

RESOURCE ARTICLE OPEN ACCESS

Improving Whole Biodiversity Monitoring and Discovery With Environmental DNA Metagenomics

Manuel Curto^{1,2,3}  | Ana Veríssimo^{2,3} | Giulia Riccioni^{1,4}  | Carlos D. Santos^{5,6} | Filipe Ribeiro¹ | Sissel Jentoft⁷  | Maria Judite Alves^{8,9} | Hugo F. Gante^{8,10,11} 

¹MARE – Marine and Environmental Sciences Center/ARNET – Aquatic Research Network, Faculty of Sciences, University of Lisbon, Lisbon, Portugal | ²CIBIO – Research Center in Biodiversity and Genetic Resources, Vairão, Portugal | ³BIOPOLIS Program in Genomics, Biodiversity and Land Planning; CIBIO, Vairão, Portugal | ⁴Università Cattolica del Sacro Cuore, Facoltà di Scienze Agrarie, Alimentari e Ambientali; DIANA Dipartimento di Scienze Animali, Della Nutrizione e Degli Alimenti, BioDNA Centro di Ricerca Sulla Biodiversità e Sul DNA Antico, Piacenza, Italy | ⁵MARE - Marine and Environmental Science Centre and ARNET - Aquatic Research Network Associate Laboratory, Department of Environmental Sciences and Engineering, NOVA School of Science and Technology, NOVA University Lisbon, NOVA University of Lisbon, Caparica, Portugal | ⁶Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell, Germany | ⁷Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway | ⁸cE3c – Centre for Ecology, Evolution and Environmental Changes/CHANGE – Global Change and Sustainability Institute, Faculty of Sciences, University of Lisbon, Lisbon, Portugal | ⁹Museu Nacional de História Natural e da Ciência, University of Lisbon, Lisbon, Portugal | ¹⁰Division Ecology, Evolution and Biodiversity Conservation, Department of Biology, KU Leuven, Leuven, Belgium | ¹¹Department of Biology, Royal Museum for Central Africa, Tervuren, Belgium

Correspondence: Manuel Curto (mcurto@cibio.up.pt) | Hugo F. Gante (hugo.gante@kuleuven.be)

Received: 9 December 2024 | **Revised:** 6 March 2025 | **Accepted:** 11 March 2025

Handling Editor: Carla Martins Lopes

Funding: This work was supported by Fundação para a Ciência e a Tecnologia, CEEC/0482/2020, DL 57/2016/CP1440/CP1646/CT0001, LA/P/0069/2020, PTDC/BIA-CBI/31644/2017, UID/04292/2020, UID/BIA/00329/2020, UIDP/50027/2020; Horizon 2020 Framework Programme, 857251; and KU Leuven, STG/21/044.

Keywords: dark diversity | metabarcoding | non-microbial eDNA | river ecosystems | shotgun sequencing | transported eDNA

ABSTRACT

Environmental DNA (eDNA) metagenomics sequences all DNA molecules present in environmental samples and has the potential of identifying virtually any organism from which they are derived. However, due to unacceptable levels of false positives and negatives, this approach is underexplored as a tool for biodiversity monitoring across the tree of life, particularly for non-microscopic eukaryotes. We present SEQIDIST, a framework that combines multilocus BLAST matches against several reference databases followed by an analysis of sequence identity distribution patterns to disentangle false positives while revealing new biodiversity and increasing the accuracy of metagenomic approaches. We tested SEQIDIST on an eDNA metagenomic dataset from a riverine site and compared the results to those obtained with an eDNA metabarcoding approach for benchmarking purposes. We start by characterising the biological community (~2000 taxa) across the tree of life at low taxonomic levels and show that eDNA metagenomics has a higher sensitivity than eDNA metabarcoding in discovering new diversity. We show that limited representation of whole genome sequences in reference databases can lead to false positives. For non-microscopic eukaryotes, eDNA metagenomic data often consist of a few sparse, anonymous sequences scattered across the genome, making metagenome assembly methods unfeasible. Finally, we infer eDNA source and residency time using read length distributions as a measure of decay status. The higher accuracy of SEQIDIST opens the discussion of the potential of eDNA metagenomics for archived samples and its implementation in long-term biodiversity monitoring at a planetary scale.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

1 | Introduction

Global declines and changes in biodiversity distribution are hallmarks of the Anthropocene (Butchart et al. 2010). Moreover, lack of taxonomic expertise and financial and technical challenges hinder accurate large-scale biodiversity assessment (Pereira et al. 2010). Importantly, much biodiversity is still undocumented, and knowledge gaps are more pronounced in underexplored regions, especially in diverse freshwater and marine environments, where biodiversity loss is likely underestimated (Schmeller et al. 2017).

The DNA present in the environment (eDNA), derived from both living microorganisms (intracellular eDNA) and remains of macroscopic eukaryotes (extracellular eDNA), is a powerful source of biodiversity information (Taberlet et al. 2012). In the context of biodiversity monitoring, eDNA appears as an emergent alternative to capture-based methods that are often invasive, require taxonomic expertise, and are inefficient in detecting rare or elusive species (Goldberg et al. 2016). Moreover, eDNA samples harbour information across the tree of life, widening their applicability to multispecies monitoring projects (Stat et al. 2017). In this context, rivers are a particularly interesting source of biodiversity information since they accumulate eDNA molecules originating not only from local aquatic organisms but also from upstream and nearby terrestrial biota (Deiner et al. 2016) through long-distance transport, runoff, and possibly also deposition of airborne eDNA (Clare et al. 2022; Littlefair et al. 2023; Lynggaard et al. 2022). Therefore, riverine eDNA samples have the potential to provide biodiversity information at the landscape level.

The analysis of eDNA samples by amplifying and sequencing specific barcodes, known as eDNA metabarcoding, has revolutionised biodiversity monitoring due to its relatively simple and cost-effective implementation (Deiner et al. 2017). Nevertheless, this method has known limitations that need to be circumvented, such as biases in taxon amplification and decay rate (Jo et al. 2017; Kelly et al. 2019). In turn, environmental metagenomics (or shotgun metagenomics (Quince et al. 2017)) sequences DNA molecules from an environmental sample (e.g., water, soil, air) with limited taxonomic biases or prior knowledge of species composition. By doing so, eDNA metagenomics is the methodology that can unlock the largest amount of information from eDNA samples.

Despite its proven success in characterising microbial communities for two decades (Venter et al. 2004), application of eDNA metagenomics to non-microscopic eukaryotes has been surprisingly rare and severely constrained by analytical limitations leading to high prevalence of false positives and negatives (Simon et al. 2019; Stat et al. 2017). eDNA from non-microscopic eukaryotes is usually extracellular, fragmented and present in the environment in low abundance (Nagler et al. 2022), making the assembly-based methodologies, already established in the field of microbiology, ineffective in reconstructing larger contigs (Quince et al. 2017). Thus, the detection of non-microscopic eukaryotes has to rely on directly comparing anonymous short reads to reference databases (Manu and Umapathy 2023; Singer et al. 2020). Because reference sequence databases are incomplete in

terms of taxonomic representation, most of the data remains unmapped, resulting in false negatives. Furthermore, some regions of the genome are conserved across species and attract heterospecific matches that may raise false positive rates (Simon et al. 2019). The rare studies describing multicellular eukaryotic communities using eDNA metagenomics were conducted at high taxonomic levels, and taxon detection was based on individual sequence matches (Coward et al. 2018; Manu and Umapathy 2023; Singer et al. 2020). Here, we address the lack of resolution and detection bias of using single sequence matches by developing a new framework that makes eDNA metagenomics a viable and powerful tool for whole biodiversity monitoring, including for non-microscopic organisms.

In this study, we propose using eDNA metagenomics coupled with a dedicated bioinformatics workflow, named SEQIDIST. Instead of considering individual sequence matches as indicators of positive detection, which can lead to false positives, SEQIDIST combines percent identity information from all BLAST matches to a taxon. This approach assumes that, while some regions of the genome may be conserved across species boundaries, this will not be the case for most of the genome. Therefore, sequence identity distribution patterns should differ between intraspecific and interspecific comparisons. We provide proof-of-principle for SEQIDIST's ability to detect false positives. By applying SEQIDIST to an eDNA metagenomic dataset from the Ave River (NW Iberian Peninsula), we demonstrate that SEQIDIST can characterise freshwater communities across the tree of life with high accuracy and greater sensitivity than traditional eDNA metabarcoding. In this context, we focus on fish taxa as a key case study to evaluate the ability of metagenomics to accurately describe eDNA diversity at the species level. Teleost fishes are the best documented taxon in the Ave River and the group of organisms with the largest amount of genomic resources in public databases. Additionally, we test if read length contains information on eDNA decay, which could improve the characterisation of eDNA origin (e.g., intra- vs. extracellular and habitat), illustrating the potential of metagenomics to extract additional levels of information from eDNA samples.

2 | Material and Methods

2.1 | Multilocus SEQIDIST Detection of True and False Positives

Most genomic reads are expected to match genome assemblies of conspecific taxa at high sequence identity, with no or very few mismatches. Nevertheless, some proportion of reads can also match heterospecific taxa with high sequence identity (e.g., >99%), which would result in false positives if individual matches of each read are considered. This can happen because of missing representative genomes in the reference database, incomplete assemblies or a high degree of conservation of some genomic regions. Nevertheless, only a small part of the genome should be conserved across species, making heterospecific matches with high sequence identity much less frequent than heterospecific matches with lower sequence identity. Therefore, we expect that the multilocus sequence

identity distribution of metagenomic reads to a certain taxon can provide information on whether its detection is a true or false positive. Specifically, false positives should show proportionally more multilocus matches with lower identity values. We tested this expectation by evaluating the sequence identity distributions of BLAST hits from shotgun sequence data against genome assemblies across a range of intraspecific, intrageneric and intrafamilial comparisons. Shotgun Illumina sequencing datasets without any type of enrichment were selected for a few case-study fish species likely to be detected in our study area (described in Section 2.2 below). These included at least one sample of species represented in the short-read archive (SRA) database from GenBank for the following genera (Table S1): *Gobio*, *Gambusia*, *Lepomis*, *Micropterus*, *Luciobarbus*, *Perca*, *Sander*, *Salmo* and *Squalius*.

Short reads were downloaded with the program fastq-dump from the sra toolkit v. 3.0.0 (Sequence Read Archive Toolkit) (<https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft>), and filtered for quality with trimmomatic (Bolger et al. 2014) (details in Section 2.5. Quality control of shotgun reads). Possible PCR duplicates were removed with the script clumpify.sh from BBMap (Bushnell 2014). To mimic the low coverage data obtained with metagenomics data, 1000 random reads were selected from the dataset with seqtk (Li 2012). Paired reads were then merged with PEAR (Zhang et al. 2014) with default parameters and converted to fasta format. Reads were then compared to individual reference genomes (Table S1) using megablast searches, and matches were processed as described in Section 2.7: *Taxonomic assignment of metagenomics reads*. The output is a table with the BLAST match parameters per read and the corresponding taxonomic assignment. These were used in R (R Core Team 2023) to produce histograms based on match sequence identity using the library ggplot2 (Wickham et al. 2016). Several bin sizes were tested (bin-width = 0.1, 0.5, 1.0, 1.25), with the binwidth parameter of 1.25 providing maximum differentiation between intra- and inter-specific comparisons, although the overall pattern did not change for other values. The bin encompassing the highest sequence identity included matches of ID > 99.5%. The remaining bins corresponded to intervals of 1% sequence identity.

2.2 | Study Area and Sampling Site

The Ave River is a small drainage located in the northwestern Iberian Peninsula, flowing for about 91 km from the Cabreira mountain to the Atlantic Ocean at Vila do Conde, Portugal (in a North-East to South-West direction). This drainage (41°15'N–40°45'N; 7°45'W–8°45'W) encompasses an area of 1473 km², ranging up to 1261 m altitude. The climate is mildly Mediterranean (Köppen-Geiger classification, Csb) with cool summers and a strong Atlantic influence. Mean annual rainfall is about 1400 mm (ranging from 900 to 3900 mm) and the annual mean temperature is 12°C–15°C. The basin drains mostly from limestone rocks, with high groundwater input, and land cover is mostly forest (57%, mainly pine and eucalyptus) and farmland (34%). However, the lower and middle sections of the drainage are heavily modified and degraded due to heavy industrialisation and urbanisation. The area around this drainage is inhabited by more than 400 people per km²,

being more densely inhabited in the middle and lower sections. Despite the significant infrastructure efforts in water treatment, the Ave River basin is still considered one of the most polluted rivers in Portugal; yet, the upper Ave watershed and main tributaries are generally pristine areas, populated mostly by brown trout (*Salmo trutta*) and native leuciscids (Collares-Pereira et al. 2021). In contrast, the lower and middle sections of the Ave watershed are generally dominated by non-native fish species, with a small number of migratory species, such as European eel (*Anguilla anguilla*) and thinlip grey mullet (*Chelon ramada*) (Collares-Pereira et al. 2021).

Water samples were collected in the lower section of the Ave River (41°21'6.93" N; 8°40'54.83" W) at ~9.5 km from the river mouth, and just downstream from its fourth barrier (an abandoned watermill). Downstream barriers are semi-permeable and in high winter river flow conditions allow fish to move upstream. Fish communities in this region are composed of the Pyrenean gudgeon (*Gobio lozanoi*), pumpkinseed sunfish (*Lepomis gibbosus*), Iberian barbel (*Luciobarbus bocagei*), European eel, eastern mosquitofish (*Gambusia holbrooki*), northern Iberian chub (*Squalius carolitertii*), northern straight-mouth nase (*Pseudochondrostoma duriense*), ruivaco (*Achondrostoma oligolepis*), thinlip grey mullet, and roach (*Rutilus rutilus*) (Collares-Pereira et al. 2021; Ribeiro and Verissimo 2014). Although migratory species such as *Alosa* spp. and *Salmo trutta* may occur, their presence is unlikely since their migration is highly constrained by the barriers present downstream of the sampling site.

2.3 | Sampling and DNA Isolation

Five 2L water samples were collected on site using autoclaved water bottles, transported on ice and in the dark to the laboratory and filtered within 6 h upon collection. Filtration was performed using high-capacity nitrocellulose filters (700 cm² filter area) with 0.45 µm pore diameter (GoPro, Proactive environmental products, Lakewood Ranch, Florida, USA), attached to sterile tubing and a peristaltic Millipore vacuum pump. Upon filtration, the filters were filled with a 40 mL solution of lysis buffer (67 mM Tris HCl, 26.5 mM EDTA, 143.9 mM guanidine thiocyanate, 228.9 mM trisodium phosphate dodecahydrate, 34.2 mM sodium chloride; pH=9) and water lysis additive (1.2% SDS) (ratio of 3:1), following the protocol of Sellers et al. (2018), and sealed with sterile rubber stoppers. All filters were kept in the fridge until eDNA extraction (up to 5 months). Details on DNA isolation are available in Appendix S1: Supplementary methods.

2.4 | Shotgun Library Preparation and Sequencing

Library preparation and sequencing was performed by the Norwegian Sequencing Centre (www.sequencing.uio.no), University of Oslo, Norway. Library preparation followed the SMARTer ThruPLEX (Takara Bio, Kusatsu, Shiga, Japan) low-input sample prep protocol, including sonication of samples on a Covaris LE220 aiming at ~350 bp fragments. The final pooled libraries were sequenced using two runs of the Illumina MiSeq v2 500 cycles aiming at a total of 4 million reads per extraction (all samples run in both runs).

2.5 | Quality Control of Shotgun Reads

Shotgun sequencing reads were quality-trimmed with Trimmomatic (Bolger et al. 2014) by excluding adapters at the 3'-end of the reads and regions with quality scores below 20. The resulting paired reads were merged using PEAR (Zhang et al. 2014) with default parameters. For further analyses, both merged and unmerged reads longer than 50 bp were considered.

2.6 | Metagenomics Reference Databases

Taxonomic assignment was done against four reference databases (Table S1): (a) all non-redundant *nucleotide* sequences from GenBank (accessed in December 2020); (b) all transcriptomes of fish families present in Portuguese freshwater bodies available in GenBank (*transcriptomes*, accessed in May 2021) and detected in the *nt* database; (c) whole genome sequences of species of fish families found in Portuguese freshwater bodies available in GenBank and detected in the *nt* database (*genomes_v1*; accessed in May 2021); (d) an extended fish genome database including the most recent whole genome assemblies of species in fish families detected with databases (a) to (c) (*genomes_v2*; accessed in July 2022), plus newly generated whole genome sequence data from fish taxa present in Portuguese river drainages (Table S1). The *nt* database was chosen to allow taxonomic assignment of reads across the tree of life at different taxonomic levels. Databases (b) to (d) focused on the identification of fish species only to evaluate database requirements for classification at the species level. During data analysis and writing of this manuscript, several new fish genomes were produced by our team and others became available on GenBank, filling some of the gaps existing in the *genomes_v1* database (c). Thus, the extended *genomes_v2* database was used to evaluate how completeness of genomic reference databases might affect detections at the species level. In summary, the *transcriptomes* database (b) is composed of 233 transcriptomes from 28 fish families and 120 species. The *genomes_v1* and *genomes_v2* databases (c) and (d) contain 155 and 172 species (one genome per species) covering 35 and 20 fish families, respectively. Thus, while the extended *genomes_v2* database contains fewer families, it is more targeted at the diversity present in the Ave River (more species).

To test our ability to distinguish between true and false positives and between true and false negatives, the *transcriptomes* and both *genomes_v1* and *genomes_v2* databases included species known to occur at the sampling site (whenever available; i.e., included as potential true positives if they are present in the eDNA sample) but also all other species from all fish families known to occur in Portuguese freshwater systems (Collares-Pereira et al. 2021) with available transcriptome and genome sequence data. Additionally, marine fish families commonly present in human diet (e.g., Gadidae, Moronidae) were included in the databases since their eDNA may be present in the water samples. Finally, several additional species not expected to have any match were added as negative controls. In case no genome or transcriptome sequences were available for a given target fish family (as in families Cobitidae, Atherinidae and Loricariidae), members of the next higher taxonomic group were included. In contrast, given the large number of genomes

available for Cichlidae (> 500), only species from the subfamily Cichlasomatinae, with one representative introduced to Portugal (Chameleon cichlid, *Australoheros facetus*), were included to save computational resources.

Possible contaminants in databases (b) to (d) were detected with Kraken2 (Wood et al. 2019) by comparing the sequences with common contaminants plus some curated fish genomes. Putative contaminants in the genomes were surveyed using the following kraken databases: bacteria, archaea, viruses, protozoa, fungi, human, plasmid and UniVec. Genome sequences from 18 fish species were included to attract specific matches (Table S2). Since non-assigned scaffolds can contain contaminants as well, only scaffolds categorised as chromosomes from these 18 species were included. Genome sequences matching any non-fish taxa were excluded from the assemblies using the `extract_kraken_reads.py` script from KrakenTools (Lu et al. 2022). The resulting genome sequences were compiled in a final BLAST database where low complexity regions were masked with DustMasker (Morgulis et al. 2006).

2.7 | Taxonomic Assignment of Metagenomics Reads

Taxonomic assignment of metagenomics sequence reads was performed by megablast (Korf et al. 2003) searches against the four reference databases, using both the merged and unmerged reads > 50 bp. For the latter, the read 1 and read 2 files were blasted separately and combined afterwards. BLAST searches only considered hits with a minimum identity value of 90%, query coverage of 90%, and E-value of 1^{-10} . These were saved in tabular format containing the following information: query id, subject id, taxonomic id, percentage of identity, query length, alignment length, E-value and bit score.

Outputs from megablast were processed in four consecutive steps using custom scripts available at [github](https://github.com/mcurto/eDNA_Metagenomics) (https://github.com/mcurto/eDNA_Metagenomics) to obtain a final taxon list from the samples. These steps consisted of (1) combining information of hits from unmerged reads, (2) selecting the best matches for each read, (3) extracting the lineage information for each match, and (4) summarising the lineage information up to the most recent common ancestor across all matches. Details can be found in the Appendix S1: Supplementary methods.

In addition to the classification for each database, we combined the results from the four databases by finding the best hit per read across them. In cases of multiple best hits, the most recent common ancestor was taken as the final taxon. The SEQDIST approach was then implemented to exclude false positives by considering sequence identity distributions with a mode below 99.5% as false positives.

Computational analyses were performed on the Saga Cluster owned by the UiO and the Norwegian metacenter for High Performance Computing (NOTUR) and operated by the UiO Department for Research Computing (<https://www.hpc.uio.no>).

2.8 | Filtering Metagenomic Matches to Minimise Database Contaminants

Despite filtering out contaminants in genome assemblies, BLAST matches of the “cleaned read data” still showed sub-optimal hits (non-best hits) solely to non-fish taxa, namely bacteria. Therefore, as an additional quality control criterion, only reads matching fish sequences in the *nucleotide* database (a) were considered for downstream analyses. This decision is based on the expectation that, given the large number of genomic resources for fish species included in the reference database, any read matching a fish locus should match an orthologous region from other fish taxa with a lower sequence identity.

In a preliminary experiment, we observed that highly conserved regions, such as some ribosomal loci and low-complexity regions, can generate false positive species detections, especially when reference databases are incomplete. These conclusions were drawn from mapping the metagenomics reads to two fish genomes: *A. anguilla* (GCF_013347855.1) that is present in the study river system, and *Oreochromis niloticus* (GCF_001858045.2) that is absent. We assume that genomic regions with higher-than-average read mapping depth are hotspots for unspecific matches, which was the case for ribosomal loci and low-complexity regions (data not shown). Based on these observations, low-complexity regions were masked in the reference databases with DustMasker while building the blast database and reads matching ribosomal loci in the *nt* database were excluded from analyses.

2.9 | Reciprocal and Recursive BLAST Comparisons

For three genera (*Gobio*, *Luciobarbus* and *Squalius*), we found matches in the metagenomics data for closely related species. In principle, differences in assembly quality among very closely related species may contribute to false positives. Fragmented and incomplete assemblies of species that are truly present may fail to attract some of their conspecific matches, which would in turn align with high identity to a nonspecific better-quality interspecific assembly. If that is the case, we expect that reads matching the higher quality assembly (i.e., an assembled region missing in the assembly of a close relative) will not match the lower quality one. To test this hypothesis, we conducted reciprocal BLAST of the eDNA reads detected between genome assemblies of species pairs where this might be happening: *G. gobio* (higher quality) versus *G. lozanoi* (lower quality); *L. bocagei* versus *L. comizo* (both low quality); *S. cephalus* (higher quality) versus *S. squalus* (lower quality); *S. cephalus* (higher quality) versus *S. carolitertii* (lower quality).

2.10 | DNA Metabarcoding With 12S Mitochondrial Gene

We evaluated the sensitivity of fish species detection using eDNA metagenomics by comparing its performance versus the current “gold standard” produced by eDNA metabarcoding,

using the MiFish-U primers from Miya et al. (2015). Amplicon library preparation was done with the dual index strategy similar to Kozich et al. (2013) with a protocol adapted to the primers used (Appendix S1: Supplementary methods).

The obtained reads were used to define amplicon sequence variants (ASVs) and taxonomic assignment was done via BLAST to the *nt* database and a custom database comprising 12S sequences of freshwater fish species missing in the *nt* database (see Appendix S1: Supplementary methods). Errors originating from index swapping were minimised by applying a correction factor as described by Hambäck et al. (2021). We also used the Bayesian approach described by these authors to define variable thresholds of the minimum number of reads. The sequencing run included eDNA samples from marine ecosystems amplified using the same marker. To minimise the false detection of marine taxa due to index swapping, organisms detected in the Ave River samples with higher read numbers in the marine samples were excluded; thus, only matches to species that spend part of their life cycle in freshwater remained as likely true positives.

2.11 | Read Length as a Lens Into Residency Time and Source of eDNA

The DNA found in environmental samples is a mixture of molecules from different organisms and sources, with variable lengths depending on the time and agent(s) of decay (Jo and Yamanaka 2022). Upon release of intact DNA molecules into the environment, eDNA fragmentation is expected to increase over time, whereby recently shed eDNA will consist of longer molecules (and read lengths) compared to older eDNA (Jo et al. 2017). Beyond residency time, the level of eDNA decay (and read length) may vary depending on the source organisms: living microorganisms will be present in any environmental sample and will contribute intact intracellular DNA molecules to the eDNA sample (Pathan et al. 2021). Their eDNA will exhibit longer read lengths than extracellular eDNA originating from lysed cells. Finally, environmental samples may include DNA of organisms from habitats different from the one sampled. Freshwater systems receive input material from the whole drainage area (i.e., of terrestrial origin) as well as from airborne material (Câmara et al. 2021; Deiner et al. 2016). Thus, eDNA originating from habitats other than the one sampled will undergo decay during transport (Jo et al. 2017). Decay is faster in larger molecules than in shorter ones, meaning that size heterogeneity may be greater in eDNA fragments released more recently into the environment than in those residing for longer periods (Jo et al. 2017). Accordingly, read length standard deviation (SD) of matches to microbes and freshwater organisms should be higher than that of matches to macroscopic and marine taxa. These expectations were tested by comparing read length across taxa and habitats of origin.

All taxa detected with identity distributions congruent with being true positives were classified according to size as either microorganisms (all unicellular organisms or planktonic multicellular organisms) or macroorganisms (all other), and according to habitat as freshwater, terrestrial or marine. Some organisms

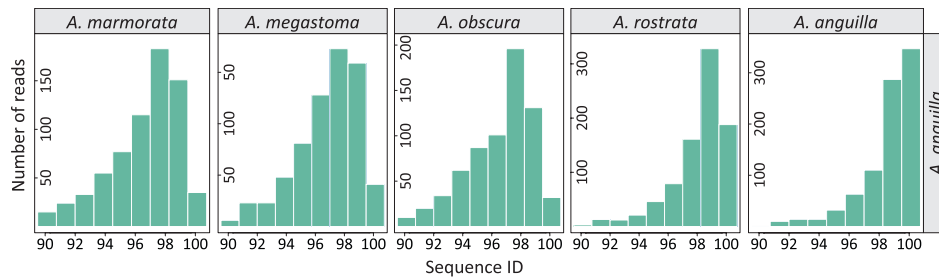


FIGURE 1 | Intra- and interspecific identity distributions of 1000 shotgun whole-genome sequences. Short-reads of *A. anguilla* (SRR12854530) are compared to the genomes of five eel species. The intraspecific comparison (right) shows a distribution with mode at 100% sequence identity, while interspecific comparisons display modes < 99.5% and are less left-skewed. For additional comparisons, see Figure S1.

may be found in multiple habitats and were not considered in the analysis. Very few organisms from brackish and airborne environments were detected (Table S2), thus, they were excluded to prevent low sample size biases in the analyses. All these classifications were based on literature searches using known databases (e.g., <https://www.algaebase.org/>, WoRMS—World Register of Marine Species, NCBI, GBIF) and google scholar searches using organism names as query. For (microbial) taxa represented by specific strains, the information was taken at the strain level. Taxa missing information for size and/or habitat were not included in the analysis. Read length information taken from matches to all databases was retrieved for statistical analysis. Models for read length are Generalised Linear Mixed Models (GLMMs) fitted with Gaussian distributions and with species identity included as a random intercept factor. Models for read standard deviation are Generalised Linear Models (GLMs) fitted with Gaussian distributions. Models were fitted in R (R Core Team 2021) using the function *lmer* of lme4 package (Bates et al. 2014) for GLMMs and the function *glm* of stats package (R Core Team 2021) for GLMs.

3 | Results

3.1 | Accuracy of Multilocus SEQIDIST Taxon Identification at Species Level

As a proof-of-principle, we used publicly available shotgun whole-genome sequence data (sgWGS) from different fish species to show that sequence identity distributions of blast hits against assemblies of conspecifics display a highly left-skewed distribution peaking at 100% sequence identity. In turn, as the genetic distance between species increases, BLAST hits against heterospecific genomes peak at < 99.5% sequence identity. These expectations were tested through 139 comparisons of short read data from genomic DNA against reference genomes (Table S1). From an initial 1000 genomic DNA reads used per sample, between 85 to 966 reads matched a certain reference genome with at least 90% identity. All intraspecific comparisons resulted in distributions with a mode at 100% identity, the true positives, as most reads of a species did not differ from its reference genome (Figure 1; Figure S1). Most interspecific comparisons, however, resulted in distributions with a mode below 100% (Figure 1; Figure S1). The exceptions being: *Luciobarbus bocagei* vs. *L. comizo*, *Micropterus salmoides* vs. *M. nigricans*, *Salmo trutta* vs. both *S. ischchan* and *S. caspius*, *Squalius carolitertii* vs. *S. pyrenaicus* and *S. squalus* vs. *S. cephalus* (Figure S1).

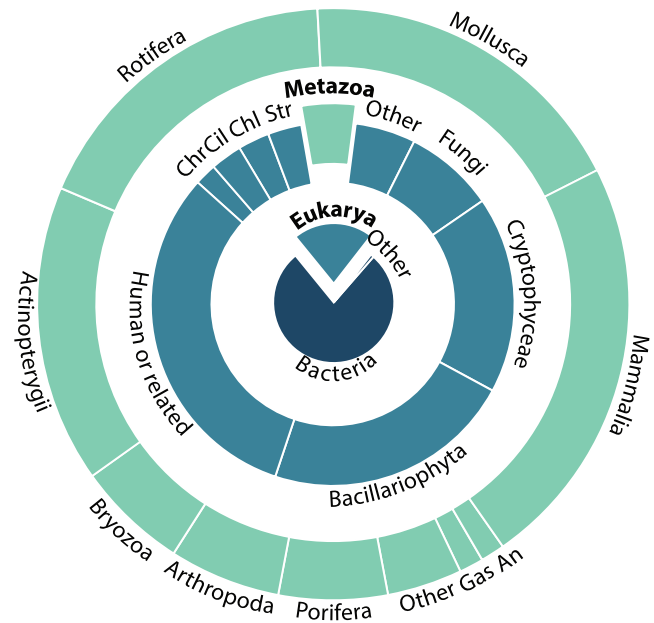


FIGURE 2 | Proportion of metagenomic reads assigned to taxa across the tree of life using the GenBank nucleotide (*nt*) database. Human or related (other primates) are plotted independently from the remaining metazoans to facilitate the visualisation of the relative abundances of animal taxa. Some taxon names are abbreviated: An—Annelida; Chl—Chlorophyta; Chr—Chrysophyceae; Cil—Ciliophora; Gas—Gastrottricha; Str—Streptophyta.

3.2 | Whole Community Biodiversity Assessment

Of the 31,519,078 shotgun sequencing reads produced, 84.5% (26,627,319) passed the quality control and 82.1% (21,855,943) were successfully merged (Table S3). Nevertheless, only 4% (1,167,349) and 0.8% (210,547) matched the *nt* database at the > 90% and > 99% identity thresholds, respectively. The SEQIDIST approach recovered 1933 taxa, from 73,579 reads with identity above 99%, across all domains of life (Figure 2), encompassing 66 phyla, 427 families and 1413 species of micro- and macroscopic organisms (Table S2). Most of the reads are assigned to bacteria (77.7%), with 21.4% assigned to eukaryotes and 0.63% and 0.24% to archaea and viruses, respectively (Figure 2). Within Eukarya, we detect 12 animal (metazoan) phyla (25 classes, 54 families, 65 genera), 2 green plant phyla (8 classes, 58 families, 81 genera), 5 fungi phyla (14 classes, 26 families, 26 genera) and 11 protist phyla (24 classes, 76 families, 108 genera). Microalgae (e.g.,

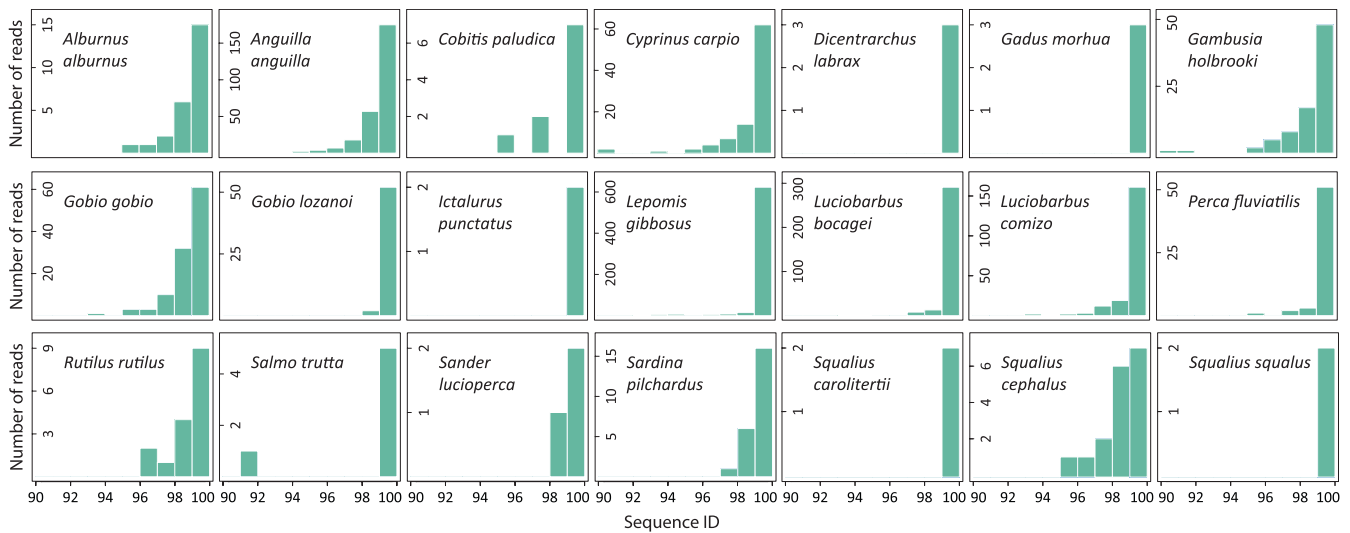


FIGURE 3 | Sequence identity distributions of metagenomics reads matching fish taxa of the Ave River. Only left-skewed distributions with mode at 100% are shown (i.e., congruent with true positives), using the combined result of the four reference databases. Three are marine species common in human diet whose eDNA likely reached the system through wastewater (*Dicentrarchus labrax*, *Gadus morhua* and *Sardina pilchardus*), while *L. comizo* and *S. cephalus* are likely false positives due to a combination of ILS and introgression.

Bacillariophyta and Cryptophyceae) were the most abundant eukaryotic taxa detected. We detect not only freshwater organisms but also taxa from nearby marine (20 phyla, 31 classes, 49 orders, 56 families and 69 genera), brackish (5 phyla, 8 classes, 10 orders, 10 families and 9 genera) and terrestrial environments (30 phyla, 57 classes, 118 orders, 182 families and 316 genera) (Table S2).

3.3 | Fish Diversity and Database Comparison

Considering the top blast hits of eDNA metagenomics reads against the four databases, 3346 and 2342 reads match fish taxa with >90% and >99% identity, respectively. The *genomes_v1* and *genomes_v2* databases provide the highest fraction of the blast hits matching fish taxa, although the extended *genomes_v2* database clearly outperforms the others (96.6%, vs. 28.6% with *genomes_v1*). Indeed, most of the fish reads (67.4%) match only the extended *genomes_v2* database (Figure S2). The *transcriptomes* (9.5%) and *nt* (6.1%) databases have fewer than 10% of the fish reads producing blast hits above the identity cutoffs.

The combination of all four databases resulted in matches to 151 fish taxa with identity above 90% and 48 above 99%. However, only 21 of these have multilocus identity distributions consistent with true positive detections using SEQIDIST (Figure 3; Table S4). These included three marine species common in the local human diet (*Dicentrarchus labrax*, *Gadus morhua* and *Sardina pilchardus*), whose eDNA likely reached the system through wastewater. Thus, while their detection is true, their presence in the riverine system is not.

Three genera (*Luciobarbus*, *Gobio* and *Squalius*) showed distributions with modes at 100% for two congeners. We conducted reciprocal and recursive blast searches to exclude the possibility of false positives caused by differences in assembly quality

between very closely related species. All reciprocal blast distributions had modes at lower identity values, typical of interspecific comparisons (Figure S3). Once we ran recursive blasts, in which the reads of the reciprocal blast are remapped to the first genome, the results show similar identity distributions typical of intraspecific comparisons (Figure S4). Altogether, a hypothetical artefact stemming from differences in assembly quality can be excluded, thus ruling out the possibility that assembly quality is driving these detections.

The intraspecific comparisons with identity distribution modes at 100% included species that have been previously detected for the Ave River using independent methods (*Luciobarbus bocagei*, *G. lozanoi*, *S. carolitertii*, *S. squalus*), and taxa that would constitute new records (*L. comizo*, *S. cephalus* and *G. gobio*). For the proof-of-principle test using sgWGS, intraspecific comparisons with identity distribution modes at 100% were already found for *L. bocagei* vs. *L. comizo* and *S. squalus* vs. *S. cephalus*. These are likely a consequence of introgression or incomplete lineage sorting (ILS) indicating that one of these species per pair is likely a false positive (Note S1). This was considered to be the case for *L. comizo* and *S. cephalus* since they are not historically present in the Ave River.

After excluding *L. comizo* and *S. cephalus*, as well as the three marine species, a final list of 16 species detected was obtained with the SEQIDIST metagenomics approach for the Ave River samples. All of these but one (*Ictalurus punctatus*) were detected by the *genomes_v2* database (Figure S2; Table S4) outperforming the *genomes_v1* that detected only 10 species. Both *genomes* databases outperformed the *nt* and *transcriptome* databases that retrieved only five and four taxa, respectively.

The addition of new genomes in the *genomes_v2* database reduced the number of matches at lower identity in distributions of related species. This was evident in *Cyprinus carpio*, which had a multimodal distribution for the *nt* and *genomes_v1*

databases and a unimodal distribution peaking at 100% for the *genomes_v2* database (Figure S5). Another example was the case of *Micropterus nigricans* (Centrarchidae), which, at the *genomes_v1* database, showed an identity distribution congruent with a false positive (ID mode = 93%) with a high number of matches (Figure S5). By including the *L. gibbosus* (Centrarchidae) genome in our extended *genomes_v2* database, the number of reads matching the *M. nigricans* genome (even with a more complete assembly) is greatly reduced, while those matching *L. gibbosus* show a highly skewed identity distribution centered at 100% identity.

3.4 | Fish eDNA Metabarcoding Results

The MiFish-U marker was used as the “gold standard” of eDNA analysis to assess the relative performance of metagenomics. Around 1.4 M reads were obtained, of which 494,926 passed all quality control, denoise and chimera filtering steps, resulting in 198 ASVs. 103 ASVs are assigned to fishes (464,495 reads) from which 95 are kept after correcting for index swapping according to Hambäck et al. (2021). These ASVs match 40 fish taxa, of which 24 are marine taxa that exhibited the highest read numbers in the marine samples sequenced in the same run as the Ave River samples. These matches likely result from index swapping, or DNA may originate from wastewater (Table S5). Therefore, we only consider matches to species that spend part of their life cycle in freshwater as likely true positives, resulting in 14 fish species detected in the Ave River samples (Tables S4 and S5).

3.5 | eDNA Residency Time Inferred From Read Length Distributions

To test the hypothesis that read length could serve as a predictor of eDNA source, with longer residency times and greater transport distances in the environment expected to result in more decayed eDNA, we categorised all taxa detected in the matches according to their size and native habitat, resulting in: microorganisms—1,378 taxa and 44,242 reads; and macroorganisms—205 taxa and 6027 reads; freshwater—570 taxa and 30,347 reads; marine—100 taxa and 2188 reads; terrestrial—682 taxa and 12,852 reads; brackish—10 taxa and 423 reads; and airborne—1 taxon and 1 read (see Table S2 for details). Because of their low numbers, brackish and airborne taxa were not considered in further analyses. Habitat information was either not found or reported as multiple habitats for 570 taxa (27,768 reads), while size information was not found or was ambiguous for 350 taxa (73,579 reads). The final datasets of read lengths used in the following analyses can be found Appendix S1: (Supporting Information 6–7). Using Generalised Linear Mixed Models (GLMMs) for read length and Generalised Linear Models (GLMs) for read length SD, we find that reads assigned to microorganisms are significantly longer and have more heterogeneous lengths (mean = 253.7, SD = 118.6) than those of macroorganisms (mean = 190.7, SD = 105.0), while reads assigned to freshwater organisms are longer and more heterogeneous in length (mean = 264.5, SD = 119.0) than those assigned to marine (mean = 258.1, SD = 116.1) and terrestrial (mean = 210.5, SD = 115.9) organisms (Figure 4; Table S6).

A similar effect of native habitat is found for plant taxa (Figure 4; Table S6). Reads assigned to freshwater plants (12 taxa and 41 reads) are longer and have more heterogeneous lengths (mean = 241.6, SD = 124.1) than those matching terrestrial plants (82 taxa and 411 reads; mean = 153.8, SD = 92.4). Among terrestrial plants, taxa used in agriculture tend to show lower read length (mean = 140.8, SD = 86.9) compared to taxa naturally found around and along the riverbank (mean = 173.4, SD = 97.6), although the differences are not statistically significant (Figure 4; Table S6). While eDNA from wild terrestrial plants is integrated into the aquatic system through remains of individuals growing nearby, eDNA of agricultural taxa is more likely to have been transported through the sewage system in addition to going through digestive systems, and thus reside in the environment for a longer time.

Considering the specific case study of fishes, reads assigned to freshwater taxa show significantly longer and more heterogeneous lengths (14 taxa and 1577 reads; mean = 223.8, SD = 107.9) compared to reads assigned to marine taxa (3 taxa and 24 reads; mean = 107.5, SD = 43.9). Within freshwater taxa, *Salmo trutta* (mean = 148.0, SD = 62.8) and *Perca fluviatilis* (mean = 116.9, SD = 37.0) also have shorter read lengths, which may be due to increased eDNA degradation associated with transport from upstream populations and thus reflect longer residency times (Figure 4).

4 | Discussion

4.1 | Excluding False Positives With SEQIDIST

SEQIDIST solves the high prevalence of false positives of ‘traditional’ metagenomics when classifying at the species level (Simon et al. 2019). Taxon detection and discovery using individual sequence matches ultimately leads to unacceptably high rates of false positives, while simultaneously overseeing false negatives (e.g., fishes (Manu and Umapathy 2023)). Alternatively, a ‘pseudo-taxonomic approach’ could be implemented, where read classification is performed as putative molecular operational taxonomic units (Cowart et al. 2018), or as classifications to the closest higher taxon, making taxonomic assignments at higher taxonomic levels (e.g., genus (Manu and Umapathy 2023)). However, that is done at the cost of information loss and increased taxonomic composition uncertainty (increasing both false-negative and false-positive assignments). Additionally, matches to erroneous taxa can occur even at high taxonomic levels (i.e., families, orders) with deceiving statistical confidence in conserved genomic regions of high sequence identity, thus not avoiding false positives despite trading-off taxonomic certainty.

The SEQIDIST framework takes all blast hits to each individual taxon to make a species identity profile. By using sgWGS data from different fish species, we demonstrate that profiles resulting from conspecifics peak at 100% sequence identity, contrasting with heterospecific comparisons that peak at lower values. Such a property of identity distributions makes false positives become easily discernible. However, some interspecific comparisons had a distribution profile with a mode at 100%. All these correspond to very closely related species, hybrid species and

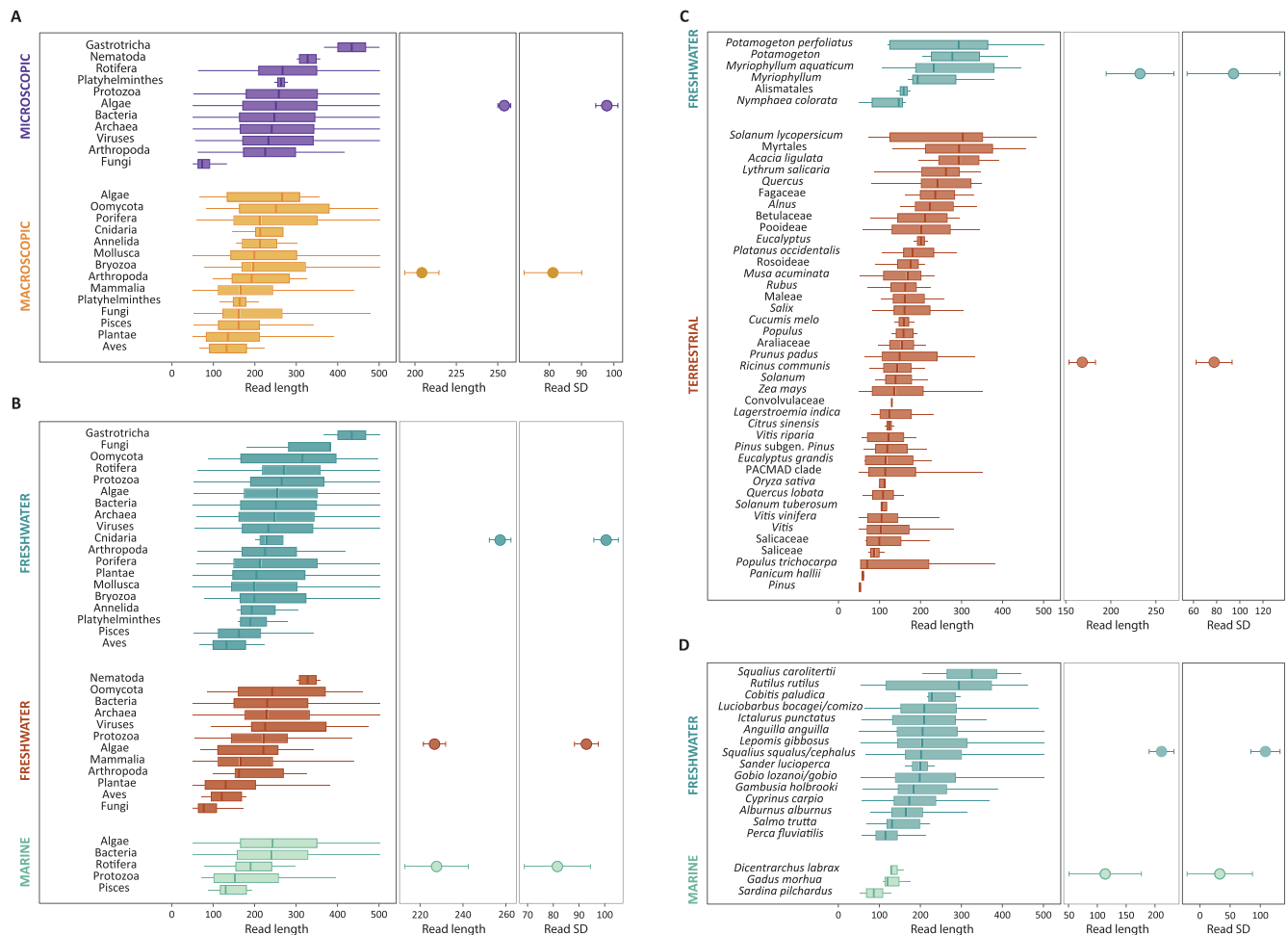


FIGURE 4 | Read length and standard deviation across groups and species. Comparisons between (A) macroscopic vs. microscopic organisms; and different habitats across (B) all organisms; (C) terrestrial plants and (D) fishes. Read length per taxa included in each category is summarised in the left panel by boxplots with median (middle bar), inter-quartile range (box) and extremes up to 1.5 times the inter-quartile range (whiskers). Middle and right panels show 95% confidence intervals based on a Generalised Linear Model (GLM) and Generalised Linear Mixed Model (GLMM), respectively. Model statistics are presented in Table S6.

species for which natural hybridisation has been reported to be common (e.g., Gante et al. 2015; Gouskov and Vorburger 2016; Segherloo et al. 2021). Thus, if the individual used for building a reference genome has been introgressed, its genome will show an enrichment of high identity reads between the hybridising species. Additionally, incomplete lineage sorting contributes to the retention of ancestral variants, resulting in patterns of shared variation across closely related taxa that do not match the species tree (Pollard et al. 2006). Therefore, the high sequence identity found between some closely related species is likely a consequence of shared genomic variation caused either by ILS or introgressive hybridisation. In such cases, there is the potential for false positives, but this can be at least partially controlled for.

Nevertheless, possible misdetections would happen mostly in the scenario of species discovery, where true species presence could be difficult to disentangle from introgression and ILS, especially in systems with many closely related species. In the case of species monitoring in better known systems using eDNA metagenomics, determination of species identity can take into account other sources of information, such as a pre-existing knowledge of species diversity at the site based on historical

sampling data, and knowledge of the phylogenetic relationships between the putative target species co-occurring in the system, or with previous evidence of introgression (e.g., eliminating off-target species with closely-related congeners in the system from databases, or monitoring detection proportions). In practice, this particular effect resembles the impact of low marker resolution, ILS and introgression also observed in (e)DNA (meta)barcoding. Thus, proper vouchering in natural history collections of specimens used to produce the reference nuclear genomes and mitogenomes (Buckner et al. 2021) is a prerequisite for accurate species identification using eDNA, and to rule out possible impacts of introgressive hybridization or other marker limitations on molecular information content.

A further challenge of the S approach is assessing distributions composed of only a few matches. These matches may, by chance, be overrepresented by conspecific sequences containing mismatches to the reference—whether due to natural variation or sequencing errors—or by heterospecific matches from conserved regions identical to the reference. Both scenarios can lead to incorrect assessments. Nevertheless, in our case, we believe that the oversampling of conserved matches by chance is the

least likely of the two scenarios, given the high number of high-quality fish genomes from multiple families included in the reference database. Reads from conserved regions are more likely to align with multiple genomes, yielding matches at higher taxonomic levels. In any case, caution should be exercised when interpreting distributions with a low number of matches, and the possibility of implementing a minimum read count threshold should be evaluated.

4.2 | Describing Diversity Across the Tree of Life

SEQIDIST accurately describes biological communities across the tree of life. We identified 1933 taxa from all domains, which is at the same order of magnitude as recent eDNA metagenomics studies from more diverse tropical brackish lagoons in India and the Southern Ocean (Coward et al. 2018; Manu and Umapathy 2021, 2023), despite our use of more stringent filtering (smaller E-value and higher sequence identity). Such a relatively high number of taxa was primarily due to the ability of SEQIDIST to confidently assign matches at species levels. Nevertheless, 96% of reads were not assigned at all, highlighting the amount of biodiversity information still unmapped in the generated dataset. Most matches found were to microorganisms, such as bacteria and diatoms that are generally abundant and largely represented by intracellular eDNA that is favourably sequenced (Singer et al. 2020; Stat et al. 2017). In addition, bacterial eDNA tends to be more easily assigned to taxa due to its low complexity and better representation in reference databases (Singer et al. 2020; Stat et al. 2017). These proportions may also reflect the high degree of eutrophication at the sampling site (A. Veríssimo and G. Riccioni, pers. observation), which favours microbial activity.

Much of the diversity detected was from non-freshwater organisms, whose presence can be explained by multiple processes. Terrestrial eDNA can be originated not only from species that contact the river directly but also from biomass that is added to the river through sewage systems (Macher et al. 2021). This assessment is supported by the high abundance of human eDNA and the detection of organisms that are commonly found in wastewater, used for human consumption or associated with agriculture and farming (Table S1). For instance, the detection of some marine vertebrate taxa (e.g., marine fishes) in freshwater samples is likely attributable to human consumption; however, occasional intrusions of some marine species, such as the European seabass (Almeida et al. 2024), may also occur. Other sources of eDNA reflect the wide spatial scale of biodiversity represented in freshwater systems. For instance, marine microalgae have been detected in freshwater lakes and other terrestrial environments near the ocean due to the aerosolisation of sea surface water and inland wind transport (Câmara et al. 2021). In fact, many organisms are part of the so-called bioaerosols and can be transported for long distances (Razjigaeva et al. 2022). In our study, this phenomenon is exemplified by the detection of microalgae like *Thalassiosira pseudonana* (Bacillariophyta) or *Crustomastix* sp. (Chlorophyta), that are marine taxa according to www.algaebase.org. Given the proximity of the sampling site to the Atlantic Ocean (~5 km in a straight line), it is plausible that the detection of marine taxa originated from marine bioaerosols.

4.3 | Discovering New Diversity With SEQIDIST

SEQIDIST's ability to exclude false positives enhances confidence in the discovery of new diversity. Of the final set of 16 species, most have been previously reported in the Ave River, but four are novel (Collares-Pereira et al. 2021), constituting new alien species for the basin (*Ictalurus punctatus*, *Perca fluviatilis*) and even for the Iberian Peninsula (*Gobio gobio*, *Squalius squalus*). Their detection is not surprising, as this drainage has been a hotspot for illegal fish introductions (Ribeiro et al. 2009; Ribeiro and Veríssimo 2014), highlighting the power of metagenomics in detecting alien species at the invasion front, generating strong hypotheses regarding species distribution and community composition. If only individual matches were considered, as many as 151 fish taxa would be detected, which would represent an unacceptably high false positive rate. Even when considering only stringent BLAST hits with >99% sequence identity and >90% read coverage, the number would decrease to 48, many of which would still be false positives. Given such uncertainty, it would be impossible to confidentially report these new occurrences to the Ave River.

The focus on fish species detection in the Ave River also unravelled the higher sensitivity of eDNA metagenomics compared to metabarcoding. eDNA metagenomics recovered 16 taxa compared to 14 using metabarcoding (Tables S4 and S5). The reference sequence database used for taxon assignment in the case of eDNA metabarcoding included a wider range of taxonomic groups than those used with the eDNA metagenomics data and contained all taxa exclusively detected by eDNA metagenomics. Thus, the poorer performance of eDNA metabarcoding in species detection was not due to an incomplete reference sequence database. The increased sensitivity of metagenomics likely reflects known limitations of metabarcoding, namely false negatives due to differences in priming efficiency across taxa (Krehenwinkel et al. 2017), PCR failure due to eDNA decay (Jo et al. 2017) since 57% of the metagenomics reads classified as fishes are shorter than the 172 bp-long barcode used, or low abundance of the barcode locus in the eDNA mix since the probability of detecting a specific mitochondrial locus is lower than anonymous nuclear regions (Note S2). Nevertheless, there are four species detected exclusively with the metabarcoding assay: *Achondrostoma oligolepis*, *Lampetra fluviatilis*, *Esox lucius* and *Alosa* sp. (Tables S4 and S5). These constitute potential false negatives in our metagenomics dataset. Among these taxa, only the genome of *Lampetra fluviatilis* was not included in the metagenomics reference databases, which could explain its non-detection. However, this is unlikely, as no matches—even at low identity—were found for the two closely related *Lampetra* species included in the database (*L. planeri* and *L. auremensis*). Thus, these likely result from a combination of low eDNA abundance (as indicated by the lowest read counts in metabarcoding) and poor genome quality in the *genomes_v2* database (with an N50 < 5 Kbp for *A. oligolepis* and *Lampetra* spp.). Rare eDNA is often only detectable through multiple replicates (Wang et al. 2022), which were used in metabarcoding but not in metagenomics. Furthermore, as mentioned above, the SEQIDIST approach may yield an incorrect assessment when only a few matches are available. For example, for *Alosa* sp., we recovered two matches with 100% and 98% identity (Supporting Information). It is possible that the lower-identity match represents a real match with a mismatch

to the reference, leading to a false positive classification. This artefact may be even more pronounced when using low-quality genomes, as they may fail to attract specific matches that help resolve identity distributions. The detection of *A. oligolepis* and *Alosa* sp. is not surprising, given the possible occurrence of the species in this region (Collares-Pereira et al. 2021). However, the occurrence of *Lampetra* sp. and *E. lucius* constitutes new records for this Portuguese drainage.

4.4 | Using Read Length as a Proxy for Decay

We detected a wide range of organisms, ranging from micro- to macroorganisms, as well as detections from nearby marine, brackish and terrestrial environments, whose eDNA molecules are expected to have different residency times and, consequently, varying degrees of decay. This allowed us to test whether read length distributions from metagenomics can be used to discriminate between local and regional transport scales and their associated residency timescales. We confirmed our expectations that eDNA of microorganisms and of freshwater organisms is on average longer and more heterogeneously fragmented. Conversely, eDNA of macroscopic organisms, and of marine and terrestrial taxa originating outside the freshwater system, is shorter and more homogeneously fragmented. These general trends are clearly illustrated by the two best-known taxonomic groups of macroscopic eukaryotes in the Ave River: plants and fishes. Aquatic plants are represented by longer reads with higher per species standard deviation than terrestrial species. Likewise, marine fish species have shorter read lengths than freshwater ones. Their DNA likely reached the riverine environment through wastewater, as they are commonly included in the local human diet. Moreover, read length distributions are congruent with differences in habitat preference: species restricted to upstream stretches are represented by shorter and more homogeneously fragmented eDNA, likely due to longer residency times in the system and higher decay. This is the case of *Salmo trutta*, whose known populations are found circa 50 km upstream of the sampling locality (Collares-Pereira et al. 2021). Discrimination between local and transported eDNA and between intracellular and extracellular eDNA is a major added value of metagenomics data for community characterisation at spatial and temporal scales. These have been pointed out as major limitations of eDNA metabarcoding as a biodiversity monitoring tool, since diversity can be overestimated by the detection of taxa absent at the sampling locality (Deiner et al. 2017).

4.5 | Impact of Database Type and Completeness on SEQDIST

Adding whole genome sequences greatly improved the detectability of fish communities with eDNA metagenomics. These results highlight the benefit of using reference databases with a good taxonomic representation of whole genome sequences. Nevertheless, genome data are still lacking for most species (Lewin et al. 2018, 2022). When using incomplete databases, eDNA sequence data originating from taxa not represented in the reference database would likely match reference genomes of related taxa, resulting in a SEQDIST distribution with a mode below 100% of sequence identity. Thus, sequence identity

distributions to a close taxon provide important information on false positives and dark diversity (false negatives) in environmental samples. This was clearly shown by the decrease in the number of matches at low identity to the *Cyprinus carpio* and *Micropterus nigricans* genomes in the *genomes_v2* database compared to the more incomplete *genomes_v1*. These matches at low identity were converted in high identity matches to newly added genomes of related species. Thus, adding genomes to reference databases (i.e., *genomes_v2* database) illuminates dark diversity, maximises species detection, and refines identification in eDNA samples. These results highlight the importance of gathering genome sequences for all diversity across the tree of life.

Our results also show that these reference genomes do not need to be of high quality. While differences in assembly quality of reference genomes of close relatives could, in principle, contribute to false positives, we determined that assembly quality is unlikely to play any major role in detection accuracy using reciprocal and recursive BLAST searches against assemblies of varying qualities available from congeners.

4.6 | Wide Applicability and Archival Potential of eDNA Metagenomics

Sensitivity and accuracy in species detection using SEQDIST improve with whole genome reference databases. Thus, the potential of eDNA metagenomics datasets as repositories of biodiversity information, akin to environmental vouchers, will only increase as more genomic resources become available globally (Ebenezer et al. 2022; Lewin et al. 2018, 2022). These resources will allow a more comprehensive description of past (Kelman and Moran 1996), present, and future biodiversity from eDNA metagenomics, offering a much-needed historical perspective of community diversity changes. eDNA metagenomics enhances whole biodiversity monitoring across the tree of life, speeds up the discovery of hidden biodiversity, and ultimately helps address the taxonomic impediment with integrative approaches (Zamani et al. 2022). Analytical strategies could provide access to hidden biodiversity by implementing approaches that describe diversity independently of reference databases (Cordier et al. 2019) and further help understand the scale of biodiversity loss the planet is currently facing.

Author Contributions

M.C., A.V., C.D.S., F.R., M.J.A. and H.F.G. contributed to the conceptualization and experimental design of the study. Field work was conducted by A.V., G.R. and F.R., while laboratory work by A.V. and G.R. Sequence data were analysed by M.C., while statistical analysis was performed by C.D.S. H.F.G. contributed throughout the analysis's steps. M.C. and H.F.G. wrote the first draft of the manuscript that was revised and edited by the remaining authors. All authors contributed to the article and approved the submitted version.

Acknowledgements

This work was funded by the Portuguese Foundation for Science and Technology (FCT) under the project ENVMETAGENOMICS (PTDC/BIA-CBI/31644/2017). Additional funds were received from FCT through the strategic plans of the research centres MARE (UIDP/04292/2020, <https://doi.org/10.54499/UIDP/04292/2020> and

UIDB/04292/2020, <https://doi.org/10.54499/UIDB/04292/2020>), cE3c (UIDP/BIA/00329/2020, <https://doi.org/10.54499/UIDP/00329/2020>) and CIBIO/InBIO (UIDP/50027/2020, <https://doi.org/10.54499/UIDP/50027/2020>), and through the project LA/P/0069/2020 (<https://doi.org/10.54499/LA/P/0069/2020>) granted to the Associate Laboratory ARNET. FCT provided individual contracts to F.R. (<https://doi.org/10.54499/2020.00482.CEECIND/CP1595/CT0001>) and to A.V. (<https://doi.org/10.54499/DL57/2016/CP1440/CP1646/CT0001>); M.C. and G.R. were supported by research contracts under the scope of ENVMETAGENOMICS. M.C.'s work was also supported by the Horizon 2020 Programme under the Grant Number 857251. H.F.G. was also supported by KU Leuven Research Fund grant STG/21/044. Finally, the work was also supported by National Funds through FCT-Fundação para a Ciência e a Tecnologia in the scope of the project LA/P/0048/2020.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Metagenomic and metabarcoding raw sequencing data are available at the European Nucleotide Archive SRA repository associated with the bioproject [PRJEB49862](https://www.ncbi.nlm.nih.gov/bioproject/PRJEB49862). Barcode sequences for the 12S marker were deposited under the accession numbers [OV522633](https://www.ncbi.nlm.nih.gov/nuccore/OV522633) to [OV522672](https://www.ncbi.nlm.nih.gov/nuccore/OV522672). Genome sequences produced in this study and corresponding short read data were deposited in GenBank under the bioproject [PRJNA902967](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA902967). We also provide Appendices S1–S2 including the sequence identity distribution for all matches found with the four databases and the combination of them (Data S1–S5) and read length information used in the comparison between taxa size scales and habitat (Data S6–S7). These files can be downloaded at https://github.com/mcurto/eDNA_Metagenomics. All scripts and codes used in the bioinformatic analysis are available and described at https://github.com/mcurto/eDNA_Metagenomics ([10.5281/zenodo.14205100](https://doi.org/10.5281/zenodo.14205100)).

Benefit-Sharing Statement

There are no benefits to report.

References

- Almeida, R., S. Tanner, C. S. Mateus, F. Ribeiro, and B. Quintella. 2024. "Not So Much a Sea Bass: Divergent European Sea Bass (*Dicentrarchus labrax* L.) Freshwater Incursions." *Journal of Fish Biology* 104, no. 4: 1241–1246.
- Bates, D., M. Mächler, B. Bolker, and S. Walker. 2014. "Fitting Linear Mixed-Effects Models Using lme4." *arXiv*. Preprint arXiv:1406.5823.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30, no. 15: 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
- Buckner, J. C., R. C. Sanders, B. C. Faircloth, and P. Chakrabarty. 2021. "The Critical Importance of Vouchers in Genomics." *eLife* 10: e68264.
- Bushnell, B. 2014. "BBMap: A Fast, Accurate, Splice-Aware Aligner." Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).
- Butchart, S. H. M., M. Walpole, B. Collen, et al. 2010. "Global Biodiversity: Indicators of Recent Declines." *Science* 328, no. 5982: 1164–1168. <https://doi.org/10.1126/science.1187512>.
- Câmara, P. E. A. S., M. Carvalho-Silva, O. H. B. Pinto, et al. 2021. "Diversity and Ecology of Chlorophyta (Viridiplantae) Assemblages in Protected and Non-Protected Sites in Deception Island (Antarctica, South Shetland Islands) Assessed Using an NGS Approach." *Microbial Ecology* 81, no. 2: 323–334.
- Clare, E. L., C. K. Economou, F. J. Bennett, et al. 2022. "Measuring Biodiversity From DNA in the Air." *Current Biology* 32, no. 3: 693–700. e5. <https://doi.org/10.1016/j.cub.2021.11.064>.

- Collares-Pereira, M., M. Alves, F. Ribeiro, et al. 2021. *Guia dos Peixes de Água Doce e Migradores de Portugal Continental*. Edições Afrontamento.
- Cordier, T., A. Lanzén, L. Apothéoz-Perret-Gentil, T. Stoeck, and J. Pawlowski. 2019. "Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring." *Trends in Microbiology* 27, no. 5: 387–397. <https://doi.org/10.1016/j.tim.2018.10.012>.
- Cowart, D. A., K. R. Murphy, and C.-H. C. Cheng. 2018. "Metagenomic Sequencing of Environmental DNA Reveals Marine Faunal Assemblages From the West Antarctic Peninsula." *Marine Genomics* 37: 148–160. <https://doi.org/10.1016/j.margen.2017.11.003>.
- Deiner, K., H. M. Bik, E. Mächler, et al. 2017. "Environmental DNA Metabarcoding: Transforming How We Survey Animal and Plant Communities." *Molecular Ecology* 26, no. 21: 5872–5895. <https://doi.org/10.1111/mec.14350>.
- Deiner, K., E. A. Fronhofer, E. Mächler, J.-C. Walser, and F. Altermatt. 2016. "Environmental DNA Reveals That Rivers Are Conveyor Belts of Biodiversity Information." *Nature Communications* 7, no. 1: 12544. <https://doi.org/10.1038/ncomms12544>.
- Ebenezer, T. E., A. W. T. Muigai, S. Nouala, et al. 2022. "Africa: Sequence 100,000 Species to Safeguard Biodiversity." *Nature* 603, no. 7901: 388–392. <https://doi.org/10.1038/d41586-022-00712-4>.
- Gante, H. F., I. Doadrio, M. J. Alves, and T. E. Dowling. 2015. "Semi-Permeable Species Boundaries in Iberian Barbels (*Barbus* and *Luciobarbus*, Cyprinidae)." *BMC Evolutionary Biology* 15: 1–18.
- Goldberg, C. S., C. R. Turner, K. Deiner, et al. 2016. "Critical Considerations for the Application of Environmental DNA Methods to Detect Aquatic Species." *Methods in Ecology and Evolution* 7, no. 11: 1299–1307.
- Gousskov, A., and C. Vorburger. 2016. "Postglacial Recolonizations, Watershed Crossings and Human Translocations Shape the Distribution of Chub Lineages Around the Swiss Alps." *BMC Evolutionary Biology* 16: 1–13.
- Hambäck, P. A., A. R. Cirtwill, D. García, et al. 2021. "More Intraguild Prey Than Pest Species in Arachnid Diets May Compromise Biological Control in Apple Orchards." *Basic and Applied Ecology* 57: 1–13.
- Jo, T., H. Murakami, R. Masuda, M. K. Sakata, S. Yamamoto, and T. Minamoto. 2017. "Rapid Degradation of Longer DNA Fragments Enables the Improved Estimation of Distribution and Biomass Using Environmental DNA." *Molecular Ecology Resources* 17, no. 6: e25–e33. <https://doi.org/10.1111/1755-0998.12685>.
- Jo, T., and H. Yamanaka. 2022. "Fine-Tuning the Performance of Abundance Estimation Based on Environmental DNA (eDNA) Focusing on eDNA Particle Size and Marker Length." *Ecology and Evolution* 12, no. 8: e9234.
- Kelly, R. P., A. O. Shelton, and R. Gallego. 2019. "Understanding PCR Processes to Draw Meaningful Conclusions From Environmental DNA Studies." *Scientific Reports* 9, no. 1: 12133. <https://doi.org/10.1038/s41598-019-48546-x>.
- Kelman, Z., and L. Moran. 1996. "Degradation of Ancient DNA." *Current Biology* 6, no. 3: 223. [https://doi.org/10.1016/S0960-9822\(02\)00460-8](https://doi.org/10.1016/S0960-9822(02)00460-8).
- Korf, I., M. Yandell, and J. Bedell. 2003. *Blast*. O'Reilly Media.
- Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform." *Applied and Environmental Microbiology* 79, no. 17: 5112–5120.
- Krehenwinkel, H., M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, and R. G. Gillespie. 2017. "Estimating and Mitigating Amplification Bias in Qualitative and Quantitative Arthropod Metabarcoding." *Scientific Reports* 7, no. 1: 17668. <https://doi.org/10.1038/s41598-017-17333-x>.

- Lewin, H. A., S. Richards, E. Lieberman Aiden, et al. 2022. "The Earth BioGenome Project 2020: Starting the Clock." *Proceedings of the National Academy of Sciences* 119, no. 4: e2115635118. <https://doi.org/10.1073/pnas.2115635118>.
- Lewin, H. A., G. E. Robinson, W. J. Kress, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences* 115, no. 17: 4325–4333. <https://doi.org/10.1073/pnas.1720115115>.
- Li, H. 2012. "Seqtk Toolkit for Processing Sequences in FASTA/Q Formats." *GitHub* 767: 69.
- Littlefair, J. E., J. J. Allerton, A. S. Brown, et al. 2023. "Air-Quality Networks Collect Environmental DNA With the Potential to Measure Biodiversity at Continental Scales." *Current Biology* 33, no. 11: R426–R428. <https://doi.org/10.1016/j.cub.2023.04.036>.
- Lu, J., N. Rincon, D. E. Wood, et al. 2022. "Metagenome Analysis Using the Kraken Software Suite." *Nature Protocols* 17, no. 12: 2815–2839.
- Lynggaard, C., M. F. Bertelsen, C. V. Jensen, et al. 2022. "Airborne Environmental DNA for Terrestrial Vertebrate Community Monitoring." *Current Biology* 32, no. 3: 701–707.e5. <https://doi.org/10.1016/j.cub.2021.12.014>.
- Macher, T.-H., R. Schütz, J. Arle, A. J. Beermann, J. Koschorreck, and F. Leese. 2021. "Beyond Fish eDNA Metabarcoding: Field Replicates Disproportionately Improve the Detection of Stream Associated Vertebrate Species." *bioRxiv*. 2021–03.
- Manu, S., and G. Umapathy. 2021. "A Novel Metagenomic Workflow for Biomonitoring Across the Tree of Life Using PCR-Free Ultra-Deep Sequencing of Extracellular eDNA." *Authorea Preprints*.
- Manu, S., and G. Umapathy. 2023. "Deep Sequencing of Extracellular eDNA Enables Total Biodiversity Assessment of Ecosystems." *Ecological Indicators* 156: 111171. <https://doi.org/10.1016/j.ecolind.2023.111171>.
- Miya, M., Y. Sato, T. Fukunaga, et al. 2015. "MiFish, a Set of Universal PCR Primers for Metabarcoding Environmental DNA From Fishes: Detection of More Than 230 Subtropical Marine Species." *Royal Society Open Science* 2, no. 7: 150088. <https://doi.org/10.1098/rsos.150088>.
- Morgulis, A., E. M. Gertz, A. A. Schäffer, and R. Agarwala. 2006. "A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences." *Journal of Computational Biology* 13, no. 5: 1028–1040.
- Nagler, M., S. M. Podmirseg, J. Ascher-Jenull, D. Sint, and M. Traugott. 2022. "Why eDNA Fractions Need Consideration in Biomonitoring." *Molecular Ecology Resources* 22, no. 7: 2458–2470.
- Pathan, S. I., P. Arfaio, E. Taskin, M. T. Ceccherini, E. Puglisi, and G. Pietramellara. 2021. "The Extracellular DNA Can Baffle the Assessment of Soil Bacterial Community, but the Effect Varies With Microscale Spatial Distribution." *FEMS Microbiology Letters* 368, no. 12: fnab074.
- Pereira, H. M., P. W. Leadley, V. Proença, et al. 2010. "Scenarios for Global Biodiversity in the 21st Century." *Science* 330, no. 6010: 1496–1501. <https://doi.org/10.1126/science.1196624>.
- Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. "Widespread Discordance of Gene Trees With Species Tree in *Drosophila*: Evidence for Incomplete Lineage Sorting." *PLoS Genetics* 2, no. 10: e173.
- Quince, C., A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. 2017. "Shotgun Metagenomics, From Sampling to Analysis." *Nature Biotechnology* 35, no. 9: 833–844. <https://doi.org/10.1038/nbt.3935>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Published online 2020.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- Razjigaeva, N., L. Ganzey, T. Grebennikova, et al. 2022. "Bioaerosols as Evidence of Atmospheric Circulation Anomalies Over the Okhotsk Sea and Shantar Islands in the Late Glacial–Holocene." *Climate* 10, no. 2: 24. <https://doi.org/10.3390/cli10020024>.
- Ribeiro, F., H. F. Gante, G. Sousa, A. F. Filipe, M. J. Alves, and M. F. Magalhaes. 2009. "New Records, Distribution and Dispersal Pathways of *Sander lucioperca* in Iberian Freshwaters." *Cybium* 33, no. 3: 255–256.
- Ribeiro, F., and A. Verissimo. 2014. "Full Westward Expansion of *Rutilus rutilus* (Linnaeus, 1758) in the Iberian Peninsula." *Journal of Applied Ichthyology* 30, no. 3: 540–542.
- Schmeller, D. S., M. Böhm, C. Arvanitidis, et al. 2017. "Building Capacity in Biodiversity Monitoring at the Global Scale." *Biodiversity and Conservation* 26, no. 12: 2765–2790. <https://doi.org/10.1007/s10531-017-1388-7>.
- Segherloo, I. H., J. Freyhof, P. Berrebi, et al. 2021. "A Genomic Perspective on an Old Question: Salmo Trout or *Salmo trutta* (Teleostei: Salmonidae)?" *Molecular Phylogenetics and Evolution* 162: 107204.
- Sellers, G. S., C. Di Muri, A. Gómez, and B. Hänfling. 2018. "Mu-DNA: A Modular Universal DNA Extraction Method Adaptable for a Wide Range of Sample Types." *Metabarcoding and Metagenomics* 2: e24556. <https://doi.org/10.3897/mbmg.2.24556>.
- Simon, H. Y., K. J. Siddle, D. J. Park, and P. C. Sabeti. 2019. "Benchmarking Metagenomics Tools for Taxonomic Classification." *Cell* 178, no. 4: 779–794.
- Singer, G. A. C., S. Shekarri, A. McCarthy, N. Fahner, and M. Hajibabaei. 2020. "The Utility of a Metagenomics Approach for Marine Biomonitoring." [Preprint]. *Ecology*. <https://doi.org/10.1101/2020.03.16.993667>.
- Stat, M., M. J. Huggett, R. Bernasconi, et al. 2017. "Ecosystem Biomonitoring With eDNA: Metabarcoding Across the Tree of Life in a Tropical Marine Environment." *Scientific Reports* 7, no. 1: 12240. <https://doi.org/10.1038/s41598-017-12501-5>.
- Taberlet, P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg. 2012. "Environmental DNA." *Molecular Ecology* 21, no. 8: 1789–1793. <https://doi.org/10.1111/j.1365-294X.2012.05542.x>.
- Venter, J. C., K. Remington, J. F. Heidelberg, et al. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea." *Science* 304, no. 5667: 66–74.
- Wang, J., P. Liu, J. Chang, C. Li, F. Xie, and J. Jiang. 2022. "Development of an eDNA Metabarcoding Tool for Surveying the World's Largest Amphibian." *Current Zoology* 68, no. 5: 608–614.
- Wickham, H., W. Chang, and M. H. Wickham. 2016. "Package 'ggplot2.' Create Elegant Data Visualisations Using the Grammar of Graphics." *Version*, 2(1), 1–189.
- Wood, D. E., J. Lu, and B. Langmead. 2019. "Improved Metagenomic Analysis With Kraken 2." *Genome Biology* 20: 1–13.
- Zamani, A., Z. F. Fric, H. F. Gante, et al. 2022. "DNA Barcodes on Their Own Are Not Enough to Describe a Species." *Systematic Entomology* 47, no. 3: 385–389. <https://doi.org/10.1111/syen.12538>.
- Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. "PEAR: A Fast and Accurate Illumina Paired-End reAd mergeR." *Bioinformatics* 30, no. 5: 614–620.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.