

REVIEW ARTICLE

R-programming: A bioinformatic tool bridging the eDNA metabarcoding and its applications in understanding the ecological complexities

Mohd Imran^{1,*}, Mohammad Afzal²

ABSTRACT: Understanding ecosystem structure and functioning requires the integration of complex interactions among species across multiple trophic levels. Recent advances in environmental DNA (eDNA) metabarcoding coupled with next-generation sequencing (NGS) have revolutionized biodiversity assessment by enabling rapid, non-invasive, and comprehensive detection of species assemblages from environmental samples. However, the enormous volume of sequence data generated through NGS requires robust bioinformatics and statistical tools for effective interpretation. The R programming language, together with its graphical user interface RStudio, has emerged as a powerful platform for analysing and visualising eDNA metabarcoding data. Specialized R packages, including vegan, phyloseq, microDecon, betapart, stats, and mgcv, facilitate sequence processing, diversity estimation, community composition analysis, ecological modelling, and multivariate ordination techniques such as Principal Component Analysis (PCA), Principal Coordinates Analysis (PCoA), Non-metric Multidimensional Scaling (NMDS), and Redundancy Analysis (RDA). These analytical tools enable researchers to quantify biodiversity, examine spatial and temporal patterns in species assemblages, and explore relationships between community composition and environmental variables. By transforming complex sequence datasets into meaningful statistical and graphical representations, R has become an indispensable tool in modern eDNA-based ecological research, supporting reproducible analyses, biodiversity monitoring, conservation planning, and ecosystem management.

Keywords: Environmental DNA, Next Generation Sequencing, R Programming, Biodiversity

Received: 11 March 2026; Revised: 23 March 2026; Accepted: 03 July 2026; Published Online: 20 July 2026

1. INTRODUCTION

Ecology, the science of ecosystems, explores the processes that determine the functions, productivity and maintenance of ecosystems. Interactions which are readily seen involve predator-prey food-chains, like in the savannah, or the balance between carnivores, birds, and herbivores in a forest habitat. Others are more elusive, occurring through various

trophic levels and involving a range of organisms from microorganisms, to elusive species, to macroinvertebrates. Elaborating a comprehension of an ecosystem thus necessitates the gathering of assemblies of species and the interactions between them together in a single analytical space. With high throughput sequencing and environmental DNA (eDNA) sampling, such assemblies of species can now be obtained, through eDNA metabarcoding [1, 2]. eDNA sampling involves detecting genetic material which is shed into the environment by resident organisms, through mucus, saliva, faeces, urine, gametes, or skin [3]. Coupled to next generation sequencing (NGS), eDNA metabarcoding allows for the simultaneous sequencing and identification of species in a single environmental sample, offering speed, ease of sampling, non-destructiveness and wide taxonomic coverage [4]. The technique involves obtaining millions of short-read

¹ Department of Zoology, Jamia Urdu College of Education, Raja Mahendra Pratap Singh University, Aligarh, U.P. India.

² Section of Genetics, Department of Zoology, Aligarh Muslim University, Aligarh, U.P. India.

* Author to whom correspondence should be addressed:
imranzoology.amu@gmail.com (M. Imran)

DNA sequences which are typically between 100 and 600 base pairs in length [5] which are compared to online databases such as BOLD [6] and GenBank [7] to assign species identity.

The conceptual flow of the eDNA sequencing process, from the field site to analysis is illustrated in Figure 1, which depicts a representative terrestrial-freshwater ecosystem in which mammals, birds, amphibians, fishes, macroinvertebrates, and microorganisms continuously shed skin cells, mucus, faeces, urine, gametes, etc. into the water, sediment or soil, which are then collected with environmental samples of water, sediment or soil. The figure traces the process of eDNA sampling and extraction through to library preparation, bioinformatics, taxonomic assignment, and ecological analysis, noting that each step builds on the previous, and that the final analytical step is where R-based tools become essential. By showing the sources of eDNA alongside the analytic pipeline, the figure demonstrates that the transition from organism biology to statistical inference is a continuous process.

The amount of data generated from eDNA metabarcoding presents significant analytical challenges. In order to transform the raw reads into an interpretable account of species assembly, spatial and temporal variation, and relationships with the environment, command based bioinformatics and statistical tools are necessary. The R

programming language, accessed via the RStudio interface, is the key player in this arena due to the fact that it allows one to tackle sequence handling, diversity estimation, multivariate ordination, etc., in a single, reproducible environment. Packages such as tidyverse [8] and ggplot2 [9] provide the data handling and visualization backbone while domain-specific packages are used, depending on the analytical needs. The vegan package is used to calculate species richness, Shannon and Simpson diversity, beta diversity and species accumulation curves [10]. phyloseq allows the integration of taxonomic abundance data, with the associated sample metadata and phylogenetic information [11]. microDecon is used for the filtration (removal of contamination signals) of reads from metabarcoding datasets [12]. Betapart is used for the partitioning of beta diversity into turnover and nestedness components, for taxonomic, functional, and phylogenetic data [13]. The stats package provides basic statistical modelling [14], while mgcv allows for generalized additive models [15]. The bioseq package is used to allow the handling of DNA, RNA, and protein sequences within R [16]. Outside of R-specific packages, the NCBI BLAST+ command line tools can be invoked from within R to perform taxonomic assignment, which would not be practical to do in web interfaces. Thus sequence identification and statistical analysis can proceed in a single workflow [17].

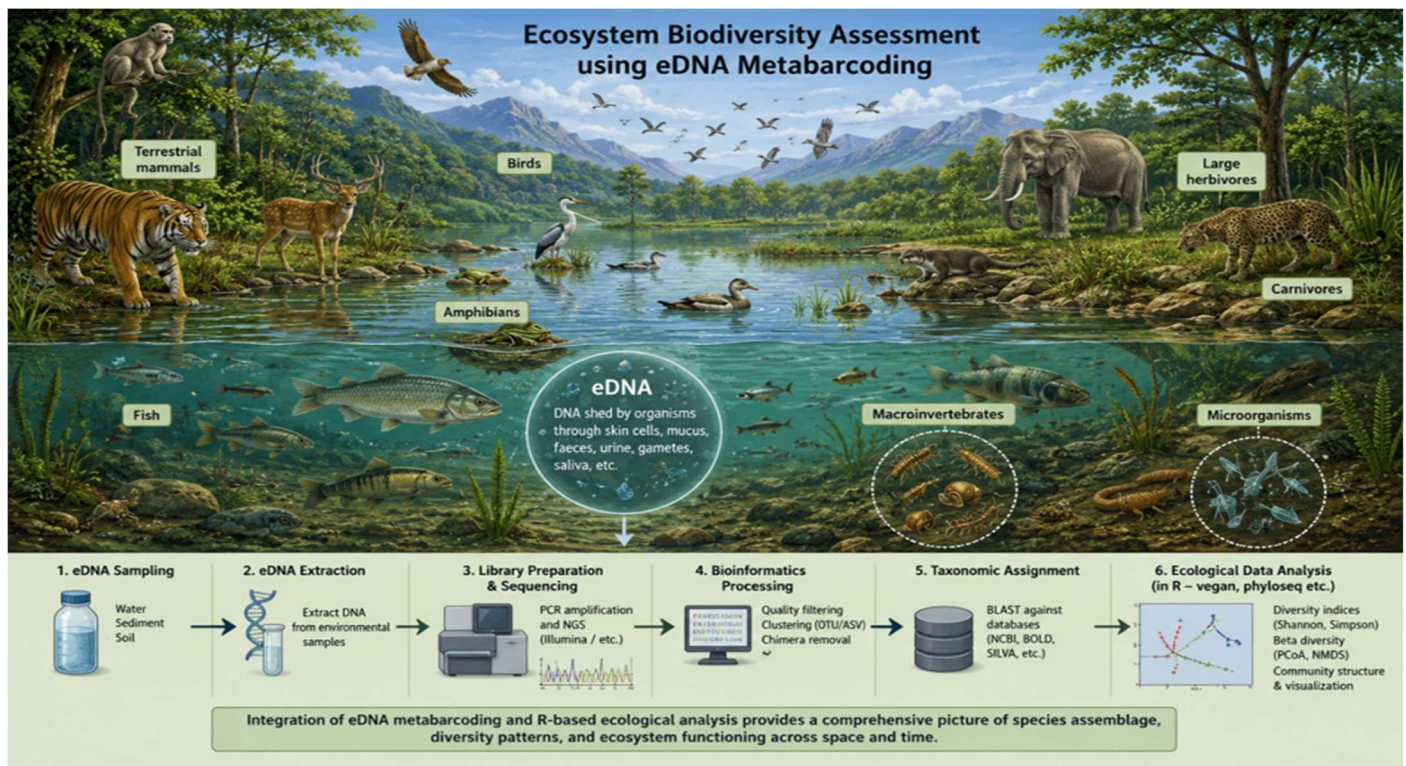


Fig. 1. Conceptual illustration of ecosystem biodiversity assessment using environmental DNA (eDNA) metabarcoding integrated with R programming-based ecological analysis. The figure depicts a representative terrestrial–freshwater ecosystem comprising diverse taxonomic groups, including terrestrial mammals, birds, amphibians, fishes, macroinvertebrates, and microorganisms, all of which continuously release environmental DNA (eDNA) into their surroundings through skin cells, mucus, faeces, urine, gametes, and other biological materials, which is collected with environmental samples of water, sediment, or soil.

Taxonomic assignment is dependent on reference databases, which differ in coverage, marker genes and intended applications. Table 1 shows the major databases used in eDNA metabarcoding studies, including NCBI GenBank, BOLD, SILVA, SILVAngs, UNITE, PR2, MIDORI, MitoFish, FishCARD, GTDB, Greengenes2, RDP, EMBL-ENA, DDBJ, WoRMS, GBIF, OBIS and iBOL. The table shows that some, such as GenBank, EMBL-ENA, and DDBJ are general nucleotide repositories which cover all organisms and markers, while others are taxon or marker-specific. BOLD and iBOL focus on DNA barcodes, SILVA and SILVAngs on ribosomal RNA, UNITE on fungal ITS, PR2 on protist 18S, MIDORI on mitochondrial metazoan genes, MitoFish and FishCARD on fish 12S and COI, GTDB, Greengenes2, and RDP on microbial taxonomy. WoRMS, GBIF, and OBIS provide taxonomic and occurrence information not related to sequence references. Knowing the database differences is important since this determines the accuracy of assignment, rates of false positives, and the taxonomic resolution one can obtain for downstream analysis, especially for invertebrates. Comparative evaluations have shown that invertebrate detection rates can differ, depending on whether BOLD or GenBank is used for COI assignment, with BOLD often having greater taxonomic richness for invertebrates [3]. Likewise, there are persistent gaps and errors in public databases, which can limit assignment accuracy for less well-known groups of organisms [13]. Thus, the use of curated, region specific reference sets is essential.

A typical eDNA metabarcoding pipeline proceeds

through extraction, library preparation, sequencing, read filtration, clustering or amplicon sequencing variant (ASV) inference, taxonomic assignment against reference databases, and ecological analysis. The choice of marker gene influences the taxonomic scope of the data, where mitochondrial 12S ribosomal DNA is widely used for vertebrate detection [18, 19] while the mitochondrial cytochrome c oxidase subunit I (COI) gene is common for macroinvertebrate assessment [20, 21]. Outputs from external platforms such as SILVAngs (NGS analysis of ribosomal DNA) can be imported into R for downstream community analysis [22-28]. Inside of R, alpha diversity is often estimated using the Shannon index to capture both species richness and evenness, while beta diversity is often evaluated through Jaccard pairwise dissimilarity for presence-absence data or Bray Curtis dissimilarity for abundance data. Ordination approaches such as Principal Component Analysis (PCA), Principal Coordinates Analysis (PCoA), Non-metric Multidimensional Scaling (NMDS), and Redundancy Analysis (RDA) allow researchers to visualize community patterns and relate them to environmental gradients. For example, PCA is applied to Hellinger-transformed abundance data, PCoA and NMDS to dissimilarity matrices, and RDA constrains community variation to measured environmental variables. These approaches, coupled to diversity partitioning and statistical modelling transforms high dimensional sequence data into graphical and quantitative output, which helps in the interpretation of ecological data.

Table 1. Major reference databases used in environmental DNA (eDNA) metabarcoding studies for sequence identification, taxonomic assignment, and biodiversity analysis. The databases differ in their taxonomic coverage, target genetic markers, and analytical applications.

Database	Primary Purpose	Common Marker Genes	Major Taxonomic Coverage	URL
NCBI GenBank	Global nucleotide sequence repository	COI, 12S, 16S, ITS, rbcL, matK, 18S	All organisms	https://www.ncbi.nlm.nih.gov/genbank/
BOLD Systems (Barcode of Life Data System)	DNA barcode identification	COI, rbcL, matK	Animals, plants, fungi	https://www.boldsystems.org/
SILVA	Curated ribosomal RNA database	16S, 18S rRNA	Bacteria, Archaea, Eukaryotes	https://www.arb-silva.de/
SILVAngs	Online NGS analysis platform using SILVA	16S, 18S	Microbial and eukaryotic communities	https://www.arb-silva.de/ngs/

UNITE	Fungal reference database	ITS	Fungi	https://unite.ut.ee/
PR2 (Protist Ribosomal Reference Database)	Protist taxonomy	18S rRNA	Protists	https://pr2-database.org/
MIDORI	Curated mitochondrial metazoan database	COI, 12S, 16S, Cyt b	Animals	https://www.reference-midori.info/
MitoFish	Fish mitochondrial genome database	12S, COI, complete mtDNA	Fishes	https://mitofish.aori.u-tokyo.ac.jp/
FishCARD	Curated fish reference database for eDNA	12S	Freshwater and marine fishes	https://fishcard.org/
Genome Taxonomy Database (GTDB)	Standardized microbial taxonomy	Whole genomes, 16S	Bacteria, Archaea	https://gtdb.ecogenomic.org/
Greengenes2	Microbial taxonomy database	16S rRNA	Bacteria and Archaea	https://github.com/biocore/greengenes2
RDP (Ribosomal Database Project)	Taxonomic classification	16S, 28S	Bacteria and fungi	https://rdp.cme.msu.edu/
EMBL-ENA (European Nucleotide Archive)	Nucleotide sequence repository	All markers	All organisms	https://www.ebi.ac.uk/ena
DDBJ	DNA Data Bank of Japan	All markers	All organisms	https://www.ddbj.nig.ac.jp/
WoRMS (World Register of Marine Species)	Marine taxonomy	Taxonomic reference	Marine organisms	https://www.marinespecies.org/
GBIF (Global Biodiversity Information Facility)	Species occurrence and distribution	Biodiversity records	All taxa	https://www.gbif.org/
OBIS (Ocean Biodiversity Information System)	Marine biodiversity records	Species occurrence	Marine organisms	https://obis.org/
iBOL (International Barcode of Life)	Global DNA barcoding initiative	COI and others	Animals, plants, fungi	https://ibol.org/

The novelty of this review lies in the treatment of the R programming ecosystem as a cohesive analytical framework for eDNA metabarcoding, rather than as a collection of tools. Previous methodological reviews often focus on laboratory protocols, sequencing platforms, or limitations of reference databases. This article aggregates the elements of sequence processing, contamination removal, diversity estimation, beta diversity partitioning, multivariate ordination, and constrained analysis, in a single space, showing how they are interconnected. The article also highlights the role of reproducible, code-based workflows in ensuring that eDNA-derived biodiversity assessments are transparent and repeatable, and the place of reproducibility in modern eDNA research.

In this review article, we report on the analytical landscape of R programming for eDNA metabarcoding data, covering the packages, statistical approaches, and ordination approaches that enable ecological interpretation. We describe the progression from raw reads through taxonomic assignment to diversity analysis and community-environment modelling, highlighting the contribution of *vegan*, *phyloseq*, *microDecon*, *betapart*, *stats*, and *mgcv*. We also discuss the application of PCA, PCoA, NMDS, and RDA to eDNA datasets, and what ecological questions can be addressed with these techniques. We conclude by considering how R-based workflows support biodiversity monitoring, conservation planning, and ecosystem management, and why reproducibility and analytical transparency have become central concerns in modern eDNA research.

2. eDNA AND R-PROGRAMMING

The sequence readout produced by next-generation sequencing of environmental DNA (eDNA) and containing the genetic information of hundreds or thousands of invertebrate and vertebrate species is challenging to work with. An individual environmental sample can produce millions of reads in total belonging to hundreds or thousands of species, and the information on what species belong to the sample is not evident from the data. Interpreting these data as a snapshot of a larger ecological picture requires a descriptive programming language that can extract relevant information from the DNA sequences. R programming language is an essential tool for this task, serving as a bridge between DNA sequencing and statistical/graphical analysis, performing calculations necessary to turn these reads into an organized description of species composition, distribution, and interactions.

In order to understand an ecosystem, it is paramount to understand the species that comprise it, as interactions between species are dictated by the species composition of the ecosystem. The species composition is not fixed, varying with environmental factors on the spatiotemporal scale larger than the one under the study, and therefore, should be averaged out across many measurements to better characterize the underlying ecological trends. While eDNA

sequencing presents enormous opportunities to learn about the species present in an ecosystem, these opportunities are largely hindered by the need for statistical analysis of the sequencing readouts to visualize the data. As such, bioinformatics plays a major role in analyzing the sequences, and the choice of what analytical platform to use affects the ability to interpret the results.

2.1. R packages

R is a coded language that has found its niche in statistical analysis, computation, and visualization [29]. In particular, its versatility and broad adoption within the scientific community have made it a popular choice for data analysis in ecological and biodiversity assessment. Its syntax is based on writing out commands in a certain format that R interprets as instructions to execute or analyze. As a programming language, R has the capacity to record a set of performed operations as a set of instructions or a script that can be later re-run on the same or similar data. Figure 2 below is an example of RStudio, one of the platforms that allow to write and execute R scripts. Script writing, execution, and analysis are performed in the console, while plots are produced in a separate plotting window. Other features include the script editor, where written code is organized, and the environment, where objects saved in the R memory can be accessed. Together, these features facilitate an iterative process of data analysis, where the raw readouts from sequencing are processed step-by-step until a desired level of graphical representation is reached. R packages serve as the core of these operations, providing the base for data manipulation, statistical analysis, and visualization. The tidyverse package [30] and its component ggplot2 [31] provide the basic framework of statistical analysis and plotting, which can be built upon using other data-specific packages. In particular, data wrangling and analysis are separated into individual packages that address specific aspects of the data analysis. For example, the bioseq package is used for working with DNA, RNA, and amino acid sequences [32]. The vegan package is used for calculating biodiversity indices and other community ecology-related statistical procedures [33]. Phyloseq [34] is another popular package used in analyzing phylogenetic sequences, providing a framework for analysis that incorporates taxonomic databases, diversity indices, and phylogenetic trees. The ability to analyze the readouts in terms of taxonomy and phylogeny is particularly useful in interpreting the results of eDNA sequencing, where the data are inherently linked to taxonomy at the level of sequence identification.

Statistical analysis in R is performed using a combination of core stats package that implements linear models and hypothesis testing [35] and the mgcv package that estimates generalized additive models, allowing to estimate non-linear trends using linear modeling procedures [36]. Similarly, the beta diversity partitioning implemented in the betapart package [37] can help to dissect the patterns in species turnover, highlighting the trends in nestedness,

which occur due to differences in the number of species, and turnover, which reflects the changes in species composition between different sites. The package is particularly useful in the context of eDNA analysis, as beta diversity often dominates the total diversity in ecological studies at larger scales, and the ability to separate the sources of beta diversity helps to highlight the patterns of interest.

The general approach to eDNA analysis typically involves extracting, preparing, filtering, and clustering the reads, performing sequence identity analysis against the sequence databases (such as NCBI or BOLD), and performing statistical analysis of the results. For example, the microDecon package can be used to filter out contaminating sequences from the readouts, allowing to remove tags jumped by the sequencer or from external sources using

negative control samples for subtraction [38]. In particular, removing contamination is of utmost importance when analyzing eDNA, as small amounts of contaminating DNA can greatly affect the diversity estimates and downstream analysis. The readouts that have undergone the filtering process can then be subjected to identity analysis using the BLAST algorithm, either through the web portal or, for larger datasets, using the command-line BLAST+ tools [39]. While not directly an R package, the BLAST+ algorithms can be used in conjunction with R, as the BLAST query results can be parsed directly in R code using the appropriate command-line utility. As such, combining taxonomy analysis using BLAST with statistical analysis in R provides the researcher with increased flexibility and control over the data while reducing the complexity of the data analysis pipeline.

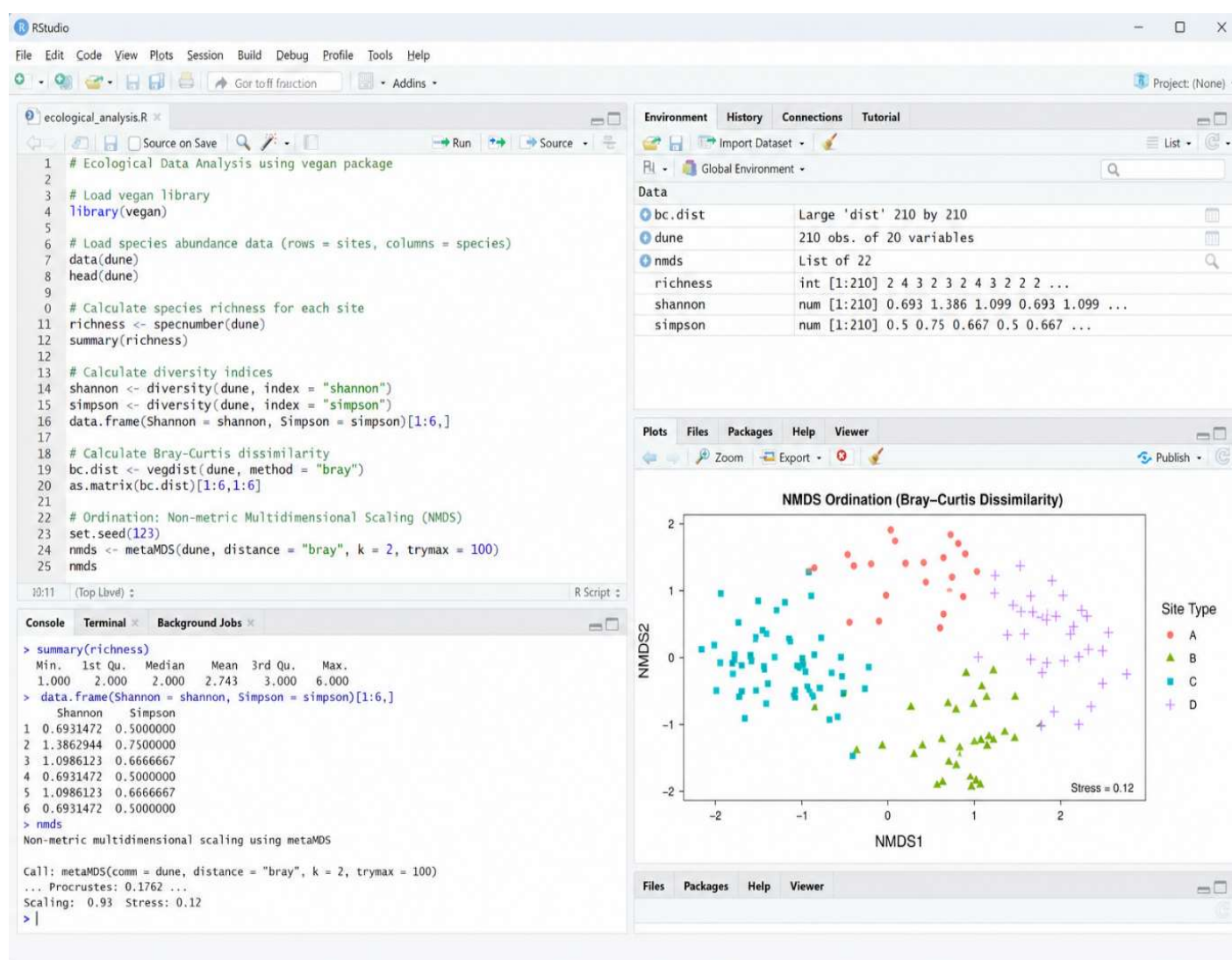


Fig. 2. Schematic illustration of an RStudio workflow. RStudio graphical user interface (GUI) illustrating ecological community analysis using the *vegan* package in R. The Source pane (top left) contains an R script for ecological data analysis, the Console pane (bottom left) displays the corresponding R commands and outputs. The Environment pane (top right) lists the imported datasets and computed objects generated during the analysis. The Plots pane (bottom right) presents the graphical ordination output.

Similarly, data from other platforms can be analyzed in R,

depending on the particular needs and available resources.

For example, SILVAngs is an online platform that allows to analyze the NGS results of ribosomal DNA [40]. The results of such an analysis can then be imported into R for further analysis, such as exploring the community richness, regularity, and divergence. As such, the flexibility of R allows the researcher to use the existing platforms for some aspects of the analysis while utilizing the customizable statistical tools available in R for other aspects.

The statistical analysis is then performed using R programming language and the RStudio platform. The analysis using the vegan package expands upon the basic tidyverse and ggplot2 package, providing the functions necessary to estimate the Shannon diversity index for alpha diversity estimation and calculate the Jaccard index for pairwise beta diversity comparison. Similar procedures can be used to estimate the Bray-Curtis dissimilarity, if the read counts are used for analysis rather than the presence/absence of species. In particular, the choice between presence-absence data and the abundance data depends on how much trust can be established in the read counts, as using abundance data provides more information but requires the assumptions about how much do these abundances reflect the true abundance in the environment. The analysis will be performed either on OTUs or ASVs depending on the marker gene used for analysis and the level of clustering chosen. In particular, the mitochondrial 12S rDNA is used for vertebrate species [41, 42], while the mtDNA COI gene is used for invertebrates [43, 44]. As such, the choice of the marker gene defines the taxonomic resolution of the analysis, as well as the downstream analysis, as different R packages provide different levels of resolution for different classification ranks. Furthermore, the choice of marker gene and the amplicon size affects the bias introduced to the analysis due to primer specificity, as well as the ability to assign taxonomy using the reference databases. As such, it is paramount to document the packages, databases, and parameters used to ensure consistency and repeatability of results across different eDNA analyses.

3. VISUALIZATION, STATISTICAL ANALYSIS AND SIGNIFICANCE USING R

Once taxonomic assignment and diversity estimation is complete, ordination techniques are the principal means of reducing dimensionality and visually examining complex sequence data. Ordination transforms a huge matrix of samples and taxa to a low-dimensional space in which samples with similar community composition appear close together and those with dissimilar composition appear far apart, facilitating the detection of clustering by site, season or habitat type, outlier detection, and relating community structure to measured environmental gradients. Three ordination methods common in eDNA-based ecological analysis through the vegan R package [45] are: Principal Component Analysis (PCA) (Figure 3), Principal Coordinates Analysis (PCoA) (Figure 4) and Non-metric Multidimensional Scaling (NMDS). Redundancy Analysis

(RDA) (Figure 5) extends these unconstrained methods by constraining community variation in measured environmental variables, allowing formal testing of community-environment relationships [46]. Together these methods make the vegan package a versatile toolkit for ecological ordination and for visualizing relationships among samples based on community composition across different ecosystems.

Principal Component Analysis is a linear ordination technique that reduces the dimensionality of abundance data by identifying orthogonal axes, termed the principal components, which successively maximize explained variance. In eDNA metabarcoding, PCA is typically applied to Hellinger-transformed OTU or ASV abundance data because Hellinger transformation reduces the influence of dominant taxa and mitigates the compositional nature of sequencing counts in which the apparent abundance of one taxon depends on the abundance of all others. Figure 3 presents a representative PCA biplot in which each point corresponds to an individual environmental sample collected from four hypothetical sampling sites, labelled Sites A to D, while colored ellipses indicate the clustering of samples with similar community composition. The first two components explain 32.4% and 18.6% of variance respectively, with 51.0% explained cumulatively, and the separation of Site A, Site B, Site C, and Site D samples along these axes indicates that community composition varies systematically among locations. Arrows represent selected environmental variables such as temperature, dissolved oxygen, conductivity, nitrate, pH, and turbidity, with their direction indicating the gradient of increasing values and their length reflecting the strength of correlation with the principal components; samples located in the direction of a vector are positively associated with that environmental variable, so that the positioning of Site A samples toward temperature and Site B samples toward dissolved oxygen suggests distinct environmental drivers of assemblage structure at those sites. The biplot therefore serves both as an exploratory visualization and as a preliminary indication of which environmental variables merit formal constrained analysis.

Principal Coordinates Analysis is a distance-based ordination technique that projects samples to a two- or three-dimensional space by using a distance or dissimilarity matrix, rather than the raw abundance matrix used in PCA. This is important because PCoA can accommodate any ecologically meaningful dissimilarity measure such as Bray-Curtis, Jaccard and UniFrac, allowing analysts to choose a metric appropriate to the data type and research question. Figure 4 presents a representative PCoA plot based on Bray-Curtis dissimilarity in which each point represents an individual environmental sample collected from four hypothetical sampling sites, Sites A to D, and the distance between points reflects differences in community composition inferred from eDNA sequence data. Samples clustering closely indicate similar species assemblages, whereas greater separation denotes increased ecological dissimilarity, and the clear grouping of samples by site in the plot indicates that spatial location is a major determinant of assemblage composition.

PCoA is particularly useful for distinguishing patterns and relationships within data that are otherwise difficult to interpret from the original distance matrix, and it works on the principle that samples closer in ordination space are genetically or ecologically similar while those further apart are more distinct [47]. As PCoA does not assume linear relationships between variables, it is often preferred when community gradients are expected to be non-linear or when dissimilarity measures other than Euclidean distance are required.

Non-metric Multidimensional Scaling is another distance-based ordination approach, which preserves only the rank order of dissimilarities rather than their absolute values, making NMDS particularly suitable for ecological data in which the magnitude of dissimilarity is less reliable than the ordering of samples, and is often used alongside PCoA to confirm that observed clustering patterns are not

artifacts of a particular ordination algorithm. The adequacy of an NMDS solution is assessed by the stress value which indicates how well the reduced-dimensional configuration reproduces the original rank-order dissimilarities, and analysts typically compare solutions across several random starts to avoid local minima. In practice, PCA, PCoA and NMDS are used together rather than in isolation, as agreement among methods strengthens confidence in the ecological interpretation while disagreement signals that the data structure may be more complex than a single ordination captures.

Redundancy Analysis links biodiversity patterns directly to environmental variables by combining ordination with constrained regression, such that the resulting axes are linear combinations of measured environmental predictors rather than unconstrained gradients.

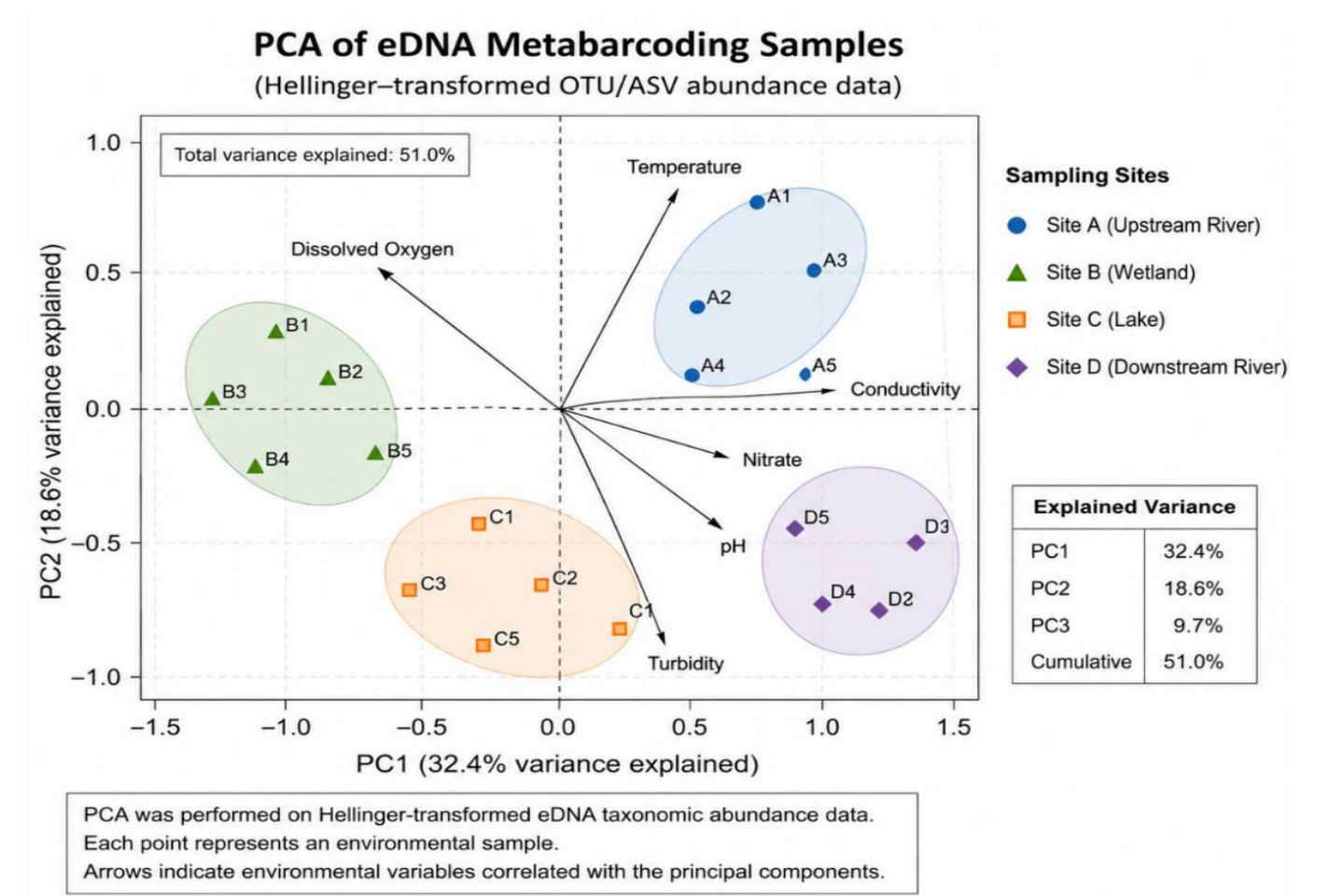


Fig. 3. Representative Principal Component Analysis (PCA) biplot of environmental DNA (eDNA) metabarcoding data based on Hellinger-transformed Operational Taxonomic Unit (OTU) or Amplicon Sequence Variant (ASV) abundance data. Each point represents an individual environmental sample collected from four hypothetical sampling sites (Sites A–D), while coloured ellipses indicate clustering of samples with similar community composition. Arrows represent selected environmental variables (temperature, dissolved oxygen, conductivity, nitrate, pH, and turbidity), with their direction indicating the gradient of increasing values and their length reflecting the strength of correlation with the principal components. Samples located in the direction of a vector are positively associated with that environmental variable.

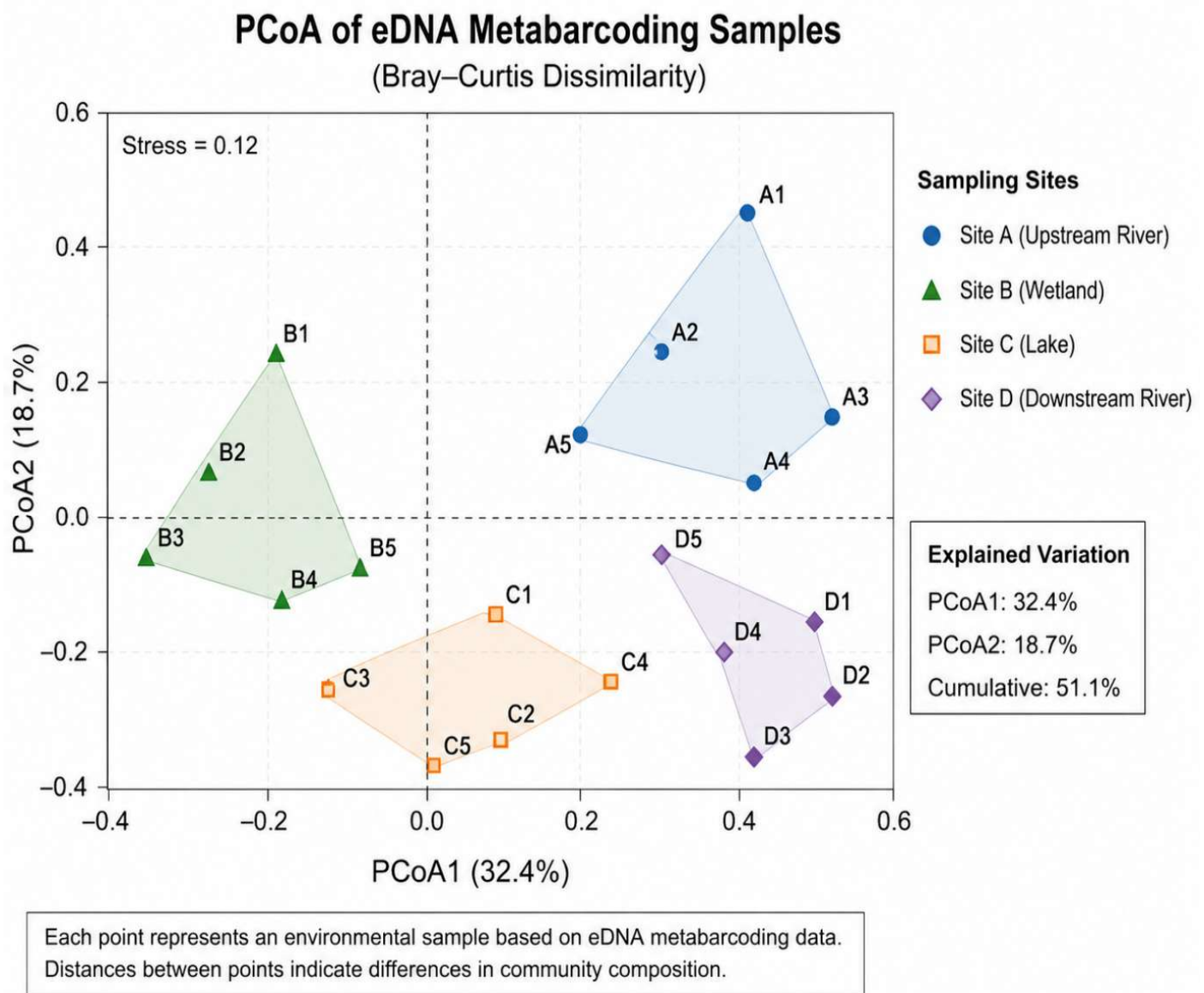


Fig. 4. Representative Principal Coordinates Analysis (PCoA) plot illustrating the ordination of environmental DNA (eDNA) metabarcoding samples based on Bray–Curtis dissimilarity. Each point represents an individual environmental sample collected from four hypothetical sampling sites (Sites A–D), while the distance between points reflects differences in community composition inferred from eDNA sequence data. Samples clustering closely indicate similar species assemblages, whereas greater separation denotes increased ecological dissimilarity.

Figure 5 presents a representative RDA biplot in which the first two constrained axes explain 38.4% and 22.7% of constrained variation, a permutation test yields $F = 4.27$ with $P = 0.002$, and constrained variation totals 61.1%. Each point represents an individual environmental sample collected from different sampling sites, while species or OTUs and ASVs are positioned according to their association with the environmental gradients. Environmental variables are represented by arrows, where the direction indicates the gradient of increasing values and the arrow length reflects the relative strength of their influence on community composition. Samples or taxa located in the direction of an environmental vector are positively associated with that

variable, whereas those positioned in the opposite direction show negative association, such that the separation of Site A samples toward temperature, Site B samples toward dissolved oxygen, Site C samples toward pH and turbidity, and Site D samples toward nitrate and conductivity translates directly into statements about which environmental factors structure each assemblage. The permutation test provides a statistical assessment of whether the observed community–environment relationship exceeds what would be expected by chance, distinguishing RDA from purely descriptive ordination and making it suitable for hypothesis-driven eDNA research.

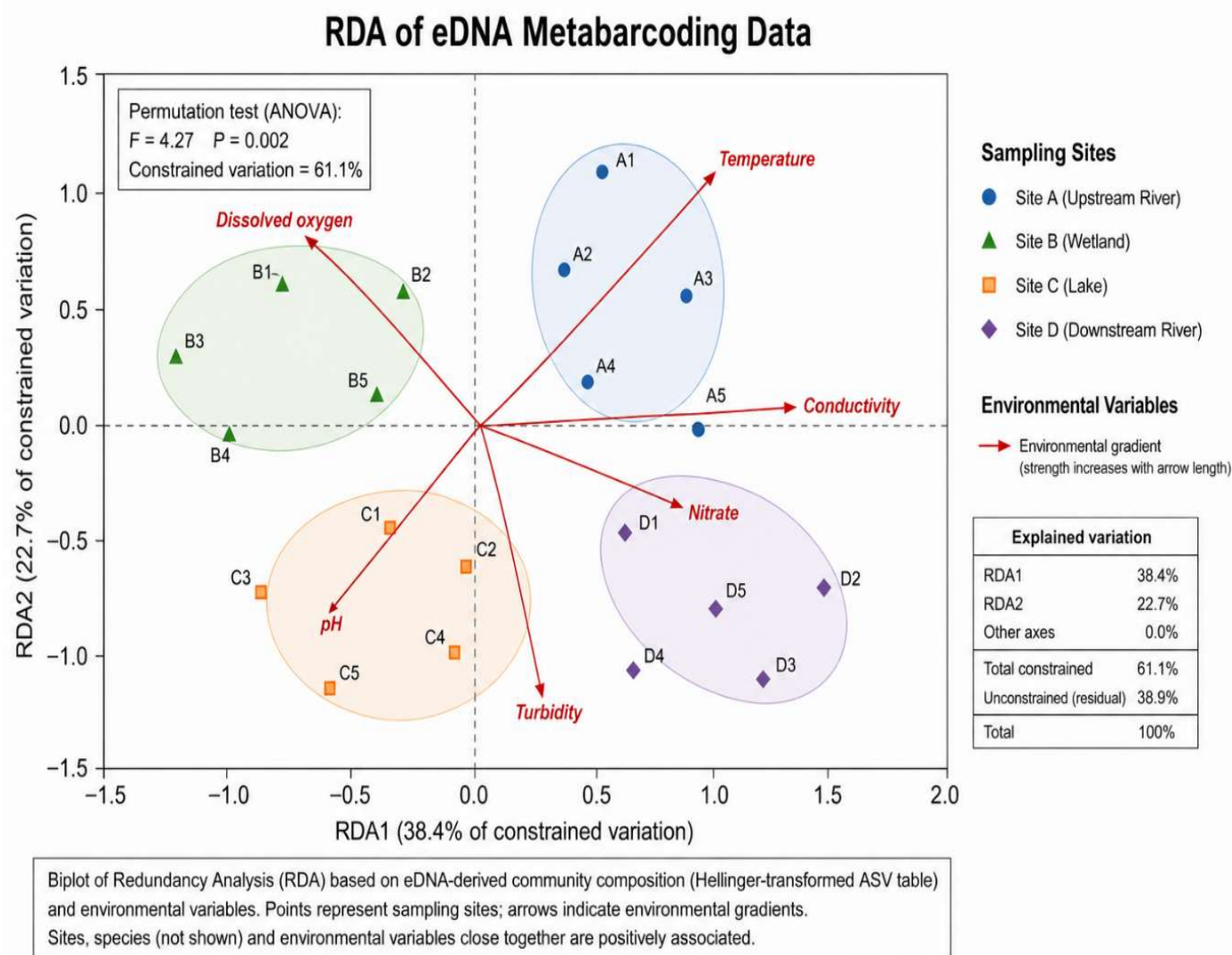


Fig. 5. Representative Redundancy Analysis (RDA) biplot illustrating the relationship between environmental DNA (eDNA)-derived community composition and measured environmental variables such as water temperature, pH, dissolved oxygen (DO), conductivity, nitrate concentration, and turbidity). Each point represents an individual environmental sample collected from different sampling sites, while species (or OTUs/ASVs) are positioned according to their association with the environmental gradients. Environmental variables are represented by arrows, where the direction indicates the gradient of increasing values and the arrow length reflects the relative strength of their influence on community composition. Samples or taxa located in the direction of an environmental vector are positively associated with that variable, whereas those positioned in the opposite direction show a negative association.

All of these analyses can be performed using the R programming language through the *vegan* package in RStudio, a graphical user interface that simplifies working with R while preserving the reproducibility of scripted analysis. The R-based statistical packages *vegan* and *phyloseq* are central to the analysis of metabarcoding data as they enable the execution of complex computational tasks necessary to organize, analyze and interpret the vast datasets characteristic of metabarcoding studies [48, 49]. PCA and PCoA are the two statistical techniques most commonly used for visualizing complex data in eDNA-based biodiversity analysis [50–52], and serve to reduce the dimensionality of large datasets, simplify complex data, and identify patterns

of genetic variation across different environmental samples that suggest species assemblage differences among areas or ecological conditions. Such analyses help generate information about species richness, how community composition changes across space and time in response to environmental gradients, and possible ecological interactions inferred from spatial clustering of samples. The usage of the R programming language has therefore made it considerably easier to interpret complex and large NGS datasets from eDNA by compiling ecosystem structure and functioning into pictorial representations that can be examined, compared and reported transparently.

4. CONCLUSION

Environmental DNA (eDNA) metabarcoding and next-generation sequencing enabled the unprecedented description of biodiversity, facilitating the comprehensive assessment of species assemblages present in terrestrial, freshwater, and marine ecosystems. However, the ability to analyze and interpret such datasets relies on the capacity of the methods used to extract valuable information from large and complicated sequences obtained from the samples. R programming language, with such packages as *vegan*, *phyloseq*, *microDecon*, *betapart*, *stats*, *mgcv* and *Rstudio* as an integrated development environment, offers a wide range of tools necessary to address the task. First of all, the language can be used for read filtration, taxonomic assignment, and diversity estimation. Furthermore, *vegan* and *phyloseq* packages, along with other tools mentioned above, provide the means for calculating beta diversity partitioning and various approaches to modeling ecological data. The methods of multivariate ordination, such as PCA, PCoA, NMDS and RDA, allowed researchers to explore community ecology by investigating the differences in species composition and the possible environmental variables influencing these differences. They can also be used to identify and visually observe the patterns in the data and display them on ordination diagrams. The combination of R programming language, with its wide range of tools for statistical analysis and graphics, enables the processing of complex datasets, such as environmental DNA, into valuable information. The language provides the opportunities for creating clear and informative visual displays of statistics and patterns in the data to explore and understand the characteristics of communities' ecology. Moreover, the ability of R to process data in the form of scripts allows for easy repetition of the calculations, ensuring the accuracy of conclusions. Overall, the analysis using R programming language is critical for biodiversity assessment and management, as well as for conservation genetics and evolutionary ecology, providing the basis for reproducible bioinformatics.

5. FUTURE DIRECTIONS

In spite of the progress, the framework described in this review still contains a number of unresolved challenges that could impede the adoption of eDNA metabarcoding for routine, large-scale biodiversity monitoring. Resolving these issues will dictate whether R-based analysis remains a tool restricted to niche players or becomes embedded in the infrastructure of the broader ecological research community.

Probably the most pressing issue is reference database incompleteness, which limits the accuracy of taxonomic assignments. Even the best ordination and diversity analysis techniques cannot compensate for the absence of reference sequences or incorrect geographical distribution of a given taxon. Especially underrepresented are invertebrates, fungi, protists, and tropical species, with most sequences coming

from temperate regions. Future work should focus on building country- or region-specific barcode libraries based on voucher specimens, implementing stricter metadata curation guidelines, and adhering to the FAIR data principles (Global eDNA Reference Library initiative, among others). It will be critical to achieve broad participation at the national level before assignment confidence becomes routine practice rather than an exception.

At the level of methodologies, consistent findings across studies are hampered by the use of different volumes of water, types and pore sizes of filters, preservation techniques, DNA extraction kits and methods, primer sets and PCR conditions, sequencing platforms, and downstream bioinformatic processing. Comparison of studies using these variables is complicated by the fact that, for the same stream, detection of up to 48 species differed by more than an order of magnitude between commercial eDNA providers. To promote comparability and facilitate large-scale analyses, standard operating procedures should be created for all aspects of a study, from field collection to bioinformatic reporting, while the development of such protocols is currently presented as an optional add-on. For R, this will mean the creation of analysis templates as packages with specific versions of functions that would minimize differences between studies and reduce the space for ad hoc decisions affecting downstream analysis. Another possibility is for R-based workflows to be used for post-publication tweaking of analyses in publications that utilized other pipelines.

It will be critical to capitalize on the opportunities presented by long-read sequencing and field assays. The Oxford Nanopore and PacBio technologies are already able to produce reads that span entire barcode regions or even mitochondrial genomes for better species resolution, especially at the level of closely related complexes. Meanwhile, the handheld MinION device enables biodiversity monitoring in remote field locations where infrastructure is unavailable. Such an approach is particularly valuable for conservation genomics, where rapid responses are often needed. To fully exploit the potential of these technologies, R will need additional packages and novel data structures that go beyond the established short-read processing tools focused on Illumina data.

A separate avenue of development is connected to interpreting eDNA findings in a broader ecological context. While detecting target species is informative, eDNA reads cannot provide information on their developmental stage, sex ratio, health, or abundance. Moreover, there is currently no consistent relationship between the number of reads and the biomass of the organisms present in a water sample that would enable quantitative modeling. An alternative to using eDNA as a substitute for traditional surveys is to utilize the shorter-lived RNA molecules, which appear to degrade faster and hence reflect a closer approximation of current biogeographical reality. The ability of R packages to analyze both eDNA and eRNA would benefit greatly from additional methods to evaluate the consistency of sequence reads with independent population genetic data.

A related priority is to leverage the promise of artificial intelligence and machine learning in accelerating the assignment process. Classifiers based on patterns in k-mer frequencies and deep learning architectures are capable of processing not only short, high-quality sequences but also longer, erroneous reads generated in next-generation sequencing assays. The combined use of database similarity searches, statistical models, and phylogenetic algorithms is likely to become the standard in the next several years. Nevertheless, the practical implementation of such models will require extensive training data and computational resources. Algorithms that successfully incorporate external taxonomic information are likely to gain wider acceptance, especially in the realm of regulatory science, where R packages would be needed to implement such algorithms or provide wrappers around existing ones.

Finally, the drive towards integrating eDNA methods into existing monitoring frameworks, such as the EU Water Framework Directive, will require recalibration of many widely used indices and a return to the fundamentals of community ecology. Occupancy modeling and other probabilistic approaches to accounting for detection bias are likely to play an important role in such analyses, as they explicitly consider the possibility of species being present but undetected. Equally critical is the need to refine reference databases and adhere to standardized protocols before eDNA metabarcoding results can be used as a complement to traditional biodiversity assessments in conservation genomics and evidence-based policymaking.

DECLARATIONS

Ethical Approval

We affirm that this manuscript is an original work, has not been previously published, and is not currently under consideration for publication in any other journal or conference proceedings. All authors have reviewed and approved the manuscript, and the order of authorship has been mutually agreed upon.

Availability of data and material

All of the data obtained or analyzed during this study is included in the report that was submitted.

Conflicts of Interest

The authors declare that there is no conflict of interests.

Authors' contributions

All authors contributed equally to this work.

REFERENCES

- [1] Deiner, K., Yamanaka, H. and Bernatchez, L., **2021**. The future of biodiversity monitoring and conservation utilizing environmental DNA. *Environmental DNA*, 3(1), pp.3-7. <https://doi.org/https://doi.org/10.3929/ethz-b-000488115>
- [2] Berran, M.S., Weeks, A.R., Coleman, R.A., Connelly, F., Song, S.V. and Tingley, R., **2025**. A comparison of environmental DNA sampling strategies for detecting anuran populations in wetlands. *Hydrobiologia*, 852(13), pp.3485-3501. <https://doi.org/10.1007/s10750-025-05826-0>
- [3] Taberlet, P., Coissac, E., Hajibabaei, M. and Rieseberg, L.H., **2012**. Environmental DNA. *Molecular ecology*, 21(8), p.1789. <https://doi.org/10.1111/j.1365-294X.2012.05542.x>
- [4] Perry, W.B., Seymour, M., Orsini, L., Jâms, I.B., Milner, N., Edwards, F., Harvey, R., de Bruyn, M., Bista, I., Walsh, K. and Emmett, B., **2024**. An integrated spatio-temporal view of riverine biodiversity using environmental DNA metabarcoding. *Nature Communications*, 15(1), p.4372. <https://doi.org/10.1038/s41467-024-48640-3>
- [5] Vamos, E.E., Elbrecht, V. and Leese, F., **2017**. Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, p.e14625. <https://doi.org/10.3897/mbmg.1.14625>
- [6] Ratnasingham, S. and Hebert, P.D., **2007**. BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular ecology notes*, 7(3), pp.355-364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>
- [7] Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D. and Karsch-Mizrachi, I., **2020**. GenBank. *Nucleic acids research*, 48(D1), pp.D84-D86. <https://doi.org/10.1093/nar/gkz956>
- [8] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.A., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J. and Kuhn, M., **2019**. Welcome to the Tidyverse. *Journal of open source software*, 4(43), p.1686. <https://doi.org/10.21105/joss.01686>
- [9] Wickham, H., **2016**. Getting Started with ggplot2. In *ggplot2: Elegant graphics for data analysis* (pp. 11-31). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_2

- [10] Morton, T., **2010**. Guest column: Queer ecology. *pmla*, 125(2), pp.273-282. <https://doi.org/10.1632/pmla.2010.125.2.273>
- [11] McMurdie, P.J. and Holmes, S., **2013**. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4), p.e61217. <https://doi.org/10.1371/journal.pone.0061217>
- [12] McKnight, D.T., Huerlimann, R., Bower, D.S., Schwarzkopf, L., Alford, R.A. and Zenger, K.R., **2019**. microDecon: a highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environmental DNA*, 1(1), pp.14-25. <https://doi.org/10.1002/edn3.11>
- [13] Baselga, A., Orme, D., Villeger, S., De Bortoli, J., Leprieux, F., Logez, M., Martinez-Santalla, S., Martin-Devasa, R., Gomez-Rodriguez, C. and Crujeiras, R.M., **2012**. betapart: Partitioning beta diversity into turnover and nestedness components. (No Title). <https://doi.org/10.32614/cran.package.betapart>
- [14] Gao, M., Ye, Y., Zheng, Y. and Lai, J., **2025**. A comprehensive analysis of R's application in ecological research from 2008 to 2023. *Journal of Plant Ecology*, 18(1), p.rtaf010. <https://doi.org/10.1093/jpe/rtaf010>
- [15] Wood, S.N., **2004**. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), pp.673-686. <https://doi.org/10.1198/016214504000000980>
- [16] Keck, F. and Mahon, A., **2020**. Handling biological sequences in R with the bioseq package. *Methods in Ecology & Evolution*, 11(12), p.1728. <https://doi.org/10.1111/2041-210X.13490>
- [17] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L., **2009**. BLAST+: architecture and applications. *BMC bioinformatics*, 10(1), p.421. <https://doi.org/10.1186/1471-2105-10-421>
- [18] Rozanski, R., Velez, L., Hocdé, R., Duhamet, A., Waldoock, C., Mouillot, D., Pellissier, L. and Albouy, C., **2024**. Seasonal dynamics of Mediterranean fish communities revealed by eDNA: Contrasting compositions across depths and Marine Fully Protected Area boundaries. *Ecological Indicators*, 166, p.112290. <https://doi.org/10.1016/j.ecolind.2024.112290>
- [19] Saenz-Agudelo, P., Delrieu-Trottin, E., DiBattista, J.D., Martínez-Rincon, D., Morales-González, S., Pontigo, F., Ramírez, P., Silva, A., Soto, M. and Correa, C., **2022**. Monitoring vertebrate biodiversity of a protected coastal wetland using eDNA metabarcoding. *Environmental DNA*, 4(1), pp.77-92. <https://doi.org/10.1002/edn3.200>
- [20] Wolany, L., Enss, J., Leese, F. and Buchner, D., **2024**. Homogenization of insect bulk samples yields more comprehensive yet comparable biodiversity data than non-destructive lysis. *Metabarcoding and Metagenomics*, 8, p.e129814. <https://doi.org/10.3897/mbmg.8.129814>
- [21] Keck, F., Hürlemann, S., Locher, N., Stamm, C., Deiner, K. and Altermatt, F., **2022**. A triad of kicknet sampling, eDNA metabarcoding, and predictive modeling to assess richness of mayflies, stoneflies and caddisflies in rivers. *Metabarcoding and Metagenomics*, 6, pp.117-131. <https://doi.org/10.3929/ethz-b-000578113>
- [22] Legendre, P., **1983**. Numerical ecology: developments and recent trends. In *Numerical taxonomy* (pp. 505-523). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-69024-2_56
- [23] Taberlet, P., Bonin, A., Zinger, L. and Coissac, E., **2018**. *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press. <https://doi.org/10.1093/oso/9780198767220.001.0001>
- [24] Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H.L., Buckley, T.R., Cruickshank, R., Dopheide, A., Handley, K.M., Hermans, S. and Kamke, J., 10Lear et al.: Environmental DNA methods REVIEW Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. <https://doi.org/10.20417/nzjecol.42.9>
- [25] Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., De Vere, N. and Pfrender, M.E., **2017**. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular ecology*, 26(21), pp.5872-5895. <https://doi.org/10.1111/mec.14350>
- [26] Fernández, S., Rodríguez, S., Martínez, J.L., Borrell, Y.J., Ardura, A. and García-Vázquez, E., **2018**. Evaluating freshwater macroinvertebrates from eDNA metabarcoding: A river Nalón case study. *PLoS One*, 13(8), p.e0201741. <https://doi.org/10.1371/journal.pone.0201741>
- [27] Lynggaard, C., Frøslev, T.G., Johnson, M.S., Olsen, M.T. and Bohmann, K., **2024**. Airborne environmental DNA captures terrestrial vertebrate diversity in nature.

- Molecular Ecology Resources*, 24(1), p.e13840. <https://doi.org/10.1111/1755-0998.13840>
- [28] Davis, S., Chariton, A., Brandis, K.J., Hejl, N. and Stow, A.J., **2025**. Tracing Waterbirds in Water: A Pilot Study on the Utility of eDNA Monitoring for Inland Australian Waterbirds. *Austral Ecology*, 50(7), p.e70094. <https://doi.org/10.1111/aec.70094>
- [29] Zavez, K. and Harel, O., **2026**. Teaching statistical concepts using computing tools: A review of the literature. *Journal of Statistics and Data Science Education*, 34(4), pp.426-437. <https://doi.org/10.1080/26939169.2024.2445541>
- [30] Patil, I., Makowski, D., Ben-Shachar, M.S., Wiernik, B.M., Bacher, E. and Lüdecke, D., **2022**. datawizard: An R package for easy data preparation and statistical transformations. *Journal of Open Source Software*, 7(78), p.4684. <https://doi.org/10.21105/joss.04684>
- [31] Wickham, H., **2016**. Programming with ggplot2. In *Ggplot2: elegant graphics for data analysis* (pp. 241-253). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_12
- [32] Philipp, E.E., Kraemer, L., Mountfort, D., Schilhabel, M., Schreiber, S. and Rosenstiel, P., **2012**. The Transcriptome Analysis and Comparison Explorer—T-ACE: a platform-independent, graphical tool to process large RNAseq datasets of non-model organisms. *Bioinformatics*, 28(6), pp.777-783. <https://doi.org/10.1093/bioinformatics/bts056>
- [33] de Ruiter, P.C., Wolters, V., Moore, J.C. and Winemiller, K.O., **2005**. Food web ecology: playing Jenga and beyond. *Science*, 309(5731), pp.68-71. <https://doi.org/10.1126/science.1096112>
- [34] Huse, S.M., Mark Welch, D.B., Voorhis, A., Shipunova, A., Morrison, H.G., Eren, A.M. and Sogin, M.L., **2014**. VAMPS: a website for visualization and analysis of microbial population structures. *BMC bioinformatics*, 15(1), p.41. <https://doi.org/10.1186/1471-2105-15-41>
- [35] Wickham, H., Bryan, J., Chang, W., McGowan, L., François, R., Golemund, G., Lemon, J., Kassambara, A., Mundt, F. and Revelle, W.R., **2020**. R: A language and environment for statistical computing. *J. Open Source Softw*, 8, pp.1-20. <https://doi.org/10.32614/CRAN.package.factoextra>
- [36] Marx, B.D. and Eilers, P.H., **1998**. Direct generalized additive modeling with penalized likelihood. *Computational statistics & data analysis*, 28(2), pp.193-209. [https://doi.org/10.1016/S0167-9473\(98\)00033-4](https://doi.org/10.1016/S0167-9473(98)00033-4)
- [37] Tuomisto, H., **2010**. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1), pp.2-22. <https://doi.org/10.1111/j.1600-0587.2009.05880.x>
- [38] Schmieder, R. and Edwards, R., **2011**. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6(3), p.e17288. <https://doi.org/10.1371/journal.pone.0017288>
- [39] Ye, J., McGinnis, S. and Madden, T.L., **2006**. BLAST: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2), pp.W6-W9. <https://doi.org/10.1093/nar/gkl164>
- [40] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O., **2012**. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1), pp.D590-D596. <https://doi.org/10.1093/nar/gks1219>
- [41] Madon, B., Haderlé, R., Arotcharen, E., David, R., Fontaine, Q., Marengo, M., Thomas, H., Torralba, A., Valentini, A. and Jung, J.L., **2025**. eDNA and Citizen Science Reveal Hidden Fish Biodiversity in Climate-Stressed Urban Ports of the Mediterranean Sea. *Environmental DNA*, 7(4), p.e70142. <https://doi.org/10.1002/edn3.70142>
- [42] Sales, N.G., McKenzie, M.B., Drake, J., Harper, L.R., Browett, S.S., Coscia, I., Wangenstein, O.S., Baillie, C., Bryce, E., Dawson, D.A. and Ochu, E., **2020**. Fishing for mammals: Landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. *Journal of Applied Ecology*, 57(4), pp.707-716. <https://doi.org/10.1111/1365-2664.13592>
- [43] Wolany, L., Enss, J., Leese, F. and Buchner, D., **2024**. Homogenization of insect bulk samples yields more comprehensive yet comparable biodiversity data than non-destructive lysis. *Metabarcoding and Metagenomics*, 8, p.e129814. <https://doi.org/10.3897/mbmg.8.129814>
- [44] Keck, F., Hürlemann, S., Locher, N., Stamm, C., Deiner, K. and Altermatt, F., **2022**. A triad of kicknet sampling, eDNA metabarcoding, and predictive modeling to assess aquatic macroinvertebrate biodiversity. *bioRxiv*, pp.2022-01. <https://doi.org/10.1101/2022.01.03.474789>

- [45] Braga, P.H.P., Hébert, K., Hudgins, E.J., Scott, E.R., Edwards, B.P., Sánchez Reyes, L.L., Grainger, M.J., Foroughirad, V., Hillemann, F., Binley, A.D. and Brookson, C.B., **2023**. Not just for programmers: How GitHub can accelerate collaborative and reproducible research in ecology and evolution. *Methods in Ecology and Evolution*, 14(6), pp.1364-1380. <https://doi.org/10.1111/2041-210X.14108>
- [46] Simberloff, D., Berthet, P., Boy, V., Cousins, S.H., Fortin, M.J., Goldberg, R., Lefkovitch, L.P., Ripley, B., Scherrer, B. and Tonkyn, D., **1987**. Novel statistical analyses in terrestrial animal ecology; dirty data and clean questions. In *Developments in Numerical Ecology* (pp. 559-572). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-70880-0_18
- [47] Anderson, M.J., **2001**. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1), pp.32-46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
- [48] Barnett, D.J., Arts, I.C. and Penders, J., **2021**. microViz: an R package for microbiome data visualization and statistics. *Journal of Open Source Software*, 6(63), p.3201. <https://doi.org/10.21105/joss.03201>
- [49] Buttigieg, P.L. and Ramette, A., **2014**. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS microbiology ecology*, 90(3), pp.543-550. <https://doi.org/10.1111/1574-6941.12437>
- [50] Ramette, A., **2007**. Multivariate analyses in microbial ecology. *FEMS microbiology ecology*, 62(2), pp.142-160. <https://doi.org/10.1111/j.1574-6941.2007.00375.x>
- [51] Bolker, B.M., **2008**. Ecological models and data in R. In *Ecological models and data in R*. Princeton University Press. <https://doi.org/10.1515/9781400840908>
- [52] Paliy, O. and Shankar, V., **2016**. Application of multivariate statistical techniques in microbial ecology. *Molecular ecology*, 25(5), pp.1032-1057. <https://doi.org/10.1111/mec.13536>