Taxonomic diversity assessment across the tree of life from extracellular environmental DNA in aquatic ecosystems

Shivakumara Manu[1] and Govindhaswamy Umapathy[1]

[1]Centre for Cellular and Molecular Biology CSIR

February 21, 2023

Abstract

Global biodiversity is declining at an alarming pace due to anthropogenic factors. It is imperative to monitor the health of ecosystems through biodiversity assessments. But existing biodiversity assessment approaches are not scalable to assess the complete diversity of all the life forms in large ecosystems. We hypothesized that the extracellular environmental DNA (eDNA) released by natural cell lysis of biological matter in aquatic ecosystems is a repertoire of genetic material from all the inhabiting organisms and deep sequencing of extracellular eDNA could enable the detection of taxa across the tree of life. We developed a lysis-free and PCR-free workflow to directly enrich and sequence extracellular eDNA from environmental water samples and demonstrate its utility for taxonomic diversity assessment across the tree of life in a large biodiverse model aquatic ecosystem (Ramsar wetland: Chilika lagoon). Using incidence-based asymptotic richness analysis, we estimated that Chilika harbors about 1071 families across the tree of life comprising approximately 799 families of Eukaryotes, 230 families of Bacteria, 27 families of Archaea, and 13 families of DNA Viruses. We also quantified the compositional changes in the relative abundance of families across the tree of life and detected a higher temporal variation (46%) across the seasons than the spatial variation (37%) across the lagoon. With the increasing genomic resources and decreasing sequencing costs, we envision the wide adoption of extracellular eDNA-based taxonomic diversity assessment across the tree of life to track the future biodiversity loss and aid the conservation, restoration, and management efforts in the Anthropocene.

INTRODUCTION

The vast biodiversity on earth is the result of billions of years of evolution. All the evolutionary lineages that make up the tree of life belong to three domains: Archaea, Bacteria, Eukaryota, and a fourth contested category of Viruses. Organisms across the tree of life have evolved and adapted to inhabit various environments on earth. Widely accepted studies estimate that about 8.7 million (±1.3 million) eukaryotic (Mora

et al., 2011) and up to a trillion species of microbes (Locey & Lennon, 2016) exist on earth. Despite over 250 years of scientific exploration, the majority of eukaryotic diversity and most of the microbial diversity remain unknown to science (The Catalogue of Life, 2022). However, there is an impending threat of sixth mass extinction due to anthropogenic factors such as pollution, land-use change, habitat loss, poaching, and climate change (Ceballos et al., 2020). The population sizes of many species have dropped significantly, and species extinction rates have increased hundreds of times compared to the background rate (Ceballos et al., 2015, 2017). Currently, about 28% of the 150,388 species of animals and plants assessed by the IUCN are threatened with extinction (IUCN Red List of Threatened Species, 2022). Using the IUCN Red List data, a global assessment report by the U.N. estimated that up to a million eukaryotic species may be threatened with extinction and might go extinct in the next few decades (Watson et al., 2019). The extinction of species is irreversible and may have long-lasting effects on ecosystem functions and services. Under the current scenario, many species may go extinct even before being cataloged. Therefore, it has become imperative to assess biodiversity at larger scales than ever before to chart conservation policies, guide restoration projects, and devise management plans.

Classical biodiversity assessment techniques are time-consuming, resource-intensive, require manual identification of specimens, and are not easily scalable to deploy in large ecosystems. The advent of DNA-based identification of species using large reference databases of standardized DNA barcodes (e.g., COI, rbcL-matK, ITS) has led to the development of high-throughput methods to assess the composition of communities from bulk and environmental samples (Creer et al., 2016). Environmental DNA (eDNA)-based biodiversity assessment techniques detect the presence of species in the ecosystem using DNA extracted directly from whole environmental samples (e.g., water, soil, air) without isolating the target organisms (Taberlet, Coissac, Hajibabaei, et al., 2012). eDNA-based bioassessment offers several advantages over classical methods (Thomsen & Willerslev, 2015). For example, filtering water samples to detect fish communities requires less time and resources than surveying fish using gill nets or electrofishing. Such physical scalability in sampling enables the collection of samples covering entire ecosystems with minimal effort. In addition, the same eDNA sample can be repurposed to detect a different set of organisms, eliminating the need for repeated sampling (Dysthe et al., 2018). By exploiting various sources of DNA in an environmental sample, eDNA-based bioassessment has emerged as a powerful new technique that has revolutionized the way we survey ecological communities (Deiner et al., 2017). The last decade witnessed tremendous strides in the methodological development of eDNA-based bioassessment techniques (Seymour, 2019). Along with the technical advances, there has also been considerable effort to understand the ecology of eDNA (Barnes & Turner, 2016; Stewart, 2019) and to clearly define the term eDNA (Pawlowski et al., 2020; Rodriguez-Ezpeleta et al., 2021). Numerous comparative studies have concluded that eDNA-based bioassessment could complement or even potentially replace classical bioassessment methods in the future (Brantschen et al., 2021; Leempoel et al., 2020).

However, the current methodologies employed in eDNA-based bioassessment are limited to approaches where a specific species or a group of related taxa are targeted. A quantitative PCR assay is routinely used to detect a species of interest with high sensitivity using species-specific primers and a PCR-based metabarcoding is employed to detect a group of related taxa using a universal primer targeting a barcode region (Bruce et al., 2021). Recently, hybridization capture by oligonucleotide probes was also employed for the targeted detection of a single species (Jensen et al., 2021) and multiple species of interest (Seeber et al., 2019). The inherent limitations of such targeted approaches are that they only permit the assessment of organisms sharing a common barcoding marker and suffer from enrichment bias leading to considerable and unpredictable dropout of species when targeting a large number of taxa (van der Loos & Nijland, 2021, 2021). But ecosystems are a continually interacting web of life and any anthropogenic effect on one group of taxa affects many other taxa in the ecological network and influences the overall ecosystem stability (Bascompte, 2009). Hence, we must transition from a targeted approach of biodiversity assessment of a specific group of taxa to a more holistic approach that encompasses all the life forms inhabiting an ecosystem. Therefore, an untargeted approach that can detect organisms irrespective of taxonomic affiliation is fundamental for scaling up the biomonitoring efforts and monitoring the fast-changing environments in the Anthropocene.

In this study, we explored whether we can effectively assess the taxonomic diversity across the tree of life

in an ecosystem using PCR-free approaches. Technically, shotgun sequencing of environmental DNA on extremely high-throughput sequencing platforms could yield billions of DNA sequences that can be used to assess biodiversity in an untargeted manner and overcome the biases and limitations of targeted approaches (Taberlet, Coissac, Pompanon, et al., 2012). Due to the absence of any target enrichment steps such as PCR with universal primers or hybridization capture with DNA/RNA probes, metagenomic approaches can provide an unbiased representation of the input library of eDNA. However, the eDNA extracted from the samples should be a good representation of the total biodiversity in the sampled location. In this regard, the ecology of eDNA encompassing the origin, state, fate, and transport of eDNA in an ecosystem has to be given due consideration (Barnes & Turner, 2016). The various sources of environmental DNA in an ecosystem offer different snapshots of biodiversity at a wide range of spatiotemporal resolutions (Fig. 1A) (Bohmann et al., 2014). The DNA released into the environment by the natural cell lysis of the organismal and extra-organismal biological entities constitute the extracellular eDNA (Nagler et al., 2022). Once released into the environment, extracellular eDNA adsorbs onto surface-reactive soil particles through cation bridging and becomes resistant to degradation (Nagler et al., 2018). In aquatic ecosystems, the particle-bound extracellular eDNA can remain suspended and spatially dispersed in the water column until sedimentation. Therefore, we hypothesized that the extracellular eDNA is a natural repertoire of the genetic material of organisms inhabiting an ecosystem and is suitable for taxonomic diversity assessment across the tree of life over large spatiotemporal scales.

To test our hypothesis, we devised a modular workflow (Fig. 1B) from sampling to data analysis and asked the following key questions in this study: (i) Can we simultaneously detect organisms across the tree of life in a single assay using extracellular eDNA? (ii) Can we estimate the total taxonomic richness of an ecosystem across the tree of life? (iii) Does extracellular eDNA provide enough spatiotemporal resolution to detect changes in biodiversity across the tree of life? Using the Chilika lagoon as a model ecosystem, we show that taxa across all the domains of life can be detected through PCR-free deep sequencing of extracellular eDNA enriched from large-volume water samples. Further, using seasonal samples, we estimate the asymptotic taxonomic richness across the tree of life and resolve the changes in biodiversity at broad spatiotemporal scales. We conclude that PCR-free deep sequencing of extracellular eDNA is an effective tool to assess the taxonomic diversity across the tree of life in large ecosystems.

METHODS

Model Ecosystem

To test our approach to assess taxonomic diversity across the tree of life, we selected Chilika lagoon, a highly biodiverse tropical brackish water ecosystem located on the east coast of India as our model ecosystem (Supplementary Fig. 1). Chilika lagoon was designated as India's first Ramsar site (no. 229) of international importance in 1981. It is the second largest brackish-water lagoon in the world, extending about 64 km in length and 20 km in width and spanning about 1100 sq. km of the area during the monsoon season. As the lagoon receives both freshwater and marine water, there exists a dynamic gradient of salinity ranging from 0 - 5 ppt in the northern sector to 5-18 ppt in the central and southern sectors. As a result, Chilika harbors a unique and diverse community assemblage consisting of marine organisms from the Bay of Bengal - the northeastern part of the Indian Ocean, and freshwater species from the tributaries of the Mahanadi River, a major river system in east-central India (Supplementary Fig. 1). It has earned an economically important status by supporting the food and livelihood of over 200,000 fisherfolk. Moreover, the fish diversity of Chilika has been well documented and a comprehensive checklist of fishes sighted in the last 100 years was recently published (Suresh et al., 2018). This checklist serves as an excellent reference to compare the results of our study. Chilika is a shallow water lagoon with an average depth of about 2 meters and experiences strong coastal winds. Therefore, we presumed that sampling the surface water should suffice for the objectives of this study as there is no strong vertical stratification in the water column.

Sample collection

First, we designed a pilot-scale sampling strategy to test the feasibility and reproducibility of our approach. We selected three equally spaced geolocated stations (S27, S28, S29) on a 10 km transect in the central sector of the lagoon (Supplementary Fig. 2). Next, we designed a spatiotemporally replicated sampling strategy with 7 geolocated stations (S1, S6, S14, S17, S26, S29, S30) that are spread across the lagoon to assess the taxonomic diversity (Supplementary Fig. 3). The pilot-scale sample collection at the 3 locations was conducted in December 2019 (Winter), while the spatiotemporal sampling was conducted during the months of March (Summer), July (Monsoon), and November (Winter) of 2020. All 7 stations were sampled in the Monsoon and 3 randomly selected stations were sampled in Summer and Winter. As we intended to use PCR-free approaches downstream, we opted to filter large volumes of water to obtain a sufficient yield of extracellular DNA. We filtered about 10 liters of water in each geolocated sampling station using the integrated eDNA sampler by Smithroot Inc. (Thomas et al., 2018). We also simultaneously measured water temperature, salinity, and pH using a water quality sonde at every sampling station (YSI, Model No. 6600, V2). We used a 0.45um mixed cellulose ester (MCE) filter membrane for water filtration as it has been shown to bind to the free form of extracellular DNA due to its chemical affinity (Liang & Keeley, 2013). The selected pore size allows more volume of water to be filtered before clogging and also retains suspended soil particles such as clay (<2um), silt (2-50um), and sand (50um-2mm) which are adsorbed by extracellular DNA (Nagler et al., 2018). A triplicate sampling module was used to maximize the rate and volume of water sampled per location. We set a maximum vacuum pressure of 10 psi to minimize cell lysis during filtration and maintain a flow rate of less than one liter per minute. The eDNA sampling system also avoids the risk of sample contamination by utilizing sterile single-use self-preserving filter holders, which were replaced for every sampling location. The self-preserving filter holder comprises desiccating plastic material that completely removes any traces of water and preserves eDNA for several weeks at room temperature (Thomas et al., 2019). Hence, we transported the filter holders to the laboratory in dark conditions at room temperature.

Extracellular eDNA enrichment

We adapted a lysis-free phosphate buffer-based DNA extraction protocol to enrich the extracellular eDNA from the filter membranes and minimize the proportion of organismal and extra-organismal DNA (Lever et al., 2015; Liang & Keeley, 2013; Taberlet, Prud'homme, et al., 2012). The main principle of the extraction protocol is to desorb the extracellular DNA bound to the surface of the MCE filter membrane and soil particles without lysing the intact cellular and subcellular particles using the saturated phosphate buffer (Nagler et al., 2022). The phosphate groups from the buffer compete with the phosphate groups of the extracellular DNA bound to the surface of soil particles via cation bridging and desorb the DNA by chemical displacement (Taberlet, Prud'homme, et al., 2012). The desorbed DNA is then isolated through a column-based DNA isolation protocol using reagents and columns from the Nucleospin soil kit (Macherey-Nagel, Germany). We performed extracellular DNA extraction from the filter membranes in a clean lab within a week of sample collection. The bench surfaces were wiped with 50% diluted commercial bleach, followed by distilled water and 70% ethanol. We used filter tips to pipette all the liquids to avoid any aerosol contamination during the extraction. The phosphate buffer was freshly prepared before extraction by mixing 0.197g of NaH2PO4 and 1.47g of Na2HPO4 in 100 ml of DNA-free water (0.12M, pH 8). Filter membranes were carefully taken out of the filter holders and rolled using sterile forceps before placing them into 15 ml falcon tubes containing 5 ml phosphate buffer and large ceramic beads (0.6-0.8mm) from the Nucleospin soil kit. The falcon tubes were shaken for 10-15 minutes by placing them on a vortex mixer with the vertical falcon holder module. The large ceramic beads help homogenize soil clumps recalcitrant to the desorption process but do not disrupt the intact cells. This process is principally different from bead-beating employed in microbiology to lyse the cells. We did not exceed the time of mixing beyond 15 minutes to avoid co-extracting a large proportion of humic acids. The homogenized mixture was immediately centrifuged at 11000 x g to precipitate particulate matter, and the supernatant was passed through the Nucleospin inhibitor removal column. The DNA-binding condition of the flow-through was then adjusted using the binding buffer and then passed through the silica

column from the Nucleospin soil kit. DNA was eluted using 150ul of warm Tris EDTA buffer with three successive elutions. We quantified the extracellular eDNA elute using a high-sensitivity double-stranded DNA assay in Qubit 4 (Thermo Fisher Scientific, USA). Samples with less than 20 ng/ul concentration were concentrated until the volume decreased to about 50ul using the SpeedVac vacuum concentrator (Thermo Fisher Scientific, USA).

Library preparation

The DNA concentration of the samples was diluted to 20ng/ul, and one microgram of DNA was taken as input for the library preparation. We chose the Illumina Truseq DNA PCR-free library preparation method to avoid PCR-induced artifacts such as substitutions, indels, and chimeras. It also helps to drastically reduce the uninformative duplicate reads that arise from library amplification. We did not include technical replicates for library preparation since there is no stochasticity from amplification due to a completely PCR-free workflow. The input DNA was first randomly sheared into 350 bp fragments using the Covaris ultrasonicator. The ends of the fragmented DNA were repaired and dA-tailed prior to ligation using unique dual index (UDI) adapters for Illumina (IDT, USA). The adapter-ligated library fragments were size-selected and purified with SPRI beads (Beckman Coulter, USA). The fragment sizes of the libraries were verified using the Agilent Bioanalyzer high-sensitivity DNA chip. It is to be noted that there is no amplification of libraries post-ligation. Therefore, the concentration of libraries was determined using a library quantification kit consisting of known concentrations of standards (Takara Bio, USA) on the ViiA7 real-time quantitative PCR (Applied Biosystems, USA). The failed libraries with less than 1nM concentration possibly due to residual inhibitors in the input DNA were excluded from sequencing.

Sequencing and quality control

We first performed a sequencing saturation analysis to determine the library complexity and target sequencing depth using a sample of 10 million reads from each of the 3 pilot-scale study samples. We used only the forward (R1) pair of the reads to count the kmers. We generated a kmer frequency histogram with the KMERCOUNTEXACT command in the BBTOOLS package (Bushnell, 2022) using 31 bp kmers having at least a 95% probability of correctness based on the base quality scores. Using the LC_EXTRAP command in PRESEQ v.3.2, we extrapolated the library complexity using a rational function approximation via continued fractions (Daley & Smith, 2013). We estimated the saturation of unique kmers with 100 bootstraps, a step size of 100 million kmers, and a maximum extrapolation of up to 100 billion kmers. We determined the saturation point where the kmer uniqueness reached a minimum threshold of 5%. We calculated the target sequencing depth as the saturation point divided by the number of 31 bp kmers in a 150 bp read. The concentrations of the extracellular eDNA libraries passing QC were then adjusted to achieve the target sequencing depth and pooled together. The pooled libraries were denatured into single-stranded DNA before loading onto an S4 patterned flow cell and sequenced for 300 cycles in paired-end mode on the Illumina Novaseq 6000 high-throughput short-read platform. The binary base call files in BCL format were demultiplexed and converted into FASTQ format using the Illumina BCL2FASTQ v2.20 software. The sequences with unexpected index combinations due to cross-contamination between multiplexed samples (via tag jumps) remain unclassified during demultiplexing due to the unique dual indexing (UDI) strategy used in the library preparation. The optical duplicates and complementary strand duplicates from the demultiplexed raw reads were filtered out using the CLUMPIFY tool in the BBTOOLS package. The deduplicated sequences containing adapter sequences and low-quality ends (q<10) were trimmed, and sequences shorter than 51 bp or containing more than 3 uncalled bases (N) after trimming were filtered out using the BBDUK tool in the BBTOOLS package (Bushnell, 2022).

Taxonomic assignment

We assembled a reference set of protein sequences from all the domains of life using the Uniprot reference clusters database v2022_3. The UniRef100 consists of the representative sequences from the entire protein universe with all the redundant sequences and fragments filtered out at 100% identity. We further filtered out unclassified and artificial sequences and retained only those UniRef100 sequences mapped with a valid NCBI taxonomic ID under the domains Archaea, Bacteria, Eukaryota, and Viruses. The final set of references was used to build an FM-index of the Burrows-Wheeler transformed sequences compatible with the metagenomic classifier KAIJU (Menzel et al., 2016). We classified the paired-end reads (R1 & R2) separately by querying the six-frame translated reads against the UniRef100 database using the Maximal Exact Matches algorithm implemented in KAIJU. We retained the default parameters of minimum match length (11 aa) and low complexity filtering using the SEG algorithm. We then merged the classification of the respective R1 and R2 reads with the Lowest Common Ancestor algorithm to increase the specificity of the classification of each read pair. We considered the read pair as unclassified if one of the reads was not classified under any domains of life or if both the read classifications were based on the same protein fragment. We calculated the total abundance of reads assigned for each classified family in all the samples. We also calculated a background rate of classification for each family as the total number of reads assigned to its parent order divided by the number of families classified under the parent order. We filtered out the families with less than 1000 classified reads or the background rate of classification. We also filtered out any common contaminants showing unusually high abundance and RNA virus families that cannot be directly detected through DNA sequencing. Finally, to assess the effectiveness of the taxonomic assignment, we compared the list of families from the class Actinopteri with the checklist of known fishes of Chilika (Suresh et al., 2018) and the list of proteomes available in Uniprot.

Diversity analysis

We used the incidence-based statistical framework (Colwell et al., 2012) to estimate the asymptotic taxonomic richness using the iNEXT R package (Hsieh et al., 2016). We divided the combined metagenomic data into sampling units of 100 million reads each and calculated the incidence frequencies of each taxon in the sampling units. We then estimated the asymptotic richness (hill number of 0th order) of various taxa using statistical extrapolation of accumulation curves of the observed incidence frequencies in the sampling units with 100 bootstraps. The Jaccard similarity index was calculated with the Spader R package (Chao & Jost, 2015) using the incidence frequencies of taxa in pairwise samples with a sampling unit size of 10 million reads. Next, We used the R package Phyloseq (McMurdie & Holmes, 2013) to calculate the Bray-Curtis dissimilarity and ordination of the samples. We generated a count matrix of the read counts of all the families found in each sample and converted it into a Phyloseq object with the taxonomy and sample metadata. The raw read counts were converted into relative abundances by dividing them by the total classified read count in the respective sample. The Bray-Curtis dissimilarity between the samples was calculated using the distance function of Phyloseq. The dissimilarity matrix was used as input for the non-metric multidimensional scaling (NMDS) method of ordination. The NMDS was run for 20 iterations until the solution was reached with a stress value of less than 0.2.

RESULTS

Organisms across the tree of life can be detected by PCR-free deep sequencing of extracellular eDNA

We first investigated whether we can effectively detect taxa across the tree of life using extracellular eDNA. We designed a lysis-free and PCR-free workflow to efficiently enrich, sequence, and taxonomically classify extracellular eDNA from water samples (Fig. 1). We tested the feasibility of our workflow through a pilot study and then conducted a spatiotemporal study in a biodiverse model aquatic ecosystem (Chilika lagoon,

India) (Supplementary Figs. 1, 2, and 3). Using a random sample of 10 million reads (150 bp) from each of the three pilot study samples, we estimated the minimum target sequencing depth based on the library complexity. We observed 95% saturation of unique kmers at about 50 billion observed kmers (Supplementary Fig. 4) corresponding to a sequencing depth of 416.6 million reads. Using this as the minimum target sequencing depth, we generated a total of 3.3 trillion bases of data from 16 samples. We demultiplexed the base call files using the unique dual indexes and obtained 10.96 billion paired-end reads (150 bp x 2) with a median depth of 658.35 million reads (SD 185.98 million) (Supplementary Table 1). After removing the optical duplicates, complementary strand duplicates, and quality filtering, we retained 94.37% of the paired-end reads. We taxonomically classified the high-quality deduplicated paired-end reads by querying them against the UniRef100 database containing 162.78 million reference protein sequences from 8539 families of taxa across the tree of life (Supplementary Fig. 5). We classified a total of 6.59 billion reads under all the domains of life with a median classification rate of 64.29% (SD 8.4%) per sample (Supplementary Fig. 6).

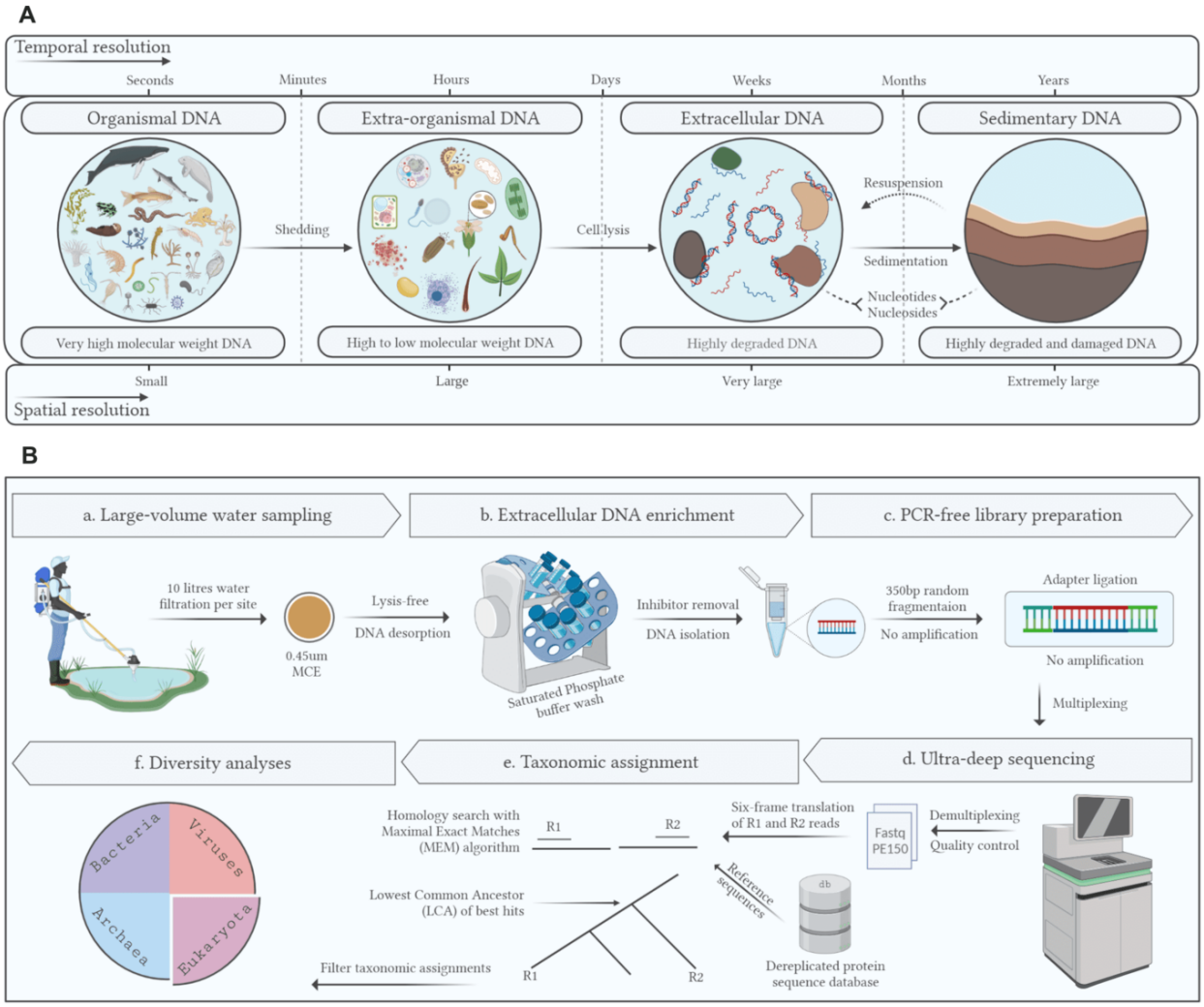


Figure 1: **Taxonomic diversity assessment across the tree of life from extracellular eDNA** A) The complex ecology of eDNA in a typical aquatic ecosystem. High molecular weight organismal and extra-organismal DNA from whole organisms and their dissociated parts is released into the environment through natural cell lysis. The extracellular DNA exists in a free form until it is completely degraded into nucleotides or otherwise adsorbs onto surface-reactive suspended soil particles that may eventually sediment down. The different types of DNA in an environmental sample provide varying spatiotemporal resolutions of biodiversity, from a few seconds to thousands of years of temporal and a few meters to several kilometers of spatial resolution depending on the origin, transport, and fate of eDNA. B) Illustration of the modular workflow from sampling to analysis. About 10 liters of water is filtered through a 0.45 um mixed cellulose ester membrane. The extracellular eDNA is desorbed from the filter membrane and enriched through a lysis-free saturated phosphate buffer method. Co-extracted inhibitors are removed and DNA is eluted using silica columns. The extracellular eDNA is randomly fragmented and ligated with adapters using PCR-free library preparation methods. Multiplexed libraries are sequenced in the paired-end mode in a high-throughput sequencer. The reads are queried against a protein database and taxonomically classified using maximal exact matches and the lowest common ancestor algorithm. The taxonomic assignments are used for various diversity analyses.

Overall, the majority of the reads were classified under the domain Bacteria (86.95%) followed by Eukaryota (5.48%), Archaea (1.01%), and Viruses (6.54%). The taxonomic resolution of the classified reads under the domains progressively decreased from 80.61% at the phylum level to 21.34% at the species level (Supplementary Fig. 7). We then inspected the taxonomic assignments of reads classified at least up to the family level. After filtering out the low-abundant families, we obtained 2.14 billion reads from all the samples classified under 1001 families across the tree of life (Fig. 2, Supplementary Data). The highest proportion of families belonged to Eukaryota (73%) followed by Bacteria (23%), Archaea (1%), and DNA Viruses (3%). Eukaryotic families were distributed in all the kingdoms, namely, Metazoa (54%), Fungi (20%), Viridiplantae (13%), and Protists (13%). Proteobacteria was the richest phylum under Bacteria with 43% of the families, followed by Actinobacteria (13%), and Firmicutes (9%). Archaea majorly consisted of families under the Phylum Euryarchaeota (63%) and Crenarchaeota (19%). The DNA Phages from the kingdom Heunggongvirae made up 62% of the viral diversity. As the non-microbial organisms are relatively low-abundant in an ecosystem than the microbes, it is technically more difficult to detect animals and plants from an environmental sample. Since we detected a high diversity of Metazoans, we assessed the sensitivity of the taxonomic assignment using a recently updated checklist of the fishes sighted in the last century in the Chilika lagoon. Upon comparison, about 44.26% of the 61 bony fish families detected in our samples matched with the checklist (Supplementary Data). We suspected the low concordance between our results and the checklist could be due to the low representation of fishes from our ecosystem in the reference database. Hence, we inspected the availability of proteomes from the annotated reference genomes in the database and their influence on the taxonomic assignment. We found that 71.74% of the 92 fish families in the checklist did not have a representative proteome in the reference database. On the contrary, we noted that about 86.88% of all the 61 fish families detected in this study were well represented in the database with complete proteomes (Supplementary Data). Hence, we re-calculated the sensitivity by only considering the 26 fish families from the checklist that were represented in the database. By accounting for the incompleteness of the reference database, the sensitivity of the taxonomic assignment drastically increased to 88.46%, almost double the previous estimate.

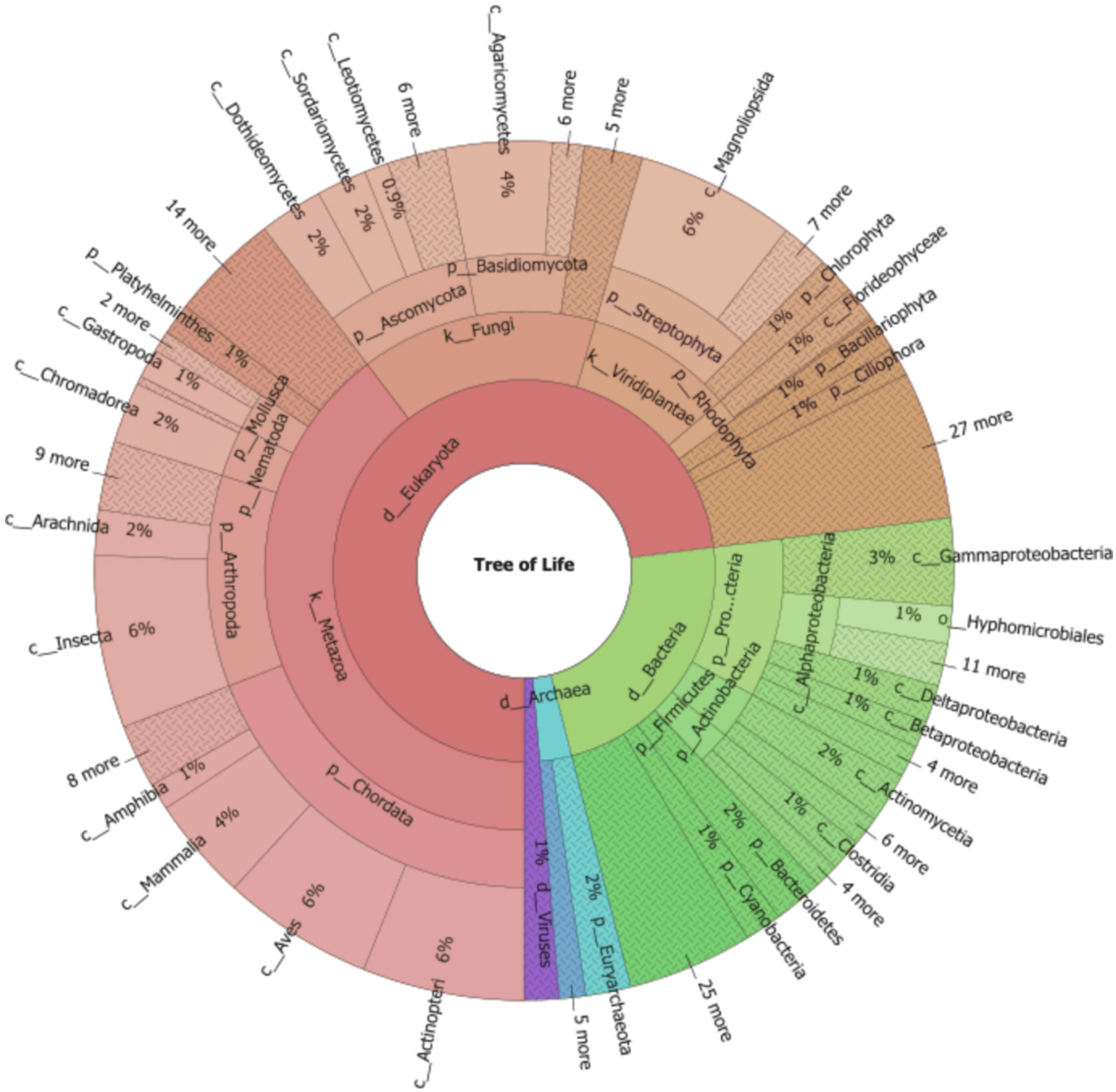


Figure 2: **FTaxonomic diversity across the tree of life** Distribution of the observed family richness (n=1001) across the tree of life detected in all the extracellular eDNA samples from Chilika lagoon. The taxon names are prefixed with a single alphabet code denoting the taxonomic ranks: d – domain, k – kingdom, p – phylum, and c – class.

Extrapolation of taxa accumulation curves provides estimates of total taxonomic richness across the tree of life

Since the observed richness of taxa in the samples is often limited by the sequencing depth, we inferred the total taxonomic richness of the ecosystem accounting for the low abundant taxa that could be potentially detected by increasing the sequencing depth. Through statistical extrapolation of the richness accumulation

curve derived from incidence frequencies of taxa, we estimated the asymptotic family richness of the Chilika lagoon across the tree of life to be 1071.49 (SEM 20.82) (Fig. 3A). Comparing the observed family richness to the estimate of asymptotic family richness, we had detected almost 93.42% (SEM 0.01%) of the taxa across the tree of life in our dataset. Most of the undetected diversity was contributed by Eukaryotes (96.79%) as the family richness accumulation curves of Archaea, Bacteria, and Viruses were nearly saturated (Supplementary Fig. 8). Further, we estimated the asymptotic taxonomic richness of all the domains and different Eukaryotic kingdoms, phyla, and classes (Fig. 3B). The tree of life consisted of 799 families (SEM 20.02) of Eukaryota, 230 families (SEM 0.0) of Bacteria, 27 families (SEM 0.14) of Archaea, and 13 families (SEM 0.69) of DNA Viruses. Metazoa was the richest kingdom in Eukaryota with about 452 families (SEM 21.67), followed by the kingdoms Fungi and Viridiplantae with about 148 families (SEM 0.49), and 111 (SEM 14.49) families, respectively. Phylum Chordata with 196 families (SEM 7.37) and Arthropoda with 114 families (SEM 8.46) made up most of Metazoa. Actinopteri was the richest class in Chordata with 65 families (SEM 5.21) followed by Aves, Mammalia, and Amphibia with 57 families (SEM 0.0), 39 families (SEM 0.46), and 10 families (SEM 1.73), respectively. We noted that the estimated asymptotic richness of 65 bony fish families (SEM 5.21) is well within the theoretical maximum of 92 fish families sighted in Chilika in the last century (Suresh et al., 2018).

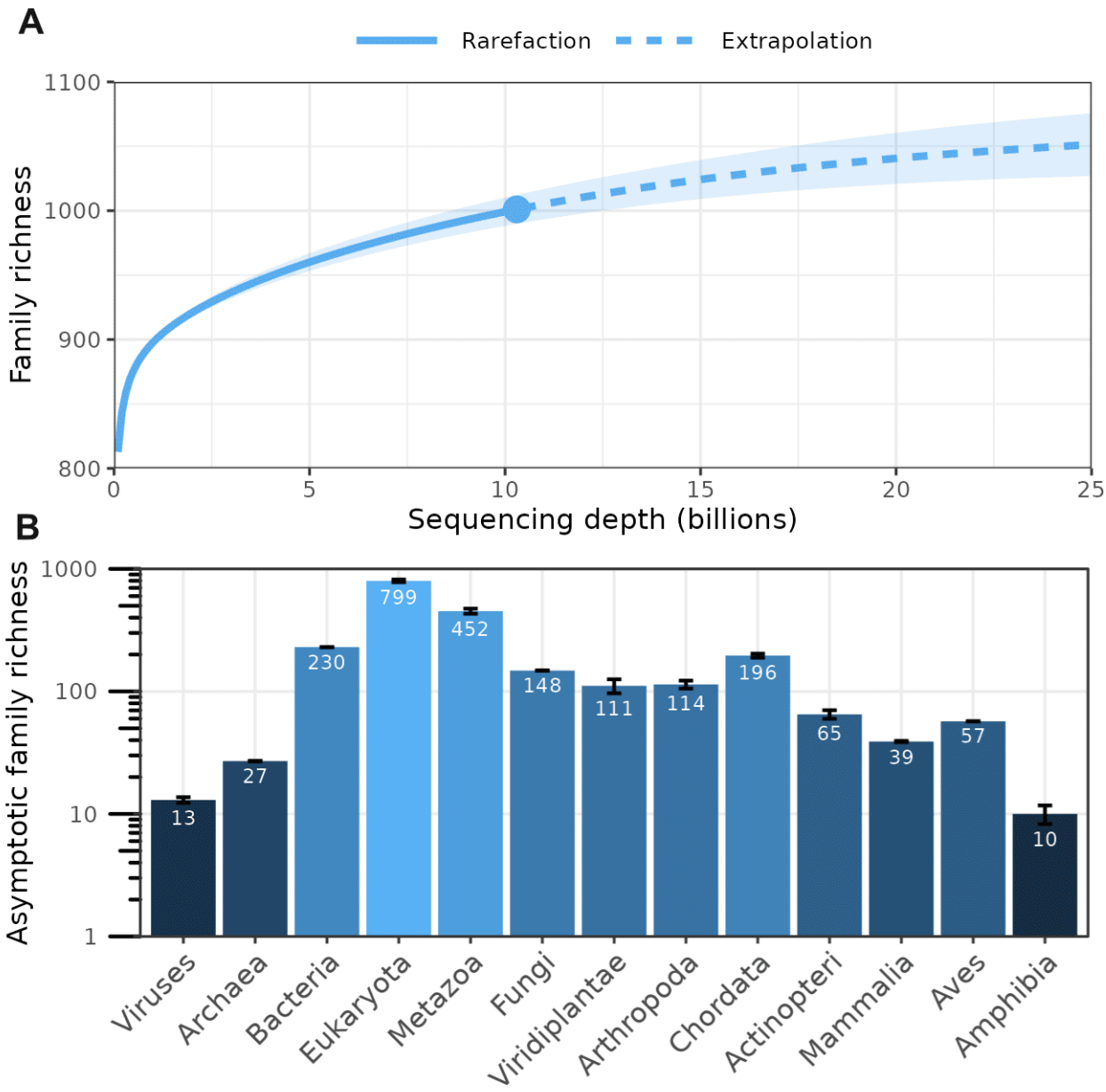


Figure 3: **The asymptotic taxonomic richness across the tree of life** A) Family richness accumulation curve as a function of sequencing depth in billions of unique paired-end reads. The solid and dotted lines represent the rarefaction curve of observed richness and the statistical extrapolation with 95% confidence intervals, respectively. B) Estimates of asymptotic richness estimates for various domains, kingdoms, phyla, and classes across the tree of life. The estimates are rounded off to the nearest integer and the error bars indicate the standard error of the estimates.

Extracellular eDNA resolves broad-scale spatiotemporal changes in biodiversity

As particle-bound extracellular eDNA can persist in the environment for a considerable amount of time depending on the abiotic conditions of the ecosystem, we examined whether extracellular eDNA can resolve broad-scale spatiotemporal changes in biodiversity across the tree of life. The sampling locations across the seasons and different sectors of the lagoon had a wide variation in abiotic conditions (Supplementary Fig. 9). The temperature of water ranged from 23 to 33.3 degrees Celsius and the salinity varied from 0.83 to 15.18 ppt. To check the variation in biodiversity at these locations, we generated a read count matrix of the 1001 families detected across the tree of life in all the samples (Supplementary Data). We then measured the spatial and temporal beta diversity with the richness-based Jaccard similarity index and the relative abundance-based Bray-Curtis dissimilarity. The Jaccard index indicated a very high degree of shared families across the tree of life among the spatiotemporal samples with a median similarity of 0.98 (SD 0.01). In contrast, the Bray-Curtis dissimilarity indicated a high variation of relative abundance of families across the tree of life among the spatial and temporal samples with median values of 0.37 (SD 0.09) and 0.46 (SD 0.10), respectively. Further, we inspected if the degree of spatiotemporal variation in relative abundances differed among the taxonomic domains. We found the highest average Bray-Curtis dissimilarity among Bacteria (0.39, SD 0.1) and Viruses (0.28, SD 0.14), compared to Archaea (0.14, SD 0.07) and Eukaryotes (0.13, SD 0.05) (Supplementary Fig. 10). Ordination of relative abundance-based beta diversity across the tree of life by NMDS resulted in the clustering of samples primarily by season and to a lesser extent by location, reflecting the changes in the biodiversity of the lagoon across space and time (Fig. 4).

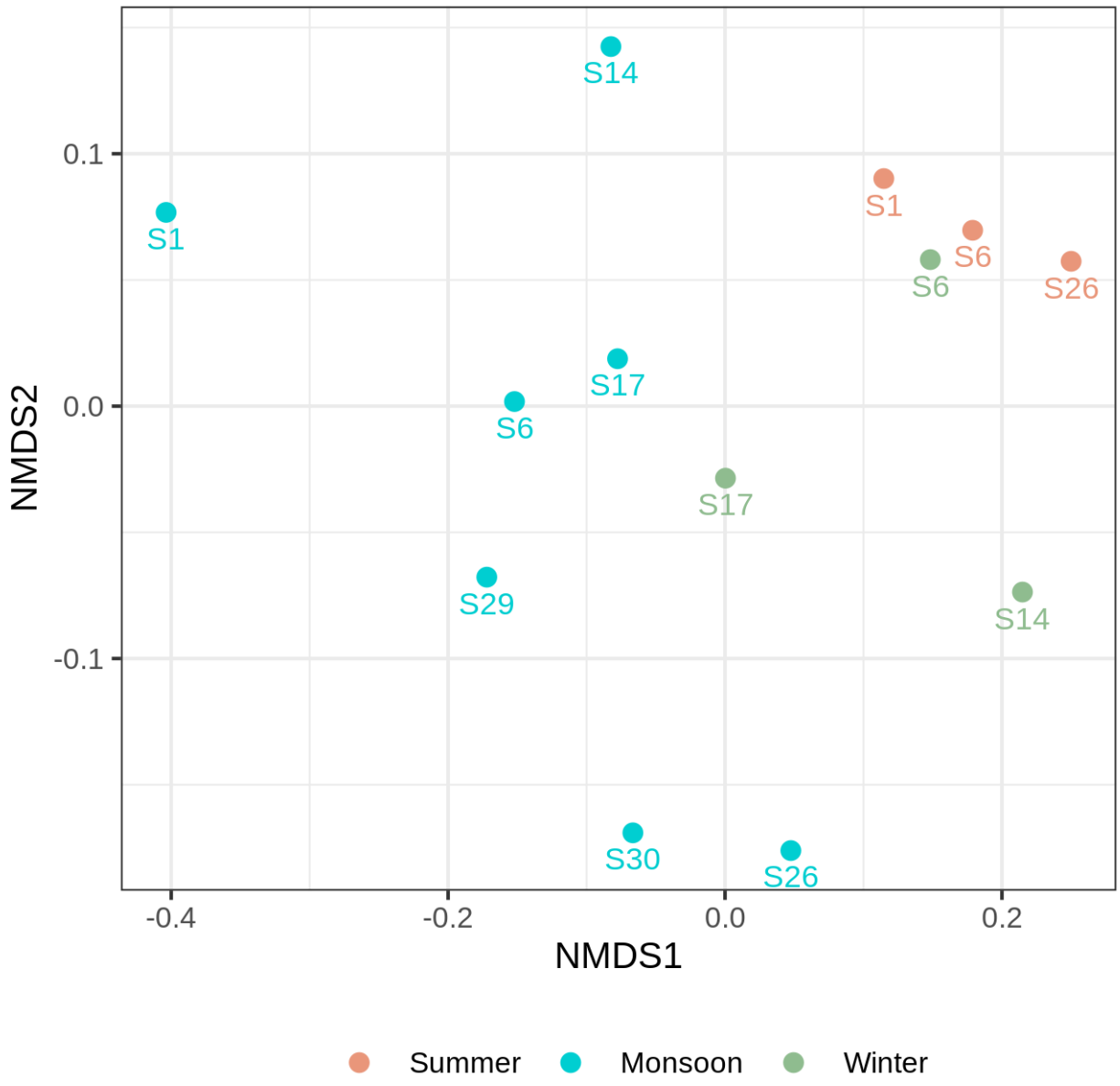


Figure 4: **Spatiotemporal variation in biodiversity across the tree of life** Ordination of Bray-Curtis dissimilarity among the spatiotemporal samples of the Chilika lagoon. The colors represent the different seasons and the labels indicate the different sampling stations. A total of 7 samples from the monsoon season and three randomly selected samples from the summer and winter seasons each were sequenced to assess the spatiotemporal variation in biodiversity.

DISCUSSION

Assessment of taxonomic diversity across the tree of life from extracellular eDNA

Our results demonstrate that PCR-free deep sequencing of extracellular eDNA is a promising approach for taxonomic diversity assessment across the tree of life in large aquatic ecosystems. By generating one of the deepest shotgun sequencing datasets of extracellular eDNA, we push the limits of biodiversity assessment to detect taxa across the tree of life, including the relatively low abundant non-microbial taxa in the ecosystem. Through statistical extrapolation of richness accumulation curves, we show that the asymptotic taxonomic richness of the ecosystem across the tree of life can be reliably estimated. Further, we also found that extracellular eDNA provides broad-scale spatiotemporal resolution to detect changes in the relative abundance of taxa across the tree of life.

We achieved these results due to the adaptations at every level of the workflow from sample collection, eDNA extraction, library preparation, sequencing, and bioinformatics. We enriched the extracellular eDNA using a lysis-free protocol rather than using the total eDNA to avoid DNA extraction bias due to the differences in lysis efficiencies between cell types from a wide range of taxa (Djurhuus et al., 2017). By eliminating the PCR from the laboratory workflow by using PCR-free library preparation methods, we achieved very low duplication rates in the sequences which otherwise may render a considerable part of the data useless by reducing the effective depth and increasing the PCR-induced artifacts (Kebschull & Zador, 2015). As the probability of detection of low-abundant taxa is determined by the depth of sequencing, we estimated the required sequencing depth by analyzing the library complexity. We then sequenced the extracellular eDNA libraries to the point of saturation by employing an extremely high-throughput sequencing platform. Further, to achieve sensitive taxonomic classifications, we derived two independent taxonomic assignments from the paired-end reads using protein-based classification algorithms and calculated the lowest common ancestor taxa for each read. The reads of bacterial origin dominated the taxonomic assignments (86.95%) due to the high abundance of Bacteria in aquatic ecosystems. However, the family richness of Eukaryotes was higher than Bacteria possibly due to a large number of eukaryotic families represented in the reference database compared to prokaryotes (Supplementary Fig. 5). Studies in the past could not detect a high diversity of Eukaryotes from shotgun sequencing of total eDNA mainly due to the shallow sequencing depth (22.3 million) and a low percentage of reads assigned to Eukaryota (0.34%) (Stat et al., 2017). We achieved over sixteen-fold more taxonomic assignments to Eukaryota (5.48%) and detected hundreds of families of Protists, Fungi, Plants, and Animals. Particularly, the high diversity of Metazoan families indicates detectable amounts of DNA from non-microbial species in the extracellular eDNA for shotgun sequencing approaches. This opens up the possibility of detecting taxa across the tree of life without using any targeted enrichment techniques such as PCR or hybridization capture that can introduce a bias toward certain taxa (van der Loos & Nijland, 2021). We also showed that statistical extrapolation of taxonomic richness accumulation curves can be used to account for the undetected taxa with very low abundances and estimate the asymptotic richness across the tree of life. The estimates of asymptotic family richness were in line with the expected richness of well characterized taxa in the ecosystem such as fishes. Such estimates of total taxonomic richness can be used to monitor the changes in taxonomic richness across the tree of life over a long period and help in identifying and prioritizing taxa for conservation. Although we did not detect any substantial change in the composition of taxonomic families among the samples, we detected high variation in the relative abundance of the families across space and time. This indicates that the taxonomic families in the ecosystem can remain largely unchanged while their relative abundance may vary in the given spatiotemporal scale. Furthermore, the genome-scale data generated using this approach can also be repurposed for assessing diversity at the gene level, mapping functional traits to specific taxa, inferring species co-occurrence patterns, and linking community changes to ecosystem functioning and services.

Limitations

The taxonomic resolution achievable through deep sequencing of extracellular eDNA is generally lower compared to approaches targeting a barcoding region in the genome. The taxonomic classification of the extracellular eDNA sequences depends upon the taxonomic resolution of various genomic loci that are stochastically captured, the sensitivity of the algorithm used to detect homology, and the availability of reference sequences from the target organisms. Different regions in the genome provide variable taxonomic resolutions depending on the sequence complexity, mutation rate, selection pressure, recombination, and evolutionary history of the species (Coissac et al., 2016). Further, sensitive alignment-based homology detection algorithms such as BLAST (Altschul, 2014) are prohibitively slow to query billions of reads against large reference databases. Alternative alignment-free kmer-based algorithms such as KRAKEN2 (Wood et al., 2019) are thousands of times faster than BLAST but far less sensitive and cannot find homology between highly divergent species (Lindgreen et al., 2016). Due to the sparsity of existing reference sequence databases, many underrepresented taxa may remain undetected and lead to underestimates of taxonomic diversity when using DNA-based classifiers. Hence, we adopted a protein-based classification algorithm as the protein sequences are more conserved than the genomic DNA sequences and offer better sensitivity with incomplete databases than DNA-based algorithms (Menzel et al., 2016). Even when the exact species is not represented in the database, the sequences can be taxonomically identified using the evolutionarily closest species present in the database as a proxy. Protein-based classification also eliminates erroneous taxonomic assignments from repetitive DNA sequences that are abundant in Eukaryotic genomes. But the trade-off of using protein-based over DNA-based classification is the lower taxonomic resolution due to the conservation of protein sequences among closely related species. However, such trade-offs are inevitable when accurate estimates of taxonomic richness are required, especially when assessing a tropical ecosystem like ours where the majority of the diversity is yet to be documented.

Sequencing costs and the availability of genome-scale data are the main limiting factors for the adoption of deep sequencing of extracellular eDNA for taxonomic assessment of ecosystems. Deep sequencing of samples to the point of saturation may quickly become infeasible for large-scale projects with hundreds of samples. Decreasing the sampling resolution and using statistical extrapolations as demonstrated in this study can bring down sequencing costs and enable the assessment of large ecosystems. Moreover, advancements in sequencing technologies are expected to decrease the sequencing cost to as less as $1 per GB in the near future which will make it more affordable. Furthermore, only a small fraction of all the known species have their genomes assembled, annotated, and archived in public sequence databases. Nevertheless, an international moonshot initiative in biology called the Earth BioGenome Project is set to change the scenario of incomplete databases by generating genomic resources for all the known eukaryotic species (about 1.5 million) in a record time of over a decade (Lewin et al., 2018). Several large-scale genome sequencing initiatives across the world have joined this massive effort targeting a wide variety of taxa. With the progress and completion of various genome sequencing initiatives, the increased availability of reference sequences in the databases will improve the sensitivity and specificity of the taxonomic assignments and provide a more accurate snapshot of taxonomic diversity.

Conclusion

Extracellular eDNA is a natural repertoire of genetic material from all the organisms inhabiting an ecosystem and is a reliable source for taxonomic diversity assessment. Organisms across the tree of life can be effectively detected through PCR-free deep sequencing of extracellular eDNA. The total taxonomic richness of the ecosystem can be estimated through statistical extrapolation of richness accumulation curves derived from incidence frequencies of taxa in the extracellular eDNA sequences. Extracellular eDNA also provides broad-scale spatiotemporal resolution of changes in biodiversity across the tree of life in an ecosystem. With plummeting sequencing costs and increasing coverage of reference databases by large-scale genome sequencing projects, we envision the wide adoption of PCR-free deep sequencing of extracellular eDNA for large-scale biodiversity assessment across the tree of life. Although there is further scope to test and opti-

mize the workflow, we believe that this study significantly advances our understanding of the capabilities and limits of extracellular eDNA for taxonomic diversity assessment. Its application to detect taxa across the tree of life is fundamental for a paradigm shift toward implementing large-scale next-generation bioassessment and biomonitoring programs for the conservation, restoration, and management of ecosystems in the Anthropocene.

REFERENCES

Altschul, S. F. (2014). BLAST Algorithm. In ELS. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0005253.pub2

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. Conservation Genetics, 17(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4

Bascompte, J. (2009). Disentangling the Web of Life. Science, 325(5939), 416–419. https://doi.org/10.1126/science.1170749

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution, 29(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Brantschen, J., Blackman, R. C., Walser, J.-C., & Altermatt, F. (2021). Environmental DNA gives comparable results to morphology-based indices of macroinvertebrates in a large-scale ecological assessment. PLOS ONE, 16(9), e0257510. https://doi.org/10.1371/journal.pone.0257510

Bruce, K., Blackman, R., Bourlat, S. J., Hellström, A. M., Bakker, J., Bista, I., Bohmann, K., Bouchez, A., Brys, R., Clark, K., Elbrecht, V., Fazi, S., Fonseca, V., Hänfling, B., Leese, F., Mächler, E., Mahon, A. R., Meissner, K., Panksep, K., ... Deiner, K. (2021). A practical guide to DNA-based methods for biodiversity assessment. Advanced Books, 1, e68634. https://doi.org/10.3897/ab.e68634

Bushnell, B. (2022). BBTools. Retrieved January 12, 2023, from https://jgi.doe.gov/data-and-tools/software-tools/bbtools/

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science Advances, 1(5), e1400253. https://doi.org/10.1126/sciadv.1400253

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proceedings of the National Academy of Sciences, 114(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Ceballos, G., Ehrlich, P. R., & Raven, P. H. (2020). Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. Proceedings of the National Academy of Sciences, 117(24), 13596–13602. https://doi.org/10.1073/pnas.1922686117

Chao, A., & Jost, L. (2015). Estimating diversity and entropy profiles via discovery rates of new species. Methods in Ecology and Evolution, 6(8), 873–882. https://doi.org/10.1111/2041-210X.12349

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. Molecular Ecology, 25(7), 1423–1428. https://doi.org/10.1111/mec.13549

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology, 5(1), 3–21. https://doi.org/10.1093/jpe/rtr044

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. Methods in Ecology and Evolution, 7(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature Methods, 10(4), Article 4. https://doi.org/10.1038/nmeth.2375

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology, 26(21), 5872–5895. https://doi.org/10.1111/mec.14350

Djurhuus, A., Port, J., Closek, C. J., Yamahara, K. M., Romero-Maraccini, O., Walz, K. R., Goldsmith, D. B., Michisaki, R., Breitbart, M., Boehm, A. B., & Chavez, F. P. (2017). Evaluation of Filtration and DNA Extraction Methods for Environmental DNA Biodiversity Assessments across Multiple Trophic Levels. Frontiers in Marine Science, 4. https://doi.org/10.3389/fmars.2017.00314

Dysthe, J. C., Rodgers, T., Franklin, T. W., Carim, K. J., Young, M. K., McKelvey, K. S., Mock, K. E., & Schwartz, M. K. (2018). Repurposing environmental DNA samples—Detecting the western pearlshell (Margaritifera falcata) as a proof of concept. Ecology and Evolution, 8(5), 2659–2670. https://doi.org/10.1002/ece3.3898

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: An R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution, 7(12), 1451–1456. https://doi.org/10.1111/2041-210X.12613

IUCN Red List of Threatened Species. (2022). [Resource]. IUCN. Retrieved January 12, 2023, from https://www.iucn.org/resources/conservation-tool/iucn-red-list-threatened-species

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., Møller, P. R., & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. Molecular Ecology Resources, 21(3), 690–702. https://doi.org/10.1111/1755-0998.13293

Kebschull, J. M., & Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Research, 43(21), e143. https://doi.org/10.1093/nar/gkv717

Leempoel, K., Hebert, T., & Hadly, E. A. (2020). A comparison of eDNA to camera trapping for assessment of terrestrial mammal diversity. Proceedings of the Royal Society B: Biological Sciences, 287(1918), 20192353. https://doi.org/10.1098/rspb.2019.2353

Lever, M. A., Torti, A., Eickenbusch, P., Michaud, A. B., Šantl-Temkiv, T., & Jørgensen, B. B. (2015). A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. Frontiers in Microbiology, 6. https://doi.org/10.3389/fmicb.2015.00476

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences, 115(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Liang, Z., & Keeley, A. (2013). Filtration Recovery of Extracellular DNA from Environmental Water Samples. Environmental Science & Technology, 47(16), 9324–9331. https://doi.org/10.1021/es401342b

Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep19233

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences, 113(21), 5970–5975. https://doi.org/10.1073/pnas.1521291113

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE, 8(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications, 7(1), Article 1. https://doi.org/10.1038/ncomms11257

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How Many Species Are There on Earth and in the Ocean? PLOS Biology, 9(8), e1001127. https://doi.org/10.1371/journal.pbio.1001127

Nagler, M., Insam, H., Pietramellara, G., & Ascher-Jenull, J. (2018). Extracellular DNA in natural environments: Features, relevance and applications. Applied Microbiology and Biotechnology, 102(15), 6343–6356. https://doi.org/10.1007/s00253-018-9120-4

Nagler, M., Podmirseg, S. M., Ascher-Jenull, J., Sint, D., & Traugott, M. (2022). Why eDNA fractions need consideration in biomonitoring. Molecular Ecology Resources, 22(7), 2458–2470. https://doi.org/10.1111/1755-0998.13658

Pawlowski, J., Apothéloz-Perret-Gentil, L., & Altermatt, F. (2020). Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. Molecular Ecology, 29(22), 4258–4264. https://doi.org/10.1111/mec.15643

Rodriguez-Ezpeleta, N., Morissette, O., Bean, C. W., Manu, S., Banerjee, P., Lacoursière-Roussel, A., Beng, K. C., Alter, S. E., Roger, F., Holman, L. E., Stewart, K. A., Monaghan, M. T., Mauvisseau, Q., Mirimin, L., Wangensteen, O. S., Antognazza, C. M., Helyar, S. J., de Boer, H., Monchamp, M.-E., … Deiner, K. (2021). Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: Comment on "Environmental DNA: What's behind the term?" by Pawlowski et al., (2020). Molecular Ecology, 30(19), 4601–4605. https://doi.org/10.1111/mec.15942

Seeber, P. A., McEwen, G. K., Löber, U., Förster, D. W., East, M. L., Melzheimer, J., & Greenwood, A. D. (2019). Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. Molecular Ecology Resources, 19(6), 1486–1496. https://doi.org/10.1111/1755-0998.13069

Seymour, M. (2019). Rapid progression and future of environmental DNA research. Communications Biology, 2(1), Article 1. https://doi.org/10.1038/s42003-019-0330-9

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. Scientific Reports, 7(1), Article 1. https://doi.org/10.1038/s41598-017-12501-5

Stewart, K. A. (2019). Understanding the effects of biotic and abiotic factors on sources of aquatic environmental DNA. Biodiversity and Conservation, 28(5), 983–1001. https://doi.org/10.1007/s10531-019-01709-8

Suresh, V. R., Mohanty, S. K., Manna, R. K., Bhatta, K. S., Mukherjee, M., Karna, S. K., Sharma, A. P., Das, B. K., Pattnaik, A. K., & Nanda, S. (2018). Fish and shellfish diversity and its sustainable management in Chilika Lake (p. 376). ICAR- Central Inland Fisheries Research Institute, Chilika Development Authority.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. Molecular Ecology, 21(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology, 21(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet, P., Prud'homme, S. M., Campione, E., Roy, J., Miquel, C., Shehzad, W., Gielly, L., Rioux, D., Choler, P., Clément, J.-C., Melodelima, C., Pompanon, F., & Coissac, E. (2012). Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Molecular Ecology, 21(8), 1816–1820. https://doi.org/10.1111/j.1365-294X.2011.05317.x

The Catalogue of Life. (2022). COL. Retrieved January 12, 2023, from https://www.catalogueoflife.org/

Thomas, A. C., Howard, J., Nguyen, P. L., Seimon, T. A., & Goldberg, C. S. (2018). eDNA Sampler: A fully integrated environmental DNA sampling system. Methods in Ecology and Evolution, 9(6), 1379–1385. https://doi.org/10.1111/2041-210X.12994

Thomas, A. C., Nguyen, P. L., Howard, J., & Goldberg, C. S. (2019). A self-preserving, partially biodegradable eDNA filter. Methods in Ecology and Evolution, 10(8), 1136–1141. https://doi.org/10.1111/2041-210X.13212

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation, 183, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

van der Loos, L. M., & Nijland, R. (2021). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology, 30(13), 3270–3288. https://doi.org/10.1111/mec.15592

Watson, R., Baste, I., Larigauderie, A., Leadley, P., Pascual, U., Baptiste, B., Demissew, S., Dziba, L., Erpul, G., & Fazel, A. (2019). Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES Secretariat: Bonn, Germany, 22–47.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 257. https://doi.org/10.1186/s13059-019-1891-0

ACKNOWLEDGEMENTS

We acknowledge the funding received by G.U. for this study from the Department of Biotechnology (DBT), Govt. of India, vide grant no. BT/PR29032/FCB/125/4/2018. S.M. was supported by a DBT-BINC PhD fellowship from the Govt. of India and an Early Career Grant from the National Geographic Society vide grant no. EC-53199T-18. We are grateful to Dr. Gurdeep Rastogi of the Chilika Development Authority for facilitating the sample collection and for his support throughout the project. We thank Manisha Ray, Divyasree Karne, Madhusmita Mohapatra, and Gopi Krishnan for their help during the sampling expeditions and DNA extraction. We also thank Tulasi Nagabandi of the CCMB NGS facility for her help with the preparation of libraries and sequencing.

DATA AVAILABILITY

All the raw sequencing data generated in this study has been submitted to NCBI SRA under the BioProject accession PRJNA691704. The SRA run accessions for all the samples are provided in Supplementary Table 1. The data will be released publicly immediately after acceptance. The Snakemake pipeline and R scripts used for analysis are available at https://github.com/manu-script/TaxDivToL. The GitHub repository will be archived, and a permanent DOI link will be provided after acceptance.

AUTHOR CONTRIBUTIONS

S.M. and G.U. conceived the idea and designed the study. S.M. collected the samples, generated the data, and analyzed the data. S.M. drafted the manuscript and G.U. edited the final manuscript.

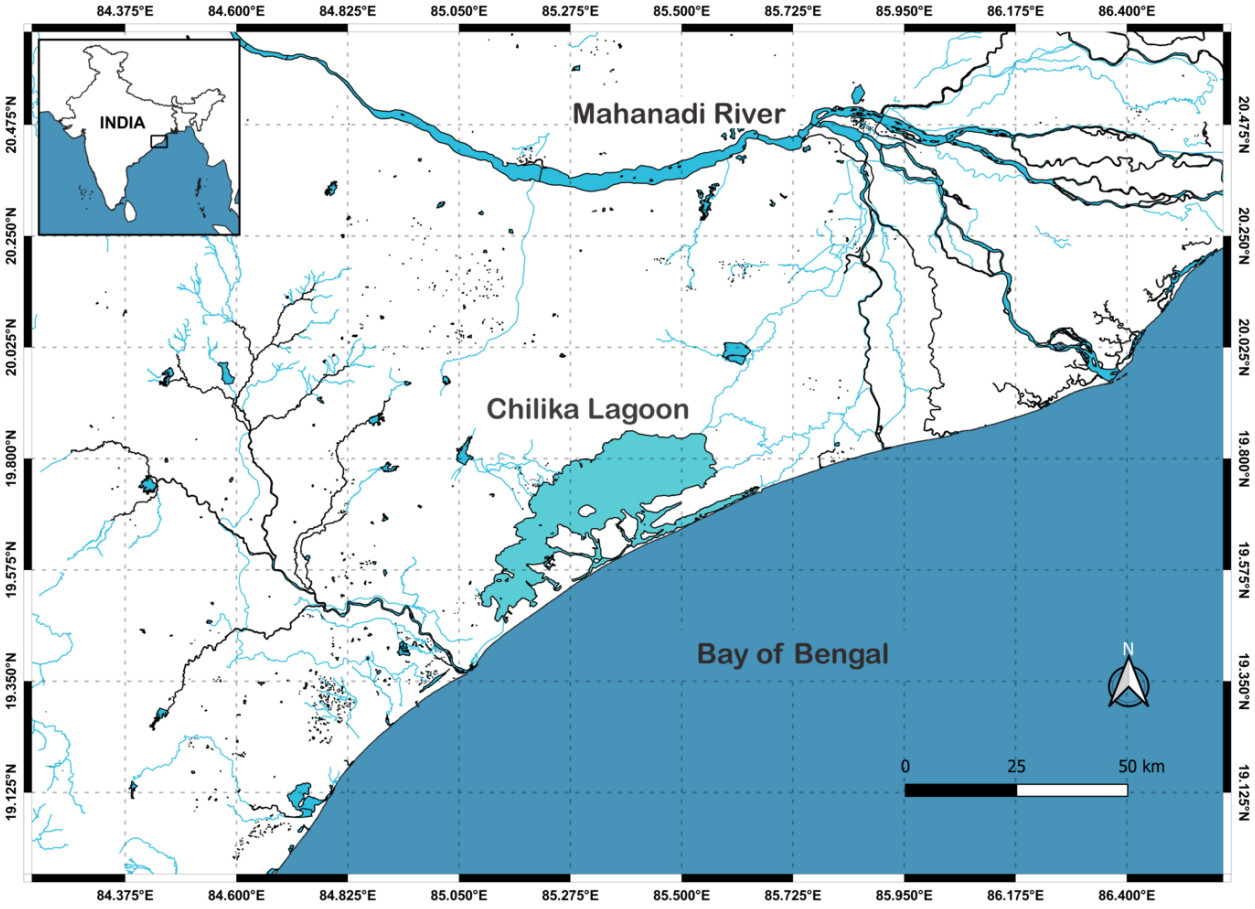
COMPETING INTERESTS

The authors declare no competing interest.

SUPPLEMENTARY FIGURES AND TABLES

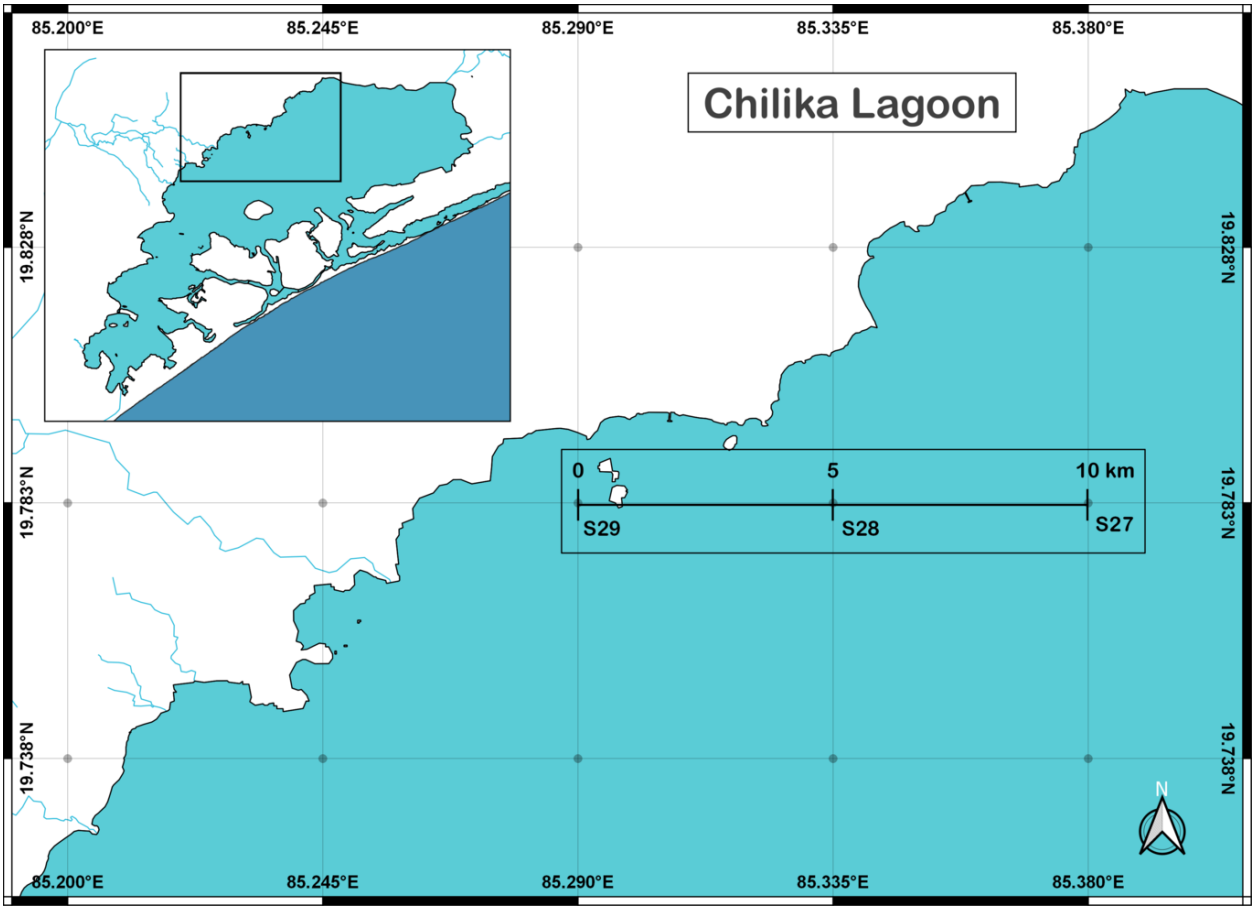
Supplementary Figure 1. Geographic overview of the Chilika lagoon.

The brackish lagoon is located on the east coast of India and receives freshwater from the Mahanadi River system and marine water from the Bay of Bengal.



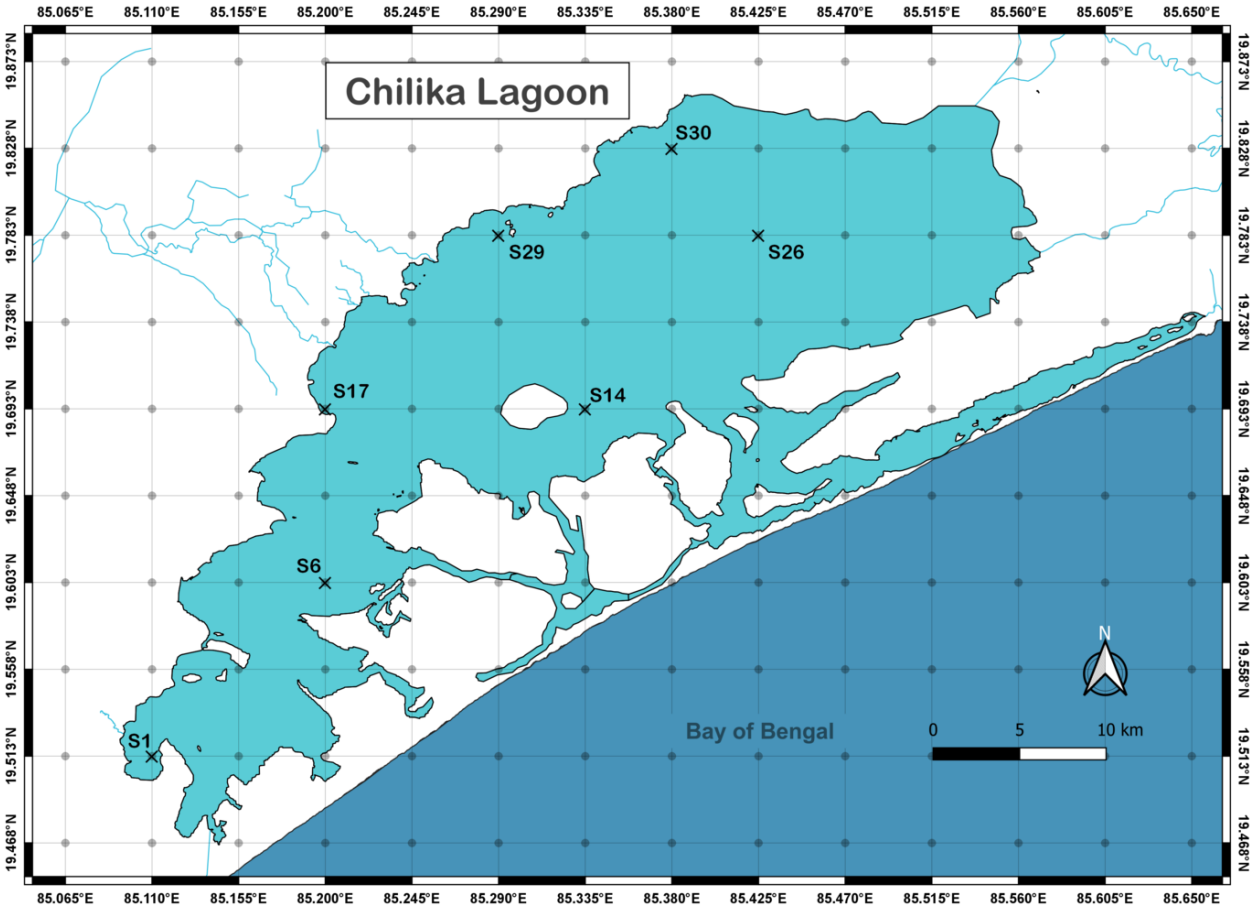
Supplementary Figure 2. Pilot study design in the Chilika lagoon.

The three geolocated sampling stations (S27, S28, and S29) are located 5 km apart on a 10 km transect in the central sector of the lagoon.



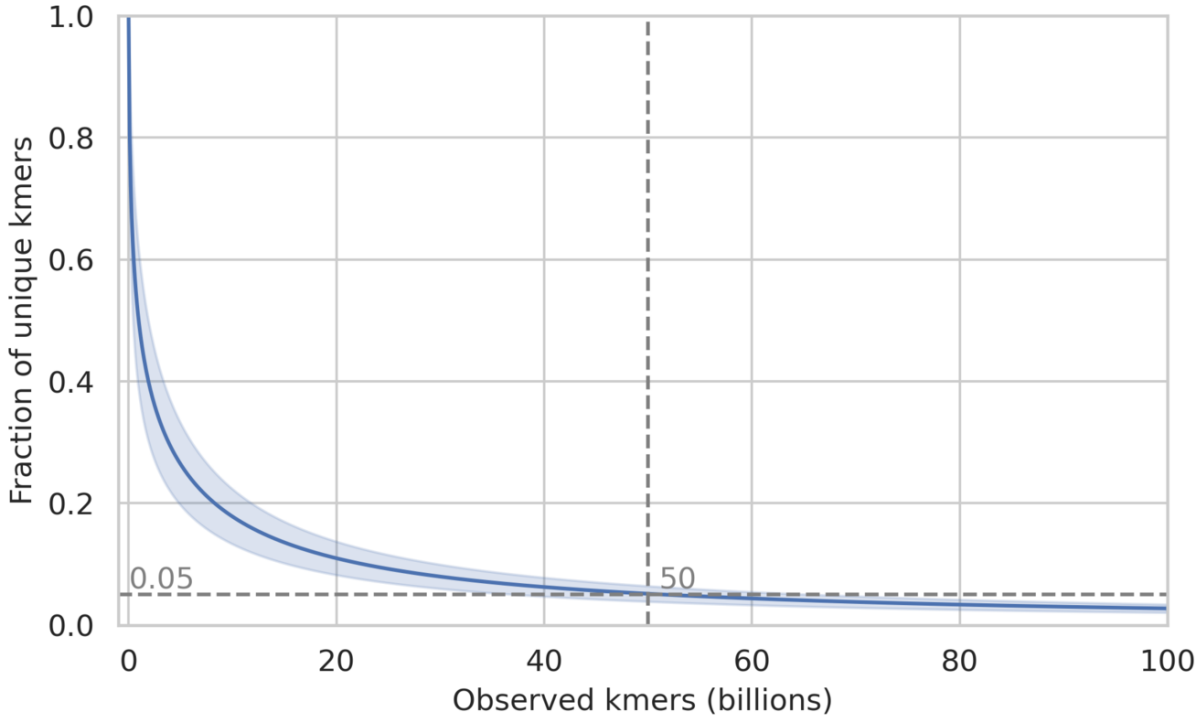
Supplementary Figure 3. Spatiotemporal sampling design.

The sampling points are spread apart at a distance of 5 to 15 km from each other across the Chilika lagoon for the spatiotemporal study.



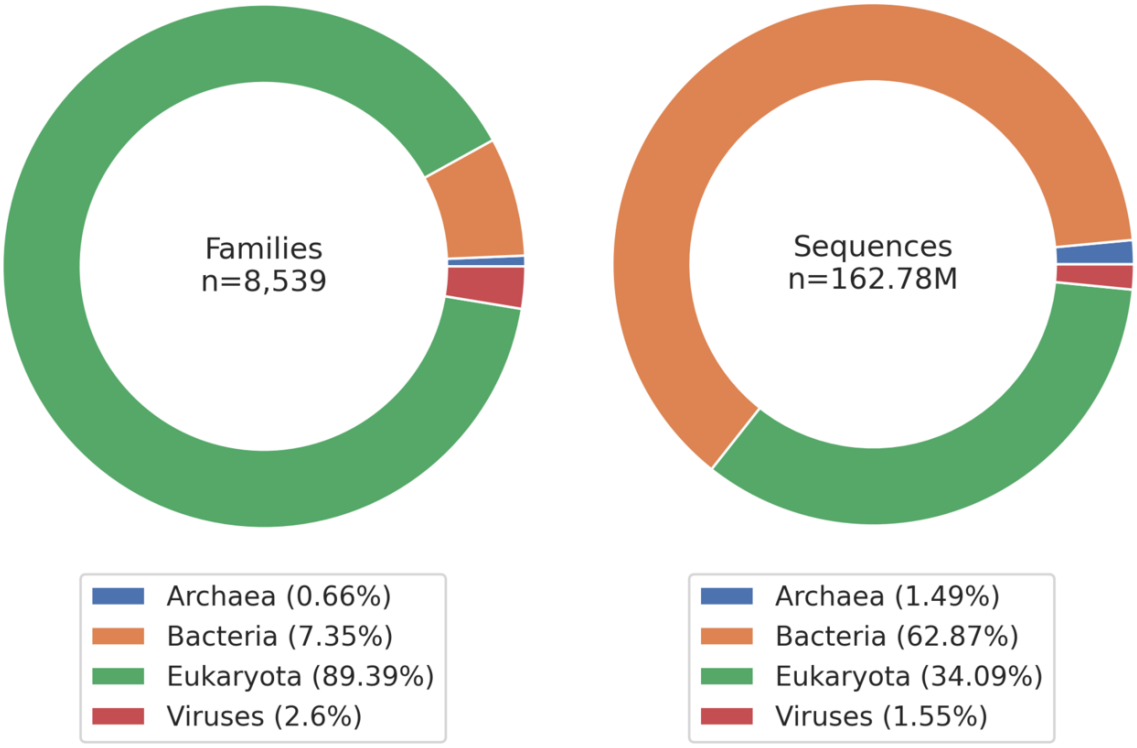
Supplementary Figure 4. Saturation of unique kmers.

Library complexity characterized by the fraction of unique kmers as a function of all observed 31 bp kmers. The solid blue line and the shaded area around the line represent the mean and standard deviation from the 3 pilot study samples. The dotted grey lines depict the 95% saturation of unique kmers at 50 billion observed kmers.



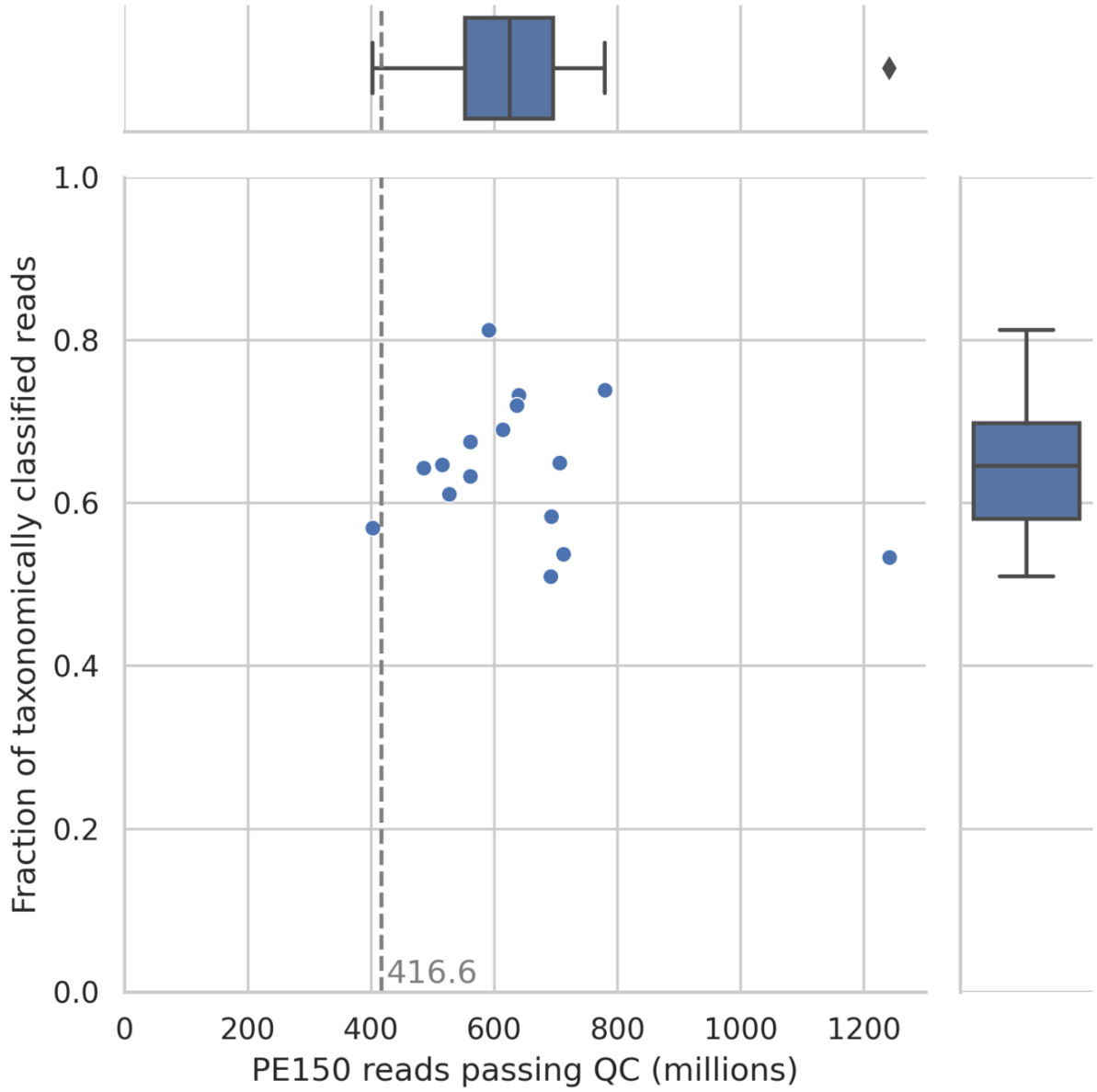
Supplementary Figure 5. UniRef100 Database composition.

Distribution of the number of taxonomic families and sequences among the different domains of life in the UniRef100-based reference database used for taxonomic classification.



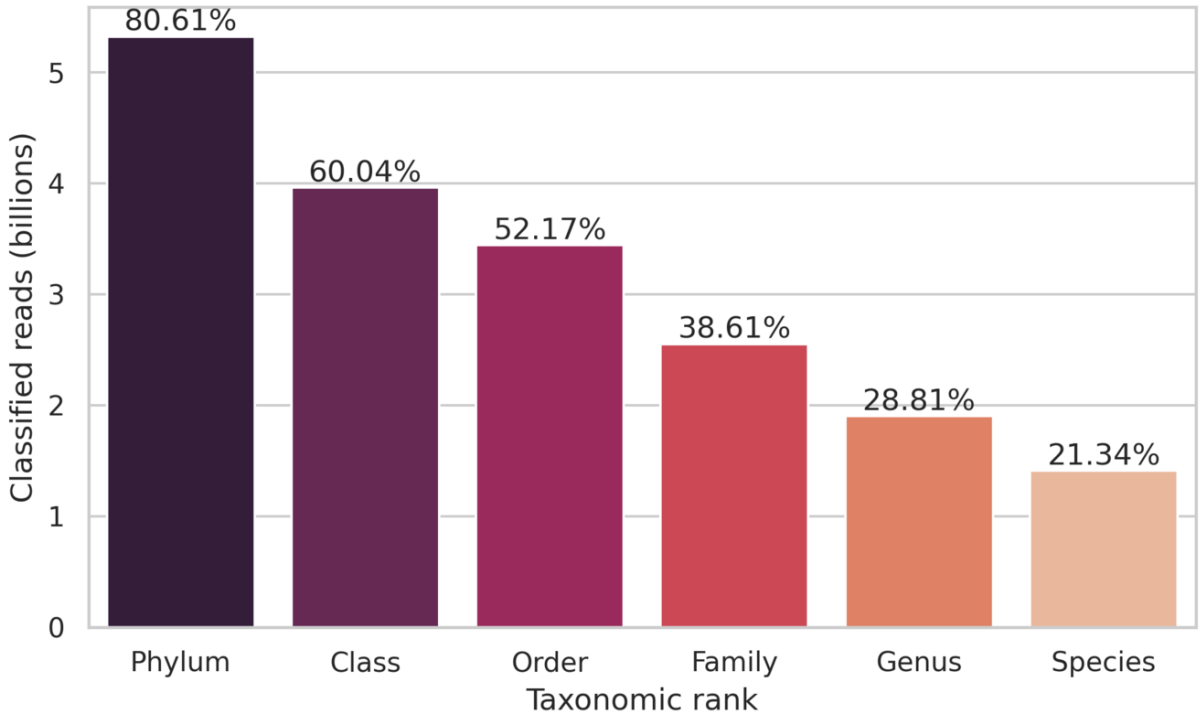
Supplementary Figure 6. Proportion of taxonomically classified reads.

Taxonomic classification rate and sequencing depth of all the samples in this study. The grey dotted line indicates the targeted sequencing depth based on the library complexity of samples from the pilot study.



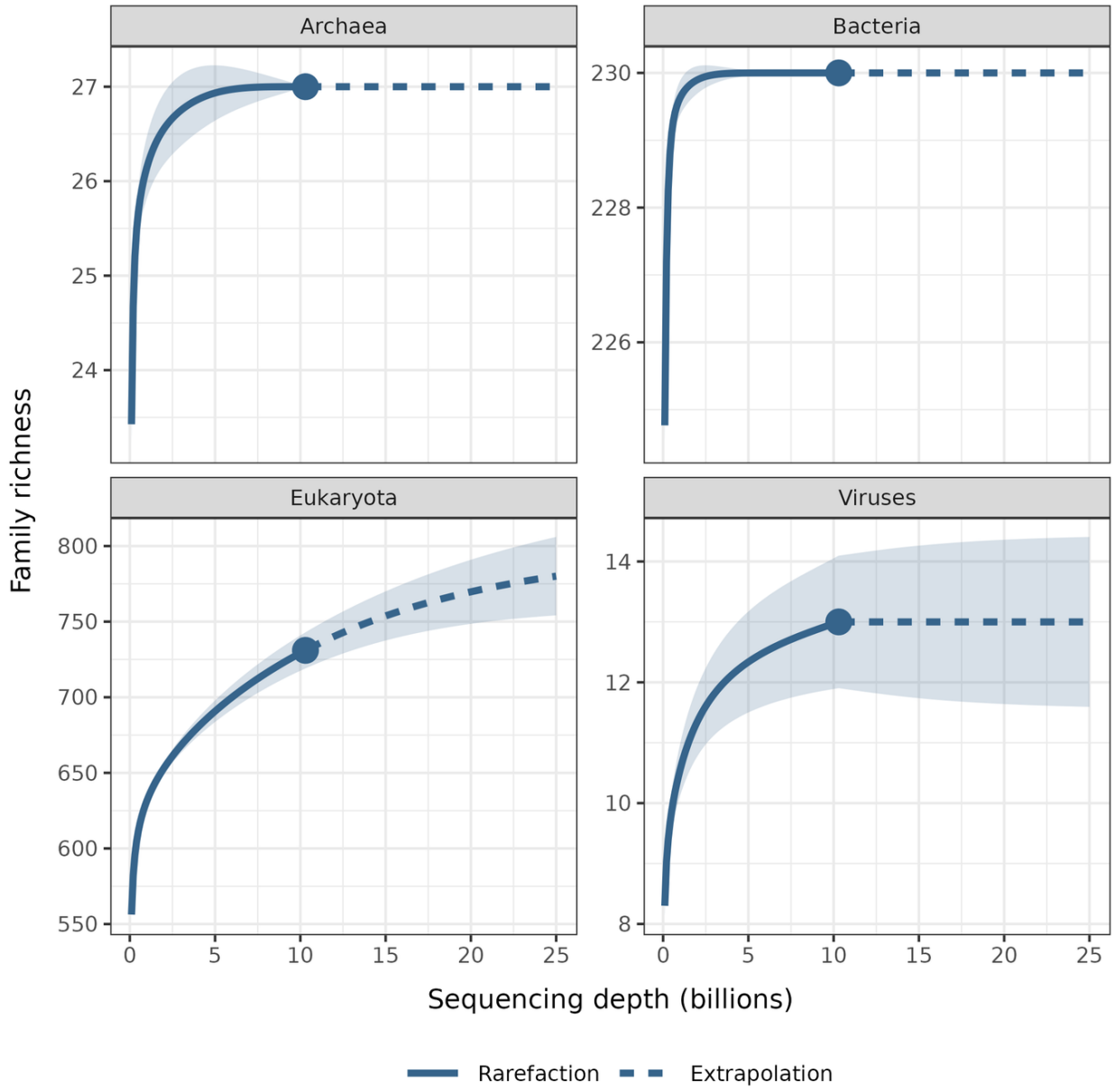
Supplementary Figure 7. Proportion of all the classified reads at different taxonomic ranks.

The percentage of classified reads from Phylum to Species level calculated with respect to the total number of classified reads at the domain level (n=6.59 billion).



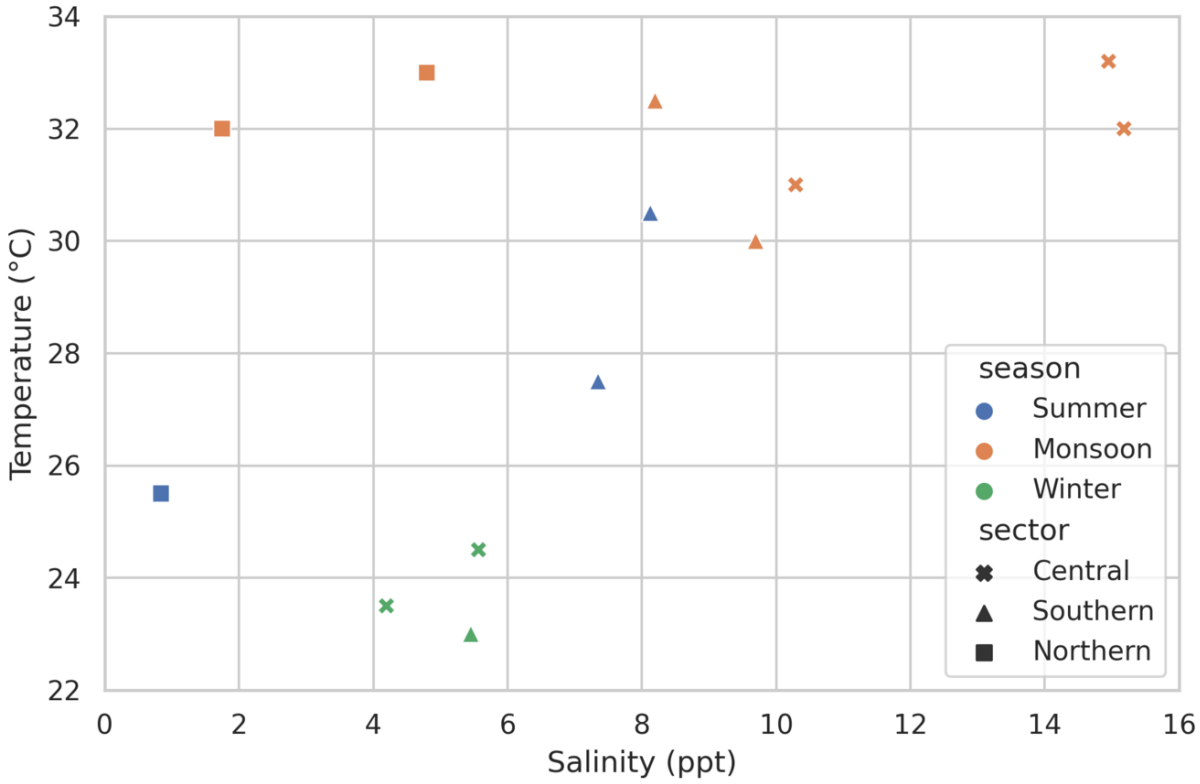
Supplementary Figure 8. Richness accumulation curves.

Family richness accumulation curves of different domains of life as a function of sequencing depth. The solid and dotted lines represent the rarefaction and extrapolation with 95% confidence intervals, respectively.



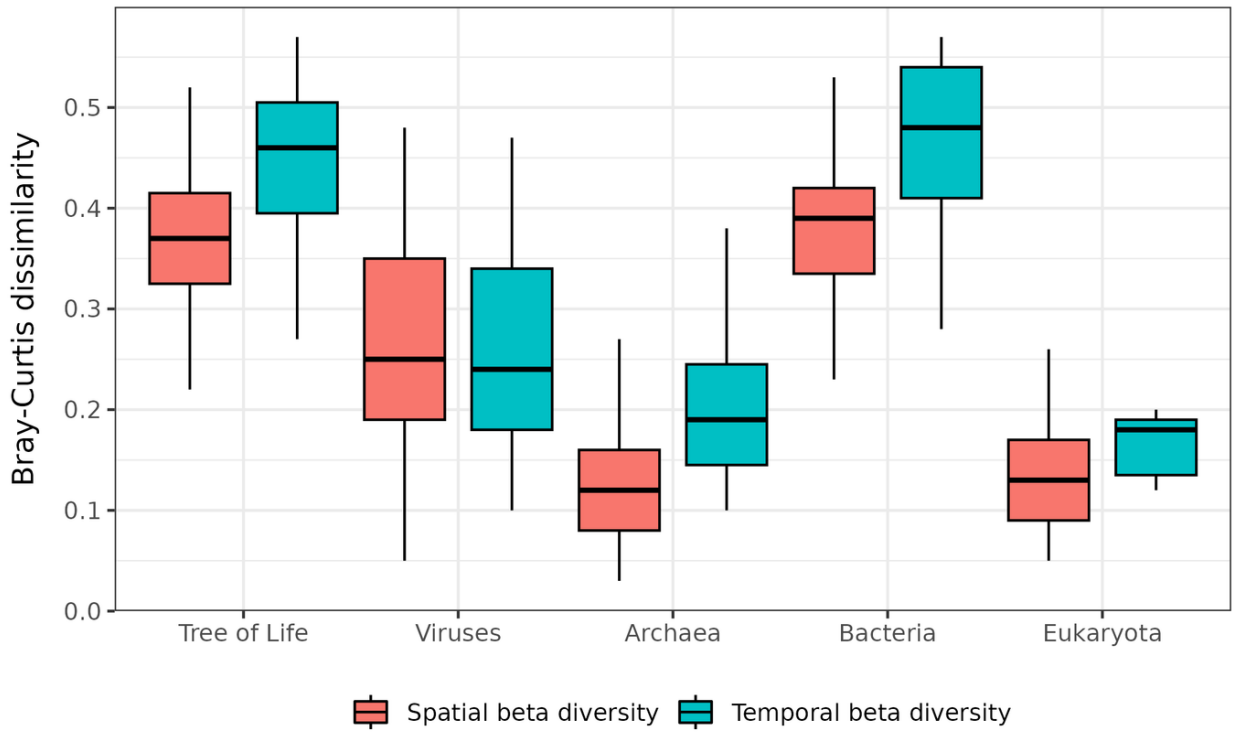
Supplementary Figure 9. Variation of environmental conditions at the sampled sites.

Scatter plot of water temperature and salinity during the time of sampling. The colors and shapes represent the different seasons and sectors of the lagoon.



Supplementary Figure 10. Spatiotemporal beta diversity.

Distribution of spatial and temporal beta diversity across the tree of life and different domains as measured by Bray-Curtis dissimilarity.



Supplementary Table 1. Geographic coordinates, sequencing depth, and NCBI SRA accession numbers of all the samples. ADDITIONAL FILES

Hosted file

`Supplementary Data.xlsx` available at https://authorea.com/users/381963/articles/510280-taxonomic-diversity-assessment-across-the-tree-of-life-from-extracellular-environmental-dna-in-aquatic-ecosystems