Jones, H. P., Navarro, L. M., ... Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. *Science*, 366(6463), 339–345. <https://doi.org/10.1126/science.aaw1620>
- Boullanger, E., Loiseau, N., Valentini, A., Arnal, V., Boissery, P., Dejean, T., Deter, J., Guellati, N., Holon, F., Juhel, J. B., Lenfant, P., Manel, S., & Mouillot, D. (2021). Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves. *Proceedings of the Royal Society B: Biological Sciences*, 288(1949), 20210112. <https://doi.org/10.1098/rspb.2021.0112>
- Boyer, F., Mercier, C., Bonin, A., le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, 16(1), 176–182. <https://doi.org/10.1111/1755-0998.12428>
- Caye, K., Deist, T. M., Martins, H., Michel, O., & François, O. (2016). TESS3: Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, 16(2), 540–548. <https://doi.org/10.1111/1755-0998.12471>
- Caye, K., Jay, F., Michel, O., & François, O. (2018). Fast inference of individual admixture coefficients using geographic data. *The Annals of Applied Statistics*, 12(1), 586–608. <https://www.jstor.org/stable/26542541>
- Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Ben Rais Lasram, F., Aguzzi, J., Ballesteros, E., Bianchi, C. N., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Froggia, C., Galil, B. S., Gasol, J. M., Gertwagen, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., ... Voultsiadou, E. (2010). The biodiversity of the Mediterranean Sea: Estimates, patterns, and threats. *PLoS One*, 5(8), e11842. <https://doi.org/10.1371/journal.pone.0011842>
- Costello, M. J., & Chaudhary, C. (2017). Marine biodiversity, biogeography, deep-sea gradients, and conservation. *Current Biology*, 27(11), R511–R527. <https://doi.org/10.1016/j.cub.2017.04.060>

- Costello, M. J., Claus, S., Dekeyser, S., Vandepitte, L., Tuama, É. Ó., Lear, D., & Tyler-Walters, H. (2015). Biological and ecological traits of marine species. *PeerJ*, 3, e1201. <https://doi.org/10.7717/peerj.1201>
- Cowman, P. F., & Bellwood, D. R. (2013). The historical biogeography of coral reef fishes: Global patterns of origination and dispersal. *Journal of Biogeography*, 40(2), 209–224. <https://doi.org/10.1111/jbi.12003>
- Crame, J. A. (2018). Key stages in the evolution of the Antarctic marine fauna. *Journal of Biogeography*, 45(5), 986–994. <https://doi.org/10.1111/jbi.13208>
- Dalongeville, A., Boulanger, E., Marques, V., Charbonnel, E., Hartmann, V., Santoni, M. C., Deter, J., Valentini, A., Lenfant, P., Boissery, P., Dejean, T., Velez, L., Pichot, F., Sanchez, L., Arnal, V., Bockel, T., Delaruelle, G., Holon, F., Milhau, T., ... Mouillot, D. (2022). Benchmarking eleven biodiversity indicators based on environmental DNA surveys: More diverse functional traits and evolutionary lineages inside marine reserves. *Journal of Applied Ecology*, 59(11), 2803–2813. <https://doi.org/10.1111/1365-2664.14276>
- Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872–5895. <https://doi.org/10.1111/mec.14350>
- Deutsch, C., Penn, J. L., & Seibel, B. (2020). Metabolic trait diversity shapes marine biogeography. *Nature*, 585(7826), 557–562. <https://doi.org/10.1038/s41586-020-2721-y>
- Duarte, S., Simões, L., & Costa, F. O. (2023). Current status and topical issues on the use of eDNA-based targeted detection of rare animal species. *Science of the Total Environment*, 904, 166675. <https://doi.org/10.1016/j.scitotenv.2023.166675>
- Duhamet, A., Albouy, C., Marques, V., Manel, S., & Mouillot, D. (2023). The global depth range of marine fishes and their genetic coverage for environmental DNA metabarcoding. *Ecology and Evolution*, 13(1), e9672. <https://doi.org/10.1002/ece3.9672>
- Eastman, J. T. (2005). The nature of the diversity of Antarctic fishes. *Polar Biology*, 28(2), 93–107. <https://doi.org/10.1007/s00300-004-0667-4>
- François, O. (2016). Running structure-like population genetic analyses with R. <https://www.bioconductor.org/>
- François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, 25(2), 454–469. <https://doi.org/10.1111/mec.13513>
- Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4), 973–983. <https://doi.org/10.1534/genetics.113.160572>
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4), 325–338. <https://doi.org/10.2307/2333639>
- Hillebrand, H., & Matthiessen, B. (2009). Biodiversity in a complex world: Consolidation and progress in functional biodiversity research. *Ecology Letters*, 12(12), 1405–1419. <https://doi.org/10.1111/j.1461-0248.2009.01388.x>
- Ji, L., Sheng, S., Shen, F., Yang, L., Wen, S., He, G., Wang, N., Wang, X., & Yang, L. (2024). Stochastic processes dominated the soil bacterial community assemblages along an altitudinal gradient in boreal forests. *Catena*, 237, 107816. <https://doi.org/10.1016/j.catena.2024.107816>
- Johnston, F. D., Simmons, S., van Poorten, B., & Venturelli, P. (2021). Comparative analyses with conventional surveys reveal the potential for an angler app to contribute to recreational fisheries monitoring. *Canadian Journal of Fisheries and Aquatic Sciences*, 79(1), 31–46. <https://doi.org/10.1139/cjfas-2021-0026>
- Jouffray, J.-B., Blasiak, R., Norström, A. v., Österblom, H., & Nyström, M. (2020). The blue acceleration: The trajectory of human expansion into the ocean. *One Earth*, 2(1), 43–54. <https://doi.org/10.1016/j.oneear.2019.12.016>
- Juhel, J.-B., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, Pouyoud, L., Dejean, T., Mouillot, D., & Hocdé, R. (2020). Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proceedings of the Royal Society B: Biological Sciences*, 287(1930), 20200248. <https://doi.org/10.1098/rspb.2020.0248>
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27. <https://doi.org/10.1007/BF02289565>
- Kulbicki, M., Parravicini, V., Bellwood, D. R., Arias-González, E., Chabanet, P., Floeter, S. R., Friedlander, A., McPherson, J., Myers, R. E., Vigliola, L., & Mouillot, D. (2013). Global biogeography of reef fishes: A hierarchical quantitative delineation of regions. *PLoS One*, 8(12), e81847. <https://doi.org/10.1371/journal.pone.0081847>
- Lamperti, L. (2024). letizialamperti/Spatial\_Matrix\_factorization\_Method\_Mixture\_Analysis: TESS3forEcology (Versione TESS3forEcology). Zenodo. <https://doi.org/10.5281/zenodo.13709658>
- Lamperti, L., Sanchez, T., Si Moussi, S., Mouillot, D., Albouy, C., Flück, B., Bruno, M., Valentini, A., Pellissier, L., & Manel, S. (2023). New deep learning-based methods for visualizing ecosystem properties using environmental DNA metabarcoding data. *Molecular Ecology Resources*, 23(8), 1946–1958. <https://doi.org/10.1111/1755-0998.13861>
- Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791. <https://doi.org/10.1038/44565>
- Legendre, P., & Legendre, L. (1998). *Numerical ecology*. Elsevier Science & Technology.
- Leibold, M. A., Govaert, L., Loeuille, N., de Meester, L., & Urban, M. C. (2022). Evolution and community assembly across spatial scales. *Annual Review of Ecology, Evolution, and Systematics*, 53(1), 299–326. <https://doi.org/10.1146/annurev-ecolsys-102220-024934>
- Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., ten Hoopen, P., Vaughan, R., ... Cochrane, G. (2011). The European nucleotide archive. *Nucleic Acids Research*, 39(suppl\_1), D28–D31. <https://doi.org/10.1093/nar/gkq967>
- Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 112(7), 2076–2081. <https://doi.org/10.5883/DS-ARMS>
- Loiseau, N., Thuiller, W., Stuart-Smith, R. D., Devictor, V., Edgar, G. J., Velez, L., Cinner, J. E., Graham, N. A. J., Renaud, J., Hoey, A. S., Manel, S., & Mouillot, D. (2021). Maximizing regional biodiversity requires a mosaic of protection levels. *PLoS Biology*, 19(5), e3001195. <https://doi.org/10.1371/journal.pbio.3001195>
- MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M. S., Ducar, M. D., Meyerson, M., & Thorner, A. R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, 19(1), 30. <https://doi.org/10.1186/s12864-017-4428-5>
- Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, 2, e593. <https://doi.org/10.7717/peerj.593>
- Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., Brennan, G., Bush, A., Canard, E., Cordier, T., Creer, S., Curry, R. A., David, P., Dumbrell, A. J., Gravel, D., Hajibabaei, M., Hayden, B., van der Hoorn, B., Jarne, P., ... Bohan, D. A. (2020). Key questions for next-generation biomonitoring. *Frontiers in Environmental Science*, 7. <https://doi.org/10.3389/fenvs.2019.00197>

- Marques, V., Guérin, P. É., Rocle, M., Valentini, A., Manel, S., Mouillot, D., & Dejean, T. (2020). Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. *Ecography*, 43(12), 1779–1790. <https://doi.org/10.1111/ecog.05049>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17(1), 10–12. <https://doi.org/10.14806/ej.17.1.200>
- Mathon, L., Marques, V., Manel, S., Albouy, C., Andreollo, M., Boulanger, E., Deter, J., Hocdé, R., Leprieur, F., Letessier, T. B., Loiseau, N., Maire, E., Valentini, A., Vigliola, L., Baletaud, F., Bessudo, S., Dejean, T., Faure, N., Guerin, P. E., ... Mouillot, D. (2023). The distribution of coastal fish eDNA sequences in the Anthropocene. *Global Ecology and Biogeography*, 32(8), 1336–1352. <https://doi.org/10.1111/geb.13698>
- Mathon, L., Marques, V., Mouillot, D., Albouy, C., Andreollo, M., Baletaud, F., Borrero-Pérez, G. H., Dejean, T., Edgar, G. J., Grondin, J., Guerin, P. E., Hocdé, R., Juhel, J. B., Kadarusman, Maire, E., Mariani, G., McLean, M., Polanco Fernandez, A., Pouyau, L., ... Manel, S. (2022). Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding. *Proceedings of the Royal Society B: Biological Sciences*, 289, 20220162.
- Mathon, L. (2023). MEGAFAUNA\_Pole2Pole\_eDNA\_data [Data set]. Zenodo, <https://doi.org/10.5281/zenodo.7805935>
- Mayor, S., Allan, E., Altermatt, F., Isbell, F., Schaepman, M. E., Schmid, B., & Niklaus, P. A. (2024). Diversity–functioning relationships across hierarchies of biological organization. *Oikos*, 2024(1), e10225. <https://doi.org/10.1111/oik.10225>
- McClenaghan, B., Compson, Z. G., & Hajibabaei, M. (2020). Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA. *PLoS One*, 15(3), 1–17. <https://doi.org/10.1371/journal.pone.0224119>
- Mori, A. S., Isbell, F., & Seidl, R. (2018).  $\beta$ -diversity, community assembly, and ecosystem functioning. *Trends in Ecology & Evolution*, 33(7), 549–564. <https://doi.org/10.1016/j.tree.2018.04.012>
- Nelson, J. S. (2016). Fishes of the world. In *Fishes of the world* (pp. i–xlii). John Wiley & Sons. <https://doi.org/10.1002/9781119174844.fmatter>
- Pellissier, L., Leprieur, F., Parravicini, V., Cowman, P. F., Kulbicki, M., Litsios, G., Olsen, S. M., Wisz, M. S., Bellwood, D. R., & Mouillot, D. (2014). Quaternary coral reef refugia preserved fish diversity. *Science*, 344(6187), 1016–1019. <https://doi.org/10.1126/science.1249853>
- Podani, J., & Schmera, D. (2011). A new conceptual and methodological framework for exploring and explaining pattern in presence–Absence data. *Oikos*, 120(11), 1625–1638. <https://doi.org/10.1111/j.1600-0706.2011.19451.x>
- Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., Roset, N., Schabuss, M., Zornig, H., & Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports*, 8(1), 10361. <https://doi.org/10.1038/s41598-018-28424-8>
- Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *AJHG*, 67(1), 170–181.
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Riva, F., Graco-Roza, C., Daskalova, G., Hudgins, E., Lewthwaite, J., Newman, E., Ryo, M., & Mammola, S. (2023). Toward a cohesive understanding of ecological complexity. *Science Advances*, 9(25), eabq4207. <https://www.science.org>
- Rognes, T. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, 4, e2584.
- Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated—Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, 15(6), 1289–1303. <https://doi.org/10.1111/1755-0998.12402>
- Sommeria-Klein, G., Watteaux, R., Ibarbalz, F. M., Pierella Karlusich, J. J., Iudicone, D., Bowler, C., & Morlon, H. (2021). Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science*, 374(6567), 594–599. <https://doi.org/10.1126/science.abb3717>
- Stuart-Smith, R. D., Edgar, G. J., & Bates, A. E. (2017). Thermal limits to the geographic distributions of shallow-water marine species. *Nature Ecology & Evolution*, 1(12), 1846–1852. <https://doi.org/10.1038/s41559-017-0353-x>
- Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press. <https://doi.org/10.1093/oso/9780198767220.001.0001>
- ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5), 1167–1179. <https://doi.org/10.2307/1938672>
- Tremblay, E. A., Ford, J. R., Black, K. P., & Swearer, S. E. (2015). Identifying the key biophysical drivers, connectivity outcomes, and meta-population consequences of larval dispersal in the sea. *Movement Ecology*, 3(1), 17. <https://doi.org/10.1186/s40462-015-0045-6>
- Uthicke, S., Robson, B., Doyle, J. R., Logan, M., Pratchett, M. S., & Lamare, M. (2022). Developing an effective marine eDNA monitoring: eDNA detection at pre-outbreak densities of corallivorous seastar (*Acanthaster cf. solaris*). *Science of the Total Environment*, 851, 158143. <https://doi.org/10.1016/j.scitotenv.2022.158143>
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4), 929–942. <https://doi.org/10.1111/mec.13428>
- Valle, D., Baiser, B., Woodall, C. W., & Chazdon, R. (2014). Decomposing biodiversity data using the latent Dirichlet allocation model, a probabilistic multivariate statistical method. *Ecology Letters*, 17(12), 1591–1601. <https://doi.org/10.1111/ele.12380>
- van Denderen, P. D., Bolam, S. G., Hiddink, J., Jennings, S., Kenny, A., Rijnsdorp, A., & van Kooten, T. (2015). Similar effects of bottom trawling and natural disturbance on composition and function of benthic communities across habitats. *Marine Ecology Progress Series*, 541, 31–43. <https://doi.org/10.3354/meps11550>
- Vigliola, L., Baletaud, F., & Mathon, L. (2022). Reef30\_New\_Caledonia\_eDNA [data set]. Zenodo, <https://doi.org/10.5281/zenodo.6381130>
- Villéger, S., Brosse, S., Mouchet, M., Mouillot, D., & Vanni, M. J. (2017). Functional ecology of fish: Current approaches and future challenges. *Aquatic Sciences*, 79(4), 783–801. <https://doi.org/10.1007/s00027-017-0546-z>
- Vozzo, M. L., Doropoulos, C., Silliman, B. R., Steven, A., Reeves, S. E., ter Hofstede, R., van Koningsveld, M., van de Koppel, J., McPherson, T., Ronan, M., & Saunders, M. I. (2023). To restore coastal marine areas, we need to work across multiple habitats simultaneously. *Proceedings of the National Academy of Sciences of the United States of America*, 120(26), e2300546120. <https://doi.org/10.1073/pnas.2300546120>
- Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, 21(2/3), 213–251. <https://doi.org/10.2307/1218190>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**Figure S1:** Cross-validation score plot (a) and scree plot (b) from the PCA on the large-scale eDNA data.

**Figure S2:** Geographical sampling areas for selection of the geographical map of Duhamet et al. (2023).

**Figure S3:** (a) Representation of the first two axes of the PCA on the large-scale eDNA data matrix. The points are coloured based on the k-means clustering performed on all the first six PCA components. (b) Geographical representation of the different points according to the k-means clustering performed on the first six PCA components.

**Figure S4:** Cross-validation score plot (a) and scree plot (b) from the PCA on the local eDNA data, the Mediterranean Sea. In (c) we reported the MOTU proportions for  $K=5$  in the Mediterranean Sea (ref. Figure 4b). In contrast to the large-scale case, the proportions represented in the Mediterranean Sea seem to be more mixed together, without having a clear proportion of one source.

**Figure S5:** Comparison diagram illustrating the analyses we can obtain from a community matrix using the spatial method of matrix factorization (left) and using beta diversity and PCoA (right).

**Appendix S1:** Algorithm and computing pipeline.

**How to cite this article:** Lamperti, L., François, O., Mouillot, D., Mathon, L., Sanchez, T., Albouy, C., Pellissier, L., & Manel, S. (2024). A spatial matrix factorization method to characterize ecological assemblages as a mixture of unobserved sources: An application to fish eDNA surveys. *Methods in Ecology and Evolution*, 00, 1–15. <https://doi.org/10.1111/2041-210X.14430>