Received: 29 March 2024 | Revised: 17 June 2024 | Accepted: 9 July 2024

DOI: 10.1002/edn3.593

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences
WILEY

ORIGINAL ARTICLE

Enriching barcoding markers in environmental samples utilizing a phylogenetic probe design: Insights from mock communities

Kevin Nota[1] | Ludovic Orlando[2] | Alexis Marchesini[3,4] | Matteo Girardi[5] | Stefan Bertilsson[1,6] | Cristiano Vernesi[4,7] | Laura Parducci[1,8]

[1]Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[2]Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT), CNRS UMR 5288, Université Paul Sabatier, Toulouse, France
[3]Research Institute on Terrestrial Ecosystems (IRET), National Research Council (CNR), Porano, Italy
[4]National Biodiversity Future Center, Palermo, Italy
[5]Conservation Genomics Unit, Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige (TN), Italy
[6]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden
[7]Forest Ecology Unit, Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige (TN), Italy
[8]Department of Environmental Biology, Sapienza University of Rome, Rome, Italy

Correspondence
Kevin Nota and Laura Parducci, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden.
Email: kevin_nota@eva.mpg.de and laura.parducci@uniroma1.it

Present address
Kevin Nota, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Funding information
H2020 European Research Council, Grant/Award Number: 101071707-Horsepower and 681605-PEGASUS; Swedish Phytogeographical Society; University Paul Sabatier (AnimalFarm IRP); CNRS

Abstract

Hybridization capture is an emerging method making use of short oligonucleotide baits to enrich DNA libraries for genomic fragments of specific organisms thus enabling detection of their presence in environmental samples. Although it offers a primer-independent alternative to metabarcoding, little empirical work has been dedicated to characterizing the underlying biases and coupled implications for biological interpretation. Moreover, few published bioinformatic pipelines are available for designing polynucleotide capture baits from a reference sequence collection. We designed RNA-baits specifically targeting two chloroplast barcoding genes *mat*K and *rbc*L to reveal the plant taxonomic diversity present in a given environmental sample. Our approach leverages the sensitivity of hybridization capture and the capacity of high-throughput DNA sequencing instruments. It builds on a new and universal method based on ancestral sequence reconstruction, ultimately limiting the number of bait-probes required and reducing experimental costs, while accessing high levels of taxonomic diversity. Our bait-set selectively targets four main plant orders (Fagales, Pinales, Asterales, and Poales), representing ~18% of all described vascular plants. This is achieved through the use of only 4084 baits, each 80 nucleotides in length (80-mer), capturing ~1.0–1.6 k nucleotide sequences from each taxon. Tests on mock communities revealed important factors influencing capture efficiency and relative abundance estimates, including GC-content, the overall target length per

Cristiano Vernesi and Laura Parducci shares joint senior authorship.

taxa, and the bait density and mean number of mismatches to the bait sequence. Our results show that hybridization capture, like metabarcoding, requires caution when interpreting results quantitatively within (paleo)-ecological studies. Biases detected in this work have the potential to be mitigated with bait designs that avoid extreme base compositional biases and balancing bait targets across taxa. However, we strongly recommend the use of mock communities and read simulations to quantify the accuracy of taxonomic representation when using new bait designs.

KEYWORDS
capture bias, DNA barcoding, hybridization capture, shotgun metagenomics, target capture, target enrichment

1 | INTRODUCTION

Over the last decade, environmental DNA (eDNA) has emerged as a new and powerful tool for cost- and time-effective characterization of the biological diversity present in a given ecosystem community (Ruppert et al., 2019; Taberlet et al., 2018). The molecular method that first revolutionized the field is metabarcoding, and still represents the most widely applied approach. Metabarcoding is based on PCR amplification and high-throughput sequencing of short regions of a gene (i.e., amplicons), whose sequence acts as a diagnostic (mini)-barcode for taxonomic identification (Meusnier et al., 2008). Taxonomic assignments are either based on sequence comparison to a reference database or alternatively, especially for microorganisms, on pre-clustering of reads into mOTUs (molecular operational taxonomic units) or ASVs (amplicon sequence variants) (Eren et al., 2013; Taberlet et al., 2012). Although no universal standards have been proposed, the underlying laboratory and bioinformatic toolkit for metabarcoding is relatively well established and streamlined, with many published laboratory protocols and bioinformatic pipelines and workflows (Taberlet et al., 2018). Important biases and limitations are fairly well characterized, and extensively studied (Nichols et al., 2018; Rodriguez-Martinez et al., 2022; Zinger et al., 2019), mostly in relation to PCR amplification, which can selectively over- or under-amplify certain alleles due to reduced polymerase fidelity, variable amplicon lengths, polymerase processivity linked to GC-content (Nichols et al., 2018) and formation of secondary structures of the targets (steric hindrance). The marker amplicon length is another potential hindrance, particularly reducing PCR success when analyzing environmental samples where DNA tends to be highly fragmented and degraded (e.g., ancient DNA). There is a trade-off between amplification success (and therefore taxonomic coverage) and the length and diagnostic value (i.e., taxonomic resolution) of the amplified region; that is, DNA fragments too short to accommodate both primer binding sites will not be amplified. Despite such limitations, metabarcoding remains a powerful and widely used method due to the increasing availability, and in some cases, relatively complete reference databases, low sequencing and laboratory costs, and general sensitivity (Taberlet et al., 2018).

Shotgun metagenome sequencing is an alternative method that aims to sequence all fragments in an eDNA sample using either single or double-stranded library preparations (Gansauge et al., 2017; Meyer & Kircher, 2010). The approach is very powerful; and for example, it has revealed the ecological diversity of microbial, plant, and animal diversity up to 2 million year (Fernandez-Guerra et al., 2023; Kjær et al., 2022) in samples where PCR metabarcoding did not yield positive results. In microbiome analyses, shotgun metagenomics can provide highly resolved insights into communities were the majority of the microbial populations are unculturable (Bendall et al., 2016; Frémont et al., 2022; Nayfach et al., 2021; Richter et al., 2022). One of the key advantages of such shotgun sequencing applied to ancient samples is the capacity to disentangle ancient templates from modern contaminants on the basis of typical sequence signatures of postmortem DNA damage (Briggs et al., 2007; Jónsson et al., 2013). However, shotgun sequencing is costly and requires a higher sequencing effort to detect low-abundance organisms. Additionally, in most studies conducted to date, over 90% of the reads remain unassigned due to the absence of reference sequences in relevant databases (Ahmed et al., 2018; Graham et al., 2016; Parducci et al., 2019; Pedersen et al., 2016; Wang et al., 2021). However, this percentage may vary depending on the field, with some prokaryote studies reporting 82% successful mapping of shotgun metagenomic reads (Mthethwa-Hlongwa et al., 2024).

Hybridization capture (also called "target capture") enriches' target reads by focusing subsequent metagenomic sequencing efforts on nucleic acid strands captured by a predefined collection of baits covering taxonomically-informative biomarkers (Armbrecht et al., 2021; Murchie, Kuch, et al., 2021; Murchie, Monteath, et al., 2021; Schulte et al., 2022; Slon et al., 2017; Vernot et al., 2021). This method requires prior knowledge of the targeted sequence diversity as well as non-targets to enable rational design of complementary and specific (RNA/DNA)-baits (also called probes) to retrieve all the targets of interest. More specifically, synthesized polynucleotide baits are biotinylated and anneal to complementary eDNA metagenomic library templates. Streptavidin-coated paramagnetic beads are immobilized with a magnet while non-targeted library molecules wash away. The immobilized library molecules are eluted for further amplification and sequenced. Multiple studies have successfully applied hybridization capture to enrich DNA library content for both full mitochondrial genomes (Slon et al., 2017) and shorter barcode genes

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences
WILEY

(e.g., Armbrecht et al., 2021; Foster et al., 2021; Lentz et al., 2021; Murchie, Kuch, et al., 2021; Murchie, Monteath, et al., 2021). For example, Slon et al. (2016, 2017) designed a set of baits to co-analyze the presence of 242 full mammalian mitochondrial genomes in cave sediments. Assuming sufficient target length, the approach preserves sequence signatures of ancient DNA damage patterns, which are paramount to data authentication. Moreover, the retrieval of fragments will not be restrained by availability of primer binding sites, or other PCR biases. Despite these potential advantages, only a few bait designs are readily available such as the "PalaeoChip Arctic-1.0 baitset," which is designed to capture full mitochondrial genomes sublimated with commonly used barcoding genes such as COI, cytb, 12S, and 16S for ~180 vertebrate taxa and *mat*K, *rbc*L, and *trn*L for over 2500 plant taxa occurring in Quaternary artic and boreal environments (Murchie, Kuch, et al., 2021). Such designs tend to be developed and validated only for particular biomes or regions since there are experimental and economic constraints on the number of baits that can be synthesized. To mitigate these issues, many bait designs reduce the number of capture sequences by clustering oligos within a set genetic distance, such as 96% identity score (Murchie, Kuch, et al., 2021). Nevertheless, these approaches might bias the results by containing baits that are most similar to the allele that was used as cluster centroid.

There is an growing number of studies reporting factors limiting the efficacy of hybridization capture (Chilamakuri et al., 2014; Cruz-Dávalos et al., 2017; Suchan, Chauvey, et al., 2022; Suchan, Kusliy, et al., 2022), identifying GC-content, complexity, read length, and reference bias, as affecting target capture outcomes. Less attention has been given to address specific biases of this method for enriching targets in environmental samples. Therefore, the extent to which the full taxonomic diversity originally present in the eDNA metagenomic library will be reflected in results from hybridization capture datasets is presently unknown. This study was designed to fill this gap in knowledge. First, we developed a universal method for designing bait sequences using online barcode sequence repositories. We then designed a bait panel targeting the chloroplast *rbc*L and *mat*K genes from four plants orders (Fagales, Pinales, Asterales, and Poales), representing approximately 18% of the total number of accepted species in the World Checklist of Vascular Plants (Govaerts et al., 2021). We experimentally tested the performance of our bait set using mock communities, and identified important biases impacting downstream abundance estimates and ecological interpretation by comparing "unbiased" shotgun and target enriched sequencing data from these mock communities.

2 | MATERIALS AND METHODS

2.1 | Ancestral sequence reconstruction

All available *mat*K and *rbc*L sequences, totaling 22,103 and 28,148 sequences respectively, were downloaded for the plant orders Fagales, Pinales, Asterales, and Poales from BOLD (October 18, 2019; Table 2) using the BOLD package (v0.8.6) in R (Chamberlain, 2021). Multiple sequence alignments (MSA) were made for each plant order and marker independently, using MAFFT with default settings with *–adjustdirection* as an additional flag (Katoh et al., 2002). Sequence alignments were visually checked to remove poorly aligned and low-quality sequences (e.g., sequences with frameshift, or non-coding triplets), and manually corrected for nucleotide calling errors when the original electropherograms were available from BOLD.

The BOLD database contains reference sequences covering either the complete or partial length of the standardized barcodes. Due to variations in primers and sequence trimming, the start- and endpoints of these records differ. For bait design, we focused on the section of the MSA that was represented by the most reference sequences, to avoid enriching for parts of the barcode gene that has little reference material. To determine the most suitable windows, we investigated the number of sequences and alignment length by sub-setting the MSA at the first and last position where a minimal percentage of the reference sequences had data (from 0 to 0.99 with 0.1 increments). Our approach aimed to maximize the number of sequences with the longest possible gene fragment size by multiplying the proportion sequences that covered the selected window by the length of the window (Figure 1). After size trimming, *rbc*L and *mat*K alignments were visually checked once again, and corrected for gaps based on the protein sequences. Finally, the alignments were concatenated into longer sequences maximizing phylogenetic signal for downstream bait design. Taxa for which only one marker was available were excluded (leaving: $N = 1721$ and $N = 2539$, for *mat*K and *rbc*L respectively, Table 2).

To avoid over-representing the most abundant sequences and to decrease redundancy, we decided to design bait sequences on reconstructed ancestral sequences. This method has been previously used to capture the mitogenomes of extinct animal species (glyptodont), for which no close relative exists (Delsuc et al., 2016). Phylogenetic trees were generated for each order using *MrBayes* (Ronquist et al., 2012) using the GTR substitution model (rates = invgamma). Bayesian MCMC analysis was run until the average standard deviation of split frequencies approached its lower plateau (~0.02–0.12, i.e., for 4e5 to 1e6 generations). Sample frequency was set to 10. *MrBayes* MCMC analysis was summarized (3e5 burn-ins, 1e6 generations), and a majority rule consensus tree was created. *HyPhy* was then used to fit the sequence data with the consensus tree using the GTR model (standard MG94 fit), before reconstructing ancestral sequences for all internal nodes of the consensus tree (Pond et al., 2005). The gene trees were obtained at order level, except for Poales and Asterales for which many reference sequences were available. For these orders, the reference sequences were split into clades that group families (e.g., Graminid, Cyperid, and Graminid for Poales). For families with a large number of species and sequences, only one species per genus was randomly selected.

2.2 | Selecting ancestral sequences for bait design

Bait-template annealing remains effective for target enrichment despite the presence of up to 10%–13% substitutions

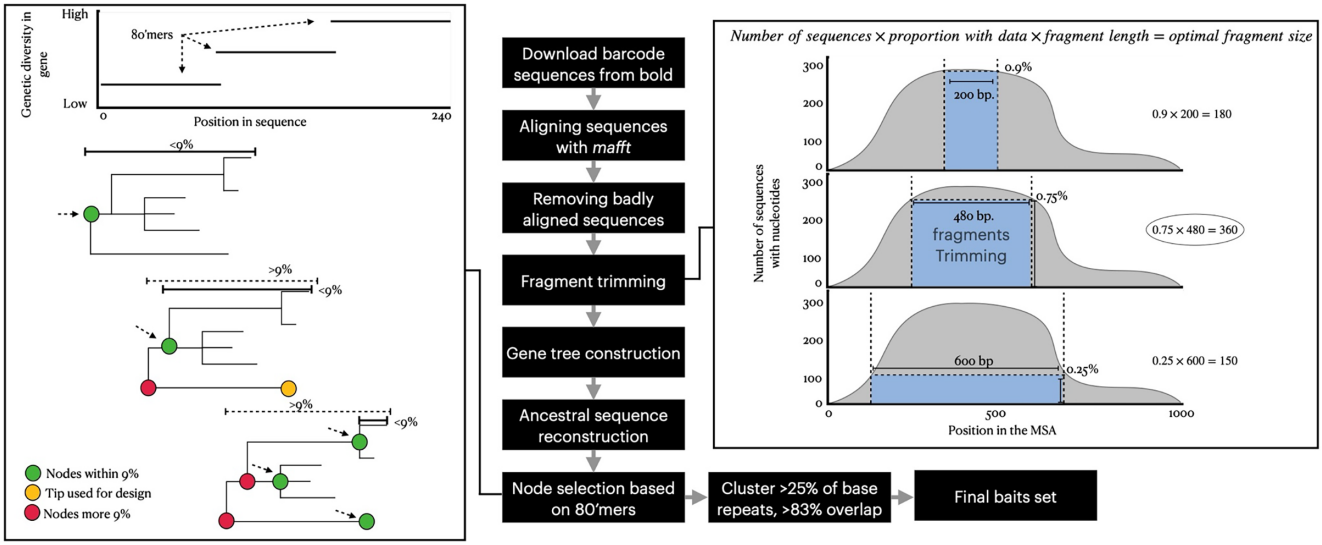

FIGURE 1 Bioinformatic pipeline for bait design. Reference sequences are obtained from BOLD and aligned with *mafft*. Poorly aligned sequences are manually discarded. From the resulting multiple sequence alignment (MSA), the fragment that maximizes the number of full-length sequences is selected while retaining sufficient reference data (left panel). A gene tree is then inferred to reconstruct ancestral sequences for each node. The nodes required for capture are selected by calculating node-to-tip distances. Nodes that can capture all tips within 9% are selected for bait design (right panel). Further filtering involves removing sequences with 25% base repeats and >83% overlap with 100% identity from the selected nodes.

TABLE 1 Overview of the mock community.

		*rbc*L		*mat*K	
Order	Taxa	GC	Total length[a]	GC-content	Total length[a]
Asterales	*Nymphoides peltata*	0.424	430	0.324	585
Fagales	*Betula pubescens*	0.434	430	0.339	585
	Quercus robur	0.444	430	0.349	630
Pinales	*Juniperus communis*	0.427	430	0.330	1130
	Larix decidua	0.449	430	0.375	1222
Poales	*Bromus madritensis*	0.434	430	0.333	567
	Glyceria notata	0.444	430	0.327	614
	Juncus effusus	0.406	430	0.295	567
	Typha latifolia	0.424	430	0.315	576

[a]The total length refers to the length of the marker that was enriched for.

(Mason et al., 2011; Paijmans et al., 2016; Peñalba et al., 2014). Therefore, it is not necessary to consider all phylogenetic nodes for capturing the large diversity of plant species found within the target orders. To decrease redundancy, we selected nodes where tips showed at most 9% nucleotide dissimilarity to the most distal ancestral sequence. Because the amount of nucleotide variation is variable within gene regions, nodes were selected within moving windows 80 nucleotides in length, representing the bait size. This was done by sub-setting the ancestral and tip sequences into 80 bases and calculating the distance from each internal node sequence to all connecting tips. The procedure was repeated until all tip sequences were matched to an ancestral node, and until the whole gene was covered with a one-base tiling. In cases where sequences were too distant from their closest ancestral node, the tip sequence was used for bait design. This procedure resulted in a multifasta file including sequences at those selected nodes and tips, which was used for designing 80-base long baits with 3× tiling. Finally, baits with >83% overlap and 100% identity were collapsed to a single random seed sequence with *cd-hit* (Li & Godzik, 2006). This was done to reduce the number of baits that were designed on ancestral nodes which showed high similarity. Furthermore, bait candidates showing more than 25% of base repeats were removed by myBaits®, Biosciences (USA) before bait array synthesis to filter for low complexity regions. The overall framework is presented on Figure 1, and an updated version of its computational implementation in python can be accessed at; https://github.com/Kevinnota/gotcha.

To validate the bait design, we simulated a total of 250,000 reads using *gargammel* (Renaud et al., 2017) using a size distribution of

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences
WILEY

TABLE 2 Counts for the BOLD dataset.

	BOLD sequences				Filtered and fragment trimmed				Ancestral Seq. Rec.			
	N Seq		*N* genera		*N* filtered Seq		*N* Uniq Seq		*N* Seq (uniq taxa)	*N* genera	Prop Seq used bait design	
Order	rbcL	matK	rbcL	matK	rbcL	matK	rbcL	matK	All	All	rbcL	matk
Asterales	7810	6538	743	885	5568	3724	1037	1523	537	431	0.52	0.35
Fagales	978	1317	37	38	747	794	139	228	126	26	0.91	0.55
Pinales	2908	2102	74	74	2385	900	319	325	251	52	0.79	0.77
Poales	16,452	12,146	741	680	9963	7860	1606	2516	1136	431	0.71	0.45
Total	28,148	22,103	1595	1677	18,663	13,278	3101	4592	2050	940	–	–

fragments typical for a degraded sample. The simulated data were restricted to the window of the gene that was used for bait design. The simulated reads were mapped to the bait sequences using *bowtie2* (Langmead & Salzberg, 2012) using the *--sensitive-local* setting. The bam file was parsed using *pysam* (https://github.com/pysam-developers/pysam) to retain the alignment length, cigar string, and number of mismatches. The cigar string was used for the reads mapping with indels to identify the longest alignment, and the number of mismatches in that region. Overall, less than 0.5% of the simulated reads did not map to any of the bait sequences, while more than 99% of the mapped reads were within 9% identity to the baits. A total of 4084 80-mer RNA baits were produced at myBaits®, Arbor Biosciences (USA; for the bait sequences see, Data S1).

2.3 | DNA extraction, mock community preparation, and sequencing

DNA was extracted from fresh leaf tissues collected from nine selected plant taxa (see Table 1). These taxa were selected based on their availability in the laboratory and to encompass a wide range of GC-content. Each plant tissue sample was extracted three times using the DNeasy Plant Mini Kit (Qiagen), following the manufacturer's instructions but with a single elution of DNA in 100 uL Elution Buffer. For each extraction batch, two extraction blanks were run alongside the samples. To generate a high number of on target template molecules, and to make mixtures of equal concentration of each taxon for investigating capture biases, we amplified the targeted loci for each species and shredded the PCR product before capture. The rbcLa gene was amplified using the rbcL-aF and rbcL-aR described in Kress & Erickson (2007). Each PCR reaction contained 1.25 U of GoTaq G2 Hot Start Taq Polymerase (Promega), 10 uL of 5× Green GoTaq Flexi Buffer, 0.4 uM of each primer, 0.2 mM of each dNTP, 2 mM of MgCl2, 2 μL of template DNA, in a 50 μL reaction. The PCR cycling protocol was as follows: 95°C for 2 min, followed by 40 cycles of 95°C for 30 s, 50°C (except for *Nymphoides peltata*, *Juniperus communis*, and *Betula pubescens*, for which annealing temperature was set to 52°C), and 72°C for 90 s, and a final extension of 5 min at 72°C. The *mat*K locus was amplified with modified primers Plant_matK413f-2 (5′-TAATTTACGATCYATTCATTCAATATTTYC-3′) and matK-1227r-1 (5′GARGATCCRCTRTRATAATGAGAAAGATTT-3′) from (Heckenhauer et al., 2016). Two Pinales species were amplified using the matK-F and matK-R described in (Kusumi et al., 2000). The PCR mixture and cycling conditions were similar to those considered for the *rbcL*a locus, with the following modifications: MgCl2 concentration was increased to 3 mM, and the annealing temperature was reduced to 48°C. For the Pinales species the MgCl2 concentration remained at 2 mM, and the annealing temperature was set to 52°C.

The PCR product of each amplification was fragmented using Covaris M200 Sonicator (Covaris) using microTUBE AFA Fiber Pre-Slit Snap-Cap 6×16 mm. Two fragmentation steps were carried out, each at 7°C, with peak incident power of 75, duty factor (%) of 25, Cycles per burst (cpb) of 215 and AVG power of 18.8. The fragmentation durations were 380, 480, and 580 s for *rbc*L, *mat*K and *mat*K-Pinales, respectively. The fragmentation efficacy was checked and quantified on an Agilent 2100 Bioanalyzer system (Agilent), (Figures S1 and S2). A total of 3.00×10^7 copy DNA/uL of each fragmented PCR product were pooled and converted into a double stranded Illumina library, using the method described in Meyer and Kircher (2010), with the adapters and indexes from NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1). DNA libraries were PCR indexed using 16 cycles, and quantified using the Collibri™ Library Quantification Kit (Invitrogen™), four times, considering 100-fold and 1,000-fold dilutions. All negative controls showed >100-fold lower copy numbers than the sample libraries. Eight identical mixtures were made using 30,000,000 copies of each library. Four of the library mixtures were captured using the myBaits plant bait set designed for this study, following the myBaits version 5.01 manual, and a hybridization temperature of 65°C. All eight mixtures, comprising four captured and four uncaptured samples, were paired-end sequenced on an iSeq 100 Sequencing System (Single end, 300 cycles, Illumina), which provided a total of 36,186–193,580 reads per mixture.

2.4 | Bioinformatic processing and analysis of sequencing data

The raw reads were paired and trimmed using PEAR with a minimal assembly length of 25 nucleotides (Zhang et al., 2014). Duplicated

reads were removed with *seqkit rmdup* (Shen et al., 2016). All reads were mapped to the *rbc*L and *mat*K reference sequences (excluding primer binding sites) using *bowtie2* with the *-sensitive* settings, and *-k* set to 5000. Taxonomic assignment was performed with the lowest common ancestor inference tool *ngsLCA* (Wang et al., 2022), considering all aligned pairs showing at least 95% similarity. All downstream GC and read length distributions were estimated on reads that were assigned to the lowest common ancestor with *ngsLCA*. The alignment length was calculated over the longest ungapped fraction from unique paired reads mapped against the bait sequences, using *bowtie2* with the *–sensitive-local* settings, keeping only the best hit (Langmead & Salzberg, 2012) using the mapping CIGAR with *pysam*.

Linear regressions were performed using the change in relative abundance of reads assigned to lowest common ancestor post-capture as a response variable with the *lm*() function from the stats package in R (version 4.3.1). The best fitting model was chosen using the *stepAIC* function from the MASS package (Venables & Ripley, 2002). The predictor factors were calculated from local alignments of reads obtained from shotgun sequencing to baits as described before, except for retaining all mappings with the *-k* flag set to 4000. The GC-content and number of mismatches were calculated over the longest ungapped fraction of the reads. For the GC-content mismatches were ignored since those bases have no impact on the binding. A median was taken to summarize the GC-content and number of mismatches for each read. The GC-values were then averaged over the different taxonomic groups and a median was used for the number of mismatches. To investigate the relative importance of the predicating variables the *calc.relimp* function from the *relaimpo* package was used (Grömping, 2006). All downstream analyses were performed in R (R Core Team, 2022) and visualized using *ggplot2* (Wickham, 2016). All scripts and data are accessible, see data statement.

3 | RESULTS

3.1 | Bait design and validation

In total, 18,663 *rbc*L and 13,278 *mat*K sequences obtained from BOLD passed initial filtering, combined representing 7947 unique plant taxa, spread across 41 families (bold taxonomic identifiers). A total of 2050 taxa were used for bait design with "gotcha," which produced a total of 4084 80-mers baits as detailed in Table 2. Of these baits, 3228 (79.0%) cover the *mat*K locus, while 856 (21.0%) covered the *rbc*L locus. We found that only 1227 of 250,000 simulated reads (0.49%) did not align against any of the 4084 bait sequences. The read size of the unmapped reads were significantly shorter compared to the simulated reads (Kolmogorov–Smirnov test, D^- = 0.53428, *p*-value < 2.2e-16). Overall, ~37.5% of the simulated reads mapped without mismatch against the bait sequences, ~49.8% of the alignments featured one or two nucleotide mismatches, while ~12.1% had three or more such mismatches. Of all the reads simulated, ~98.9% aligned over at least 30 nucleotide bases, and about 86.2% for at least 60 nucleotide bases. Based on identity score, we estimated that ~99.4% of the simulated reads are within a 9% distance from their respective bait sequences.

3.2 | Potential capture biases

The iSeq 100 sequencing run produced a total of 1,321,692 raw reads (128,887 ± 43,902) across the eight DNA libraries sequenced. After paired-end merging, 16.2 ± 0.15% (shotgun) and 22.8 ± 0.6% (capture) reads did not map against any reference sequences. Of these unmapped reads, 85.1 ± 0.9% of the shotgun and 89.3 ± 0.3% of the captured reads mapped using a local alignment, against at least two different reference sequences for more than 85% of the read length. The level of duplication was overall low, with 2.66 ± 0.54% being duplicated reads in normalized shotgun libraries, and 3.51 ± 0.79% in the captured libraries. Over the whole dataset (four libraries combined), the duplication rate of the shotgun libraries was 10.41% versus 13.08% for the captured libraries. Relative duplication rate change was highest in the *mat*K gene with an average of 73.1 ± 22.3% more duplicates in the capture library than in the shotgun library (bins 30–40 to 140–150 nt.), while the relative duplication rate was 24.5 ± 15.9% in the *rbc*L (bins 30–40 to 140–150 nt.; Figures S4 and S5). The theoretical complexity based on simulations was not exhausted, with only 9.7% of the fragments with a GC-content between 30.0% and 60.0% recovered with the present sequencing effort (see Figure S6).

The read length distribution of merged and trimmed reads after capture had a lower number of reads in the size range 30–70 nucleotides compared to the shotgun libraries (Figure 2). A minor change in read length distribution was also observed between 70 and 100 nucleotides and an increase in templates longer than 110 nucleotides (Figure 2a). We found that the number of reads showing 20–45 nucleotides overlap to a bait sequence decreased by more than half in capture experiments as compared to the shotgun reference library. The relative changes between 55 and 70 nucleotides are minimal (~±14%), while the number of libraries with a bait annealing over the whole length increased by 38.9%. The proportion of shotgun reads that did not map against any of the bait sequences was ~1.4% and ~11.7% for *rbc*L and *mat*K, respectively. Post-capture, this proportion was reduced to less than 0.4% for both markers (Figure S7).

We next investigated whether base composition of the baits influenced capture efficacy as multiple studies have reported variable on-target enrichment-folds according to bait GC-content (Cruz-Dávalos et al., 2017; Suchan, Chauvey, et al., 2022; Suchan, Kusliy, et al., 2022). To test this, we plotted the normalized read counts and the fold change between the captured and uncaptured sequence dataset, considering GC-categories in 5% intervals (Figure 2c). The relative change showed a depletion of reads with a CG content between 15% and 30% was between −65.7% and −31.3%. In contrast,

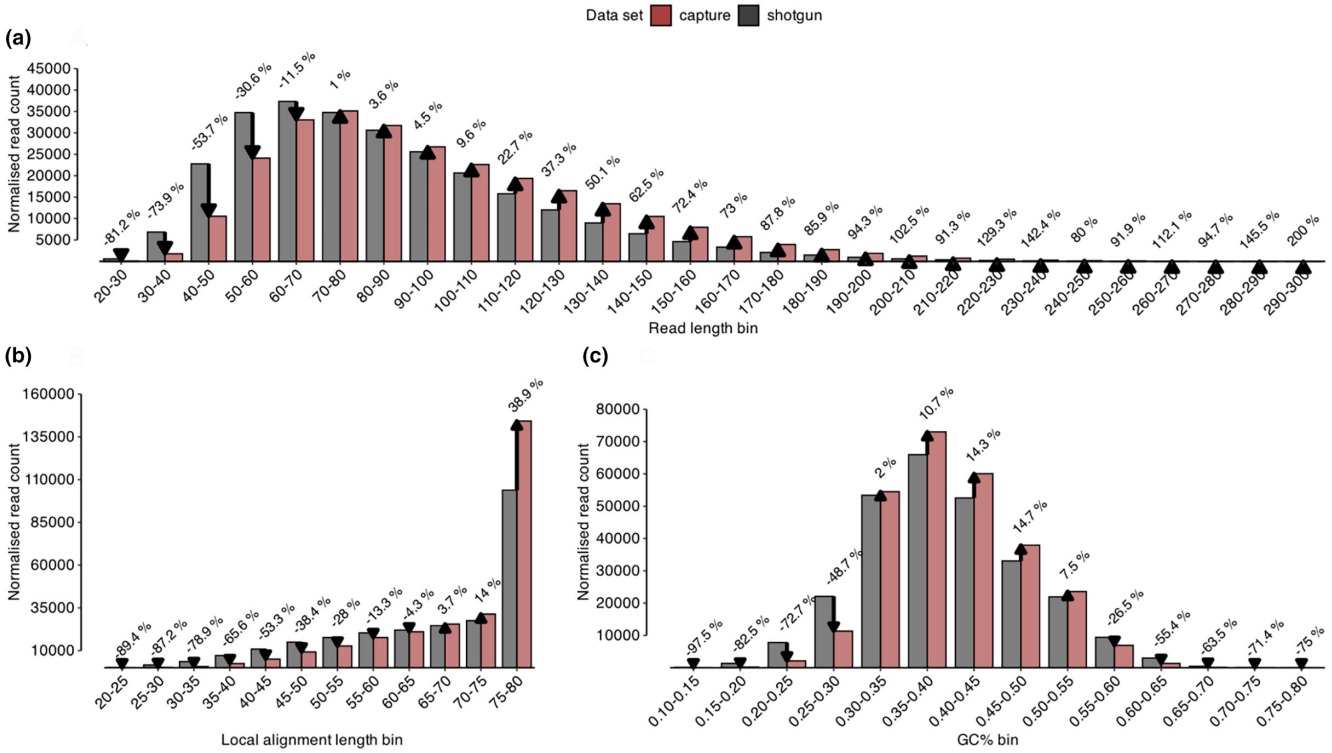

FIGURE 2 Absolute change of normalized read counts between the shotgun and the captured datasets. Arrows indicate the read counts after capture, with the values representing the relative change of reads within each bin. In the top panel, (a) the read length distribution is presented in 10 nucleotide bins. Short reads (<70 nucleotides) are under-represented post-capture, whFile reads with a length between >110 and <190 nucleotides are over-represented. Panel (b) shows the change in read length of the fragment mapping to a bait sequence. Following capture, reads below 65 nucleotides were under-represented, while reads mapped to 60–80 nucleotides were over-represented. Panel (c) shows the change observed in GC-content. Following capture, the proportion of reads with a GC-content between 35% and 55% increased, while reads with GC-content below 30% and above 55% decreased.

more reads were observed between 35% and 55% GC categories post-capture, while 55%–80% GC targets were under-represented.

3.3 | Effect of capture on relative proportion

We found that hybridization capture had an impact on the composition of DNA templates suitable for sequencing. Hence, next we assessed whether this change in composition could also affect the relative proportions of the different taxa represented in the mock communities. In the shotgun libraries, 55.8±0.2% of the reads were assigned to the *rbc*L target. However, after capture, the reads assigned to *mat*K increased by 11% and became more abundant (52.7±0.7; Figure 3b). On a taxonomy level, 89.0±0.1% of the *mat*K shotgun reads were assigned to species, compared to 61.3±0.2% for *rbc*L. After capture, the proportion of *rbc*L reads assigned to species remained identical to the shotgun library, while this fraction increased by 4.5% to 93.5±0.2% for *mat*K. Most of the *rbc*L reads that were not assigned to species could be assigned to the order level, representing 23.9±0.2% and 25.8±0.4% of shotgun and post-capture reads, respectively.

The relative taxonomic abundances at the lowest possible taxonomic level (for taxa with abundance >0.005) were highly similar between our four independent experimental replicates (Figure S8). We, thus, used the average of the four replicates for downstream analysis. Relative taxonomic abundances showed a moderate, yet significant, correlation between shotgun and post-capture libraries ($R=0.55$, p-value$=0.0052$; Figure 4a). Investigating the *mat*K and *rbc*L reads independently, we found a weak, non-significant correlation for the reads assigned to *mat*K ($R=0.36$, p-value$=0.28$; Figure 4b), but a moderate and significant positive correlation ($R=0.64$, p-value$=0.013$) was found when the low abundant taxa were included (Figure S9). The relative taxonomic abundances assessed using *rbc*L reads showed a strong, positive correlation between shotgun and reads post-capture ($R=0.83$, p-value <0.01, Figure 4c).

We next examined potential variables influencing apparent taxa distribution patterns in shotgun and captured libraries, including mean GC-content of the reads assigned to taxa the shotgun library (GC), the average number of baits overlapping a given read in shotgun libraries (ANB), median number of mismatches between overlapping read in shotgun libraries (MMS), relative taxonomic proportion in shotgun libraries (RPS), and the total length of the captured markers per species (TL; the length for other taxonomic levels was set to 0). Multiple linear regression analysis highlights that ANP, MMS, and TL were significant predictors for the change in relative abundance post-capture, with TL having the highest relative importance of 68.0% followed by MMS (20.1%) and ANB (11.9%). However, these

variables together explained only 50.58% of the observed variation ($F = 6.886$, 4 and 19 DF, p-value < 0.001; see Table 3A). When investigating the markers independently, the best fitting model for *mat*K showed that all independent variables were significant (Table 3B), explaining 94.5% of the change in relative abundances ($F = 17.46$, 5 and 5 DF, p-value = 0.003495). The variables ANB, GC, and TL were positively correlated, while MMS and RPS showed a negative effect on the change in relative read counts. The relative importance of TL and GC where the highest with 40.5% and 29.9% respectively. The three remaining variables were all nearly equally important with 11.4% (MMS), 9.4% (ANB), and 8.7% (RPS). The linear regression for the *rbc*L marker showed that 94.8% ($F = 56.1$, 4 and 8 DF, p-value < 0.001) of the change could be explained by ANB with a relative importance of 56.4% and GC with a relative importance 43.6% (Table 3C).

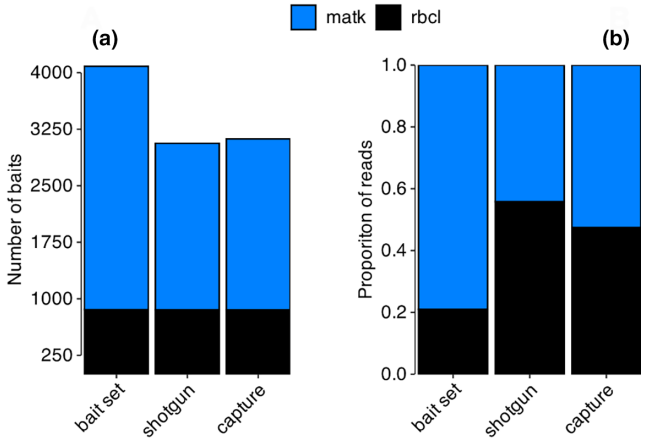

FIGURE 3 Panel (a) shows the absolute number of baits in the baits set, and divided between *mat*K and *rbc*L. The shotgun and capture columns show the total number of baits mapping the obtained reads. All *rbc*L baits mapped successfully, while 68.4% and 70.0% of the *mat*K baits mapped in shotgun and post-capture, respectively. Panel (b) shows the proportion of reads in the shotgun and post-capture that were assigned to the two different markers. The bait set column shows the relative proportion of the baits within the design. In the shotgun library, 55.8% of the reads were assigned to *rbc*L, while after capture, slightly more than half of the reads (52.7%) were assigned to *mat*K.

4 | DISCUSSION

Several bioinformatic tools have been developed for target capture bait design, which cover multiple applications. For example, *Baitfisher* helps find and filter bait sequences in multiple sequence alignments (MSA) to maximize bait design in variable regions (Mayer et al., 2016), while the recently developed *HUBDesign* delivers a unique bait set from multiple annotated genomes (Dickson et al., 2021). Additionally, *Syotti* is designed to identify bait sequences equally distant to each other (Alanko et al., 2022) and *SupeRbaits* provides bait candidate sets for population genetics studies (Jiménez-Mena et al., 2022). Some of these tools show some commonality to the approach presented in this study. For example, *HUBDesign* makes use of gene trees and reduces candidate gene sequences to a lowest common ancestor based on sequence identity (Dickson et al., 2021). However, to the best of our knowledge, none of the tools presently available are specifically tasked for designing oligonucleotide baits capable of characterizing the taxonomic composition of an environmental sample, besides clustering sequences or bait sequences at a certain threshold (Foster et al., 2021; Murchie, Kuch, et al., 2021). Our approach was designed to fill this gap, and is made available as a tool through the user-friendly, open-source python script "gotcha" available from GitHub (https://github.com/Kevinnota/gotcha).

We showed that using only sequence data from 2050 taxa (66.1% and 44.6% of all unique *rbc*L and *mat*k sequences, respectively) in the bait design was enough to map 99.9% of simulated reads coming from all 7949 taxa. The size distribution of the unmapped reads was significantly shorter than the distribution of simulated reads, which indicates that short reads, likely from diverged taxa not used

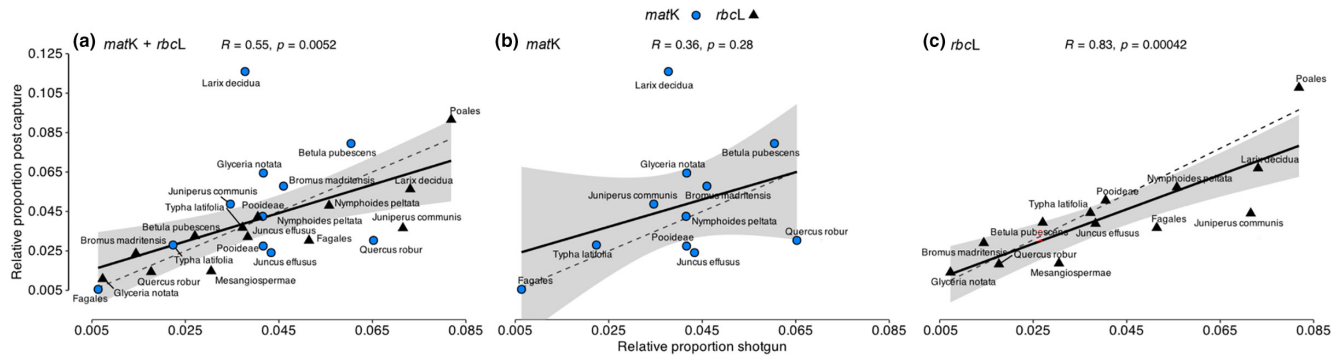

FIGURE 4 Relative proportions from the mock communities from shotgun sequencing libraries and post-capture (mean over four replicates). All taxa in the mock communities we detected in both shotgun and captured libraries. The dashed line indicates no change. Panel (a) includes all taxonomic levels (>0.005%) from both markers and shows a moderate positive linear correlation. Panels (b) and (c) show the same correlation but for matK and rbcL markers separately. The *mat*K (b) assignments are no longer significantly correlated with the reads prior to capture, although driven by outlier taxa *Larix decidua*, which almost doubles the relative abundance post-capture. The rbcL (c) detections are strongly positively correlated.

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences
WILEY

TABLE 3A Linear regression results for both markers.

Residuals:	Min	1Q	Median	3Q		Max
	−0.024967	−0.012274	0.002374	0.009362		0.032493
Coefficients	**Estimate**	**SE**	***t* value**	**Pr (>|*t*|)**	**Signif.**	**RI - lmg**
Intercept	3.40E-03	2.12E-02	0.161	0.874047		
Mean_baits_shotgun (ANB)	0.874047	1.86E-04	2.223	0.038574	*	11.9%
Median_NM_shotgun (MMS)	−1.47E-02	5.48E+00	−2.678	0.014872	*	20.1%
Relative_proportion_shotgun (RPS)	−2.52E-01	1.66E-01	−1.513	0.146629		–
Capture_length (TL)	4.H-05	9.73E-06	4.464	0.000266	***	68.0%

Note: Residual standard error: 0.01585 on 19 degrees of freedom. Multiple *R*-squared: 0.5918, Adjusted *R*-squared: 0.5058. *F*-statistic: 6.886 on 4 and 19 DF, *p*-value: 0.001332. Significance: ***$p < 0.001$, **$0.001 \le p < 0.01$, and *$0.01 \le p < 0.05$.

TABLE 3B Linear regression results for *mat*K.

Residuals:	Min	1Q	Median	3Q		Max
	−0.0109222	−0.0055359	−0.0015226	0.0021567		0.0128313
Coefficients	**Estimate**	**SE**	***t* value**	**Pr(>|*t*|)**	**Signif.**	**RI - lmg**
Intercept	−9.64E-02	3.51E-02	3.51E-02	0.04056	*	
Mean_baits_shotgun (ANB)	1.03E-03	2.38E-04	4.32	0.00757	**	9.4%
Mean_gc_shotgun (GC)	1.03E-03	1.07E-03	3.463	0.01798	*	29.9%
Median_NM_shotgun (MMS)	−2.40E-02	5.50E-03	−4.365	0.00726	**	11.4%
Relative_proportion_shotgun (RPS)	−7.22E-01	1.97E-01	−3.663	0.01455	*	8.7%
Capture_length (TL)	5.42E-05	8.57E-06	6.32	0.00146	**	40.5%

Note: Residual standard error: 0.009611 on 5 degrees of freedom. Multiple *R*-squared: 0.9458, Adjusted *R*-squared: 0.8917. *F*-statistic: 17.46 on 5 and 5 DF, *p*-value: 0.003495. Significance: ***$p < 0.001$, **$0.001 \le p < 0.01$, and *$0.01 \le p < 0.05$.

TABLE 3C Linear regression results for *rbc*L.

Residuals:	Min	1Q	Median	3Q		Max
	−0.0031958	−0.0014235	−0.0003324	0.0012415		0.0060094
Coefficients	**Estimate**	**SE**	***t* value**	**Pr(>|*t*|)**	**Signif.**	**RI lmg**
Intercept	−0.1724822	0.0182924	−9.429	1.31E-05	***	
Mean_baits_shotgun (ANB)	0.0005478	0.0000637	8.599	2.59E-05	***	56.4%
Mean_gc_shotgun (GC)	0.0033783	0.0003354	10.071	8.05E-06	***	43.6%
Median_NM_shotgun (MMS)	−0.0026946	0.0016655	−1.618	0.1443		–
Relative_proportion_shotgun (RPS)	−0.078938	0.0412027	−1.916	0.0917		–

Note: Residual standard error: 0.002995 on 8 degrees of freedom. Multiple *R*-squared: 0.9656, Adjusted R-squared: 0.9484. *F*-statistic: 56.09 on 4 and 8 DF, *p*-value: 6.831e-06. Significance: ***$p < 0.001$, **$0.001 \le p < 0.01$, and *$0.01 \le p < 0.05$.

in the design, were not enough to overlap a bait. The low number of reads that did not match a bait confirms that our strategy building on ancestral sequence reconstruction to design hybridization baits not only allows for efficiently capturing the molecular diversity in the gene tree but also sequences which were not represented in the tree. Gene trees for bait design, therefore, do not need to be fully comprehensive in terms of taxonomic diversity, as long as they include highly divergent species important for the study area. This might make bait sets designed for one geographic area be highly suitable for other regions (see Figure S3A–G for the gene trees and the selected nodes for bait design).

Testing the designed bait set on mock communities, we observed a reduction in reads not mapping to capture targets from 11.7% to 0.4% for *mat*K. This decrease in off-target reads confirms our probe set's effectiveness in capturing the desired fragments. Although, our mock communities were produced by shredding PCR product from the target genes, we expect, based on results from, Murchie, Kuch, et al. (2021) and Murchie, Monteath, et al. (2021), who also targeted chloroplast barcode regions that this observation will hold even in cases where the number of off-target molecules is much higher than in our mock samples. The reads obtained in the mock communities showed a clear enrichment for longer reads (>110

nucleotides) and a depletion of shorter reads (30–70 nucleotides). This phenomenon has been recorded in many studies and is likely a direct effect of binding potential between the bait and the library molecule, with longer templates stabilizing the hybridization (Cruz-Dávalos et al., 2017; Suchan, Chauvey, et al., 2022; Suchan, Kusliy, et al., 2022). We further observed a depletion of DNA templates with low GC-content post-capture (<30%) which is also in line with previous studies reporting enhanced capture efficiency between 35% and 60% GC-content (e.g., in exome capture experiments; Chilamakuri et al., 2014). By comparing the alignment length between the bait and the reads obtained in shotgun sequencing and post-captured libraries, we found that under the stringency conditions assessed in this study, the stability of bait-template annealing is enhanced when the alignment involves at least 50 bases. It is noteworthy that the 3× tiling used for designing baits resulted in 96.35% of simulated data aligning baits over at least 50 bases.

For ecological interpretation of capture results, it is important to assess the changes in taxonomic composition post-capture. We observed little change in assignment of reads at different ranks between the shotgun and captured dataset (e.g., similar number of reads were assigned to species before and after capture). Only a slight elevation in order level assignments with the *rbc*L marker was observed, indicating a minor enrichment for conserved reads here. The high number of reads assigned to species level in this study is due to the confinement of the database for mapping, to only the taxa in the mock communities. The taxonomic resolution for the markers will depend on the species composition of the location of the sample at hand. A potential way to increase the taxonomic resolution is to exclude baits which occur in more than one genus, or add more markers that have higher taxonomic resolution in the targeted species.

We further show a moderate, yet significant, correlation between taxon abundances in shotgun and capture reads, indicating that reads obtained after capture are somewhat following an expected pattern. However, this pattern is marker specific, since in the *mat*K marker, a non-significant correlation was observed. The difference between markers highlights that, although captured with the same bait-set, both markers experienced read abundance distortions to a different degree. When investigating the variables that could explain the observed patterns, we found that different sets of variables influence each gene. The best linear model tested on both markers together could explain only ~50% of the variation. However, when linear regression models were applied to the markers independently, the tested variables explained nearly 95% of the variation. For *mat*K, all variables tested were significant, but the highest relative importance were CL (40.5%) and GC (29.9%). The importance of the CL likely also explains the increase of the proportion of *mat*K reads after capture, since this marker had a larger capture region for each taxon (see Table 1). The increase in *mat*K reads might have been even greater if this marker had a GC-content comparable to that of *rbc*L. Further MMS, ANB, and RPS together have a relative importance of ~29.6%. For the much less taxonomically complex marker *rbc*L, ANB and GC were sufficient to explain all variation observed. The captured part of the *rbc*L gene does not have size variation across taxa and is highly conserved. This suggests that the MMS is not affecting the capture process, but rather it is the ANB which is predominantly driving any observed change.

Capture methods typically tolerate fragments showing divergence up to 10%–13% (Mason et al., 2011; Paijmans et al., 2016; Peñalba et al., 2014). Nonetheless, there is a clear indication that taxa with fewer nucleotide differences, and a higher number of baits mapping tend to be more efficiently captured. This implies that any attempt to decrease the number of baits will introduce bias into the capture process, which varies depending on the taxon. These findings collectively show that the bait design fundamentally introduces biases in reads abundance. While these biases might be minor in the case of *rbc*L, they can become significant in complex markers such as *mat*K with low overall GC and high nucleotide diversity. We used our shotgun libraries to obtain all tested variables except CL. Among these variables, only RPS is truly sample dependent, as relative abundance calculations require shotgun sequencing and are therefore unique to each sample. The relative importance of RPS was only minor with 8.7% in our mock communities. Yet, it is unclear whether this relative importance will remain low in other mixtures. To test this, only mock communities with variable taxon abundances and taxon composition are suitable. The remaining variables, such as MMS, ANB, and GC, can be calculated for each taxon using simulated data, using the read length distributions of shotgun libraries of the samples.

5 | CONCLUDING REMARKS

We successfully developed a universal set of baits for four plant orders with a set of small number of baits. Based on simulated reads, these baits are capable of capturing sequence diversity beyond what was originally used for bait design. We further experimentally show that the produced probes successfully captured all the taxa present in the assembled mock communities, over a large GC-content range. The overall distortion of relative read-proportions after capture was minimal for *rbc*L but high for *mat*K. Our results strongly suggest that hybridization capture is notably influenced by a multitude of experimental factors, ranging from bait GC-content and taxonomic imbalance (i.e., different abundances of taxa in a sample).

These confounding variables ultimately influence the results, particularly concerning apparent relative abundances. Although biases are present, the effect of read abundance changes in our mock communities was predictable based on variables that can be easily calculated. This enables potential "in silico" investigation of bait-sets before synthesis.

Based on our observations, we recommend the following strategies: (1) Consider avoiding markers with low GC content or high GC content, or conversely exclude baits with low (>30%) or high GC (60%) content to mitigate GC biases; (2) aim for balanced baits design distributing an equal number of baits per taxa with an equal number of mismatches when estimating relative abundances; (3) limit interpretation of capture data to qualitative assessments (i.e., presence–absence of taxa rather than quantitative interpretations based on abundances). We also advocate to include mock community analyses

including a wide variety of taxa to enable correction of major biases introduced during capture.

While our experimental design revealed important factors driving biases influencing taxonomic abundances, further research is required to shed light on other aspects of the methodology. These include assessing the power of target capture approaches to detect rare species, their sensitivity relative to amplicon-based metabarcoding techniques, and the influence of annealing temperatures and incubation times on the capture efficacy and taxonomic profiles. Additionally, testing multiple bait design strategies and capturing pooled libraries are recommended. Finally, it is important to evaluate the possible benefits or potentially stronger capture biases deriving from a second round of capture, a common practice for low-template DNA samples (e.g., ancient DNA).

AUTHOR CONTRIBUTIONS

K.N., L.O., L.P., and C.V. conceived the study. K.N. designed the baits, with input from L.O. The wet-lab experiments were designed by K.N., L.P., C.V., and L.O., and the laboratory work was performed by M.G. The manuscript was drafted by K.N., with contributions from L.O., C.V., and L.P. All authors contributed to data interpretation and to revising the manuscript.

ACKNOWLEDGMENTS

Brian Brunelle from Arbor Biosciences for help with the bait design. Petra Vainio, Nordic Sales Manager at TATAA Biocenter for providing advice on the capture design. The baits were funded by the Swedish Phytogeographical Society. This project has received funding from the CNRS, University Paul Sabatier (AnimalFarm IRP) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreements 681605-PEGASUS and 101071707-Horsepower). The python implication of "gotcha" (https://github.com/Kevinnota/gotcha) was in collaboration with Giobbe Forni at the University of Bologna. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

No conflicts of Interest.

DATA AVAILABILITY STATEMENT

The raw sequencing data produced in this study is available at ENA with study accession PRJEB76475, and runs accessions ERR13252521-ERR13252528. The bait sequences will be available as a Supplementary File. The python script to design phylogenetic informed probes is available on GitHub (https://github.com/Kevinnota/gotcha). The script for parsing the botwie2 mapped reads used for analyzing and plotting are available on GitHub (https://github.com/Kevinnota/capture_mock_communities), and all data files required to produce the figures are accessible on figshare (https://doi.org/10.6084/m9.figshare.26044429).

ORCID

Kevin Nota https://orcid.org/0000-0002-4744-5205

REFERENCES

Ahmed, E., Parducci, L., Unneberg, P., Ågren, R., Schenk, F., Rattray, J. E., Han, L., Muschitiello, F., Pedersen, M. W., Smittenberg, R. H., Yamoah, K. A., Slotte, T., & Wohlfarth, B. (2018). Archaeal community changes in Lateglacial lake sediments: Evidence from ancient DNA. *Quaternary Science Reviews*, *181*, 19–29. https://doi.org/10.1016/j.quascirev.2017.11.037

Alanko, J. N., Slizovskiy, I. B., Lokshtanov, D., Gagie, T., Noyes, N. R., & Boucher, C. (2022). Syotti: Scalable bait design for DNA enrichment. *Bioinformatics (Oxford, England)*, *38*(Suppl 1), i177–i184. https://doi.org/10.1093/bioinformatics/btac226

Armbrecht, L., Hallegraeff, G., Bolch, C. J. S., Woodward, C., & Cooper, A. (2021). Hybridisation capture allows DNA damage analysis of ancient marine eukaryotes. *Scientific Reports*, *11*(1), 3220. https://doi.org/10.1038/s41598-021-82578-6

Bendall, M. L., Stevens, S. L., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S. G., Bertilsson, S., Moran, M. A., Shade, A., Newton, R. J., McMahon, K. D., & Malmstrom, R. R. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal*, *10*(7), 1589–1601. https://doi.org/10.1038/ismej.2015.241

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Chamberlain, S. (2021). *bold: Interface to Bold Systems API (1.2.0)* [Computer software]. https://CRAN.R-project.org/package=bold

Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., & Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, *15*(1), 449. https://doi.org/10.1186/1471-2164-15-449

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Scheu, A., Beneke, N., Ludwig, A., Cooper, A., Willerslev, E., & Orlando, L. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources*, *17*(3), 508–522. https://doi.org/10.1111/1755-0998.12595

Delsuc, F., Gibb, G. C., Kuch, M., Billet, G., Hautier, L., Southon, J., Rouillard, J.-M., Fernicola, J. C., Vizcaíno, S. F., MacPhee, R. D. E., & Poinar, H. N. (2016). The phylogenetic affinities of the extinct glyptodonts. *Current Biology*, *26*(4), R155–R156. https://doi.org/10.1016/j.cub.2016.01.039

Dickson, Z. W., Hackenberger, D., Kuch, M., Marzok, A., Banerjee, A., Rossi, L., Klowak, J. A., Fox-Robichaud, A., Mossmann, K., Miller, M. S., Surette, M. G., Golding, G. B., & Poinar, H. (2021). Probe design for simultaneous, targeted capture of diverse metagenomic targets. *Cell Reports Methods*, *1*(6), 100069. https://doi.org/10.1016/j.crmeth.2021.100069

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111–1119. https://doi.org/10.1111/2041-210X.12114

Fernandez-Guerra, A., Borrel, G., Delmont, T. O., Elberling, B., Eren, A. M., Gribaldo, S., Jochheim, A., Henriksen, R. A., Hinrichs, K.-U., Korneliussen, T. S., Krupovic, M., Larsen, N. K., Laso-Pérez, R., Pedersen, M. W., Pedersen, V. K., Sand, K. K., Sikora, M., Steinegger, M., Veseli, I., … Willerslev, E. (2023). A 2-million-year-old microbial and viral communities from the Kap København formation in North Greenland. *bioRxiv* https://doi.org/10.1101/2023.06.10.544454

Foster, N. R., Van Dijk, K., Biffin, E., Young, J. M., Thomson, V. A., Gillanders, B. M., Jones, A. R., & Waycott, M. (2021). A multi-gene

region targeted capture approach to detect plant DNA in environmental samples: A case study from coastal environments. *Frontiers in Ecology and Evolution*, *9*, 735744. https://doi.org/10.3389/fevo.2021.735744

Frémont, P., Gehlen, M., Vrac, M., Leconte, J., Delmont, T. O., Wincker, P., Iudicone, D., & Jaillon, O. (2022). Restructuring of plankton genomic biogeography in the surface ocean under climate change. *Nature Climate Change*, *12*(4), 393–401. https://doi.org/10.1038/s41558-022-01314-8

Gansauge, M.-T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using *T4* DNA ligase. *Nucleic Acids Research*, *45*, e79. https://doi.org/10.1093/nar/gkx033

Govaerts, R., Nic Lughadha, E., Black, N., Turner, R., & Paton, A. (2021). The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Scientific Data*, *8*(1), 215. https://doi.org/10.1038/s41597-021-00997-6

Graham, R. W., Belmecheri, S., Choy, K., Culleton, B. J., Davies, L. J., Froese, D., Heintzman, P. D., Hritz, C., Kapp, J. D., Newsom, L. A., Rawcliffe, R., Saulnier-Talbot, É., Shapiro, B., Wang, Y., Williams, J. W., & Wooller, M. J. (2016). Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. *Proceedings of the National Academy of Sciences*, *113*(33), 9310–9314. https://doi.org/10.1073/pnas.1604903113

Grömping, U. (2006). Relative importance for linear regression in *R*: The package **relaimpo**. *Journal of Statistical Software*, *17*(1), 1–27. https://doi.org/10.18637/jss.v017.i01

Heckenhauer, J., Barfuss, M. H. J., & Samuel, R. (2016). Universal multiplexable matK primers for DNA barcoding of angiosperms. *Applications in Plant Sciences*, *4*(6), 1500137. https://doi.org/10.3732/apps.1500137

Jiménez-Mena, B., Flávio, H., Henriques, R., Manuzzi, A., Ramos, M., Meldrup, D., Edson, J., Pálsson, S., Ásta Ólafsdóttir, G., Ovenden, J. R., & Nielsen, E. E. (2022). Fishing for DNA? Designing baits for population genetics in target enrichment experiments: Guidelines, considerations and the new tool supeRbaits. *Molecular Ecology Resources*, *22*(5), 2105–2119. https://doi.org/10.1111/1755-0998.13598

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kjær, K. H., Winther Pedersen, M., De Sanctis, B., De Cahsan, B., Korneliussen, T. S., Michelsen, C. S., Sand, K. K., Jelavić, S., Ruter, A. H., Schmidt, A. M. A., Kjeldsen, K. K., Tesakov, A. S., Snowball, I., Gosse, J. C., Alsos, I. G., Wang, Y., Dockter, C., Rasmussen, M., Jørgensen, M. E., … Willerslev, E. (2022). A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature*, *612*(7939), 283–291. https://doi.org/10.1038/s41586-022-05453-y

Kress, W. J., & Erickson, D. L. (2007). A two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS One*, *2*(6), e508. https://doi.org/10.1371/journal.pone.0000508

Kusumi, J., Tsumura, Y., Yoshimaru, H., & Tachida, H. (2000). Phylogenetic relationships in Taxodiaceae and Cupressaceae sensu stricto based on matK gene, chlL gene, trnL-trnF IGS region, and trnL intron sequences. *American Journal of Botany*, *87*(10), 1480–1488. https://doi.org/10.2307/2656874

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lentz, D. L., Hamilton, T. L., Dunning, N. P., Tepe, E. J., Scarborough, V. L., Meyers, S. A., Grazioso, L., & Weiss, A. A. (2021). Environmental DNA reveals arboreal cityscapes at the ancient Maya Center of Tikal. *Scientific Reports*, *11*(1), 12725. https://doi.org/10.1038/s41598-021-91620-6

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Mason, V. C., Li, G., Helgen, K. M., & Murphy, W. J. (2011). Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research*, *21*(10), 1695–1704. https://doi.org/10.1101/gr.120196.111

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A software package for multispecies target DNA enrichment probe design. *Molecular Biology and Evolution*, *33*(7), 1875–1886. https://doi.org/10.1093/molbev/msw056

Meusnier, I., Singer, G. A., Landry, J.-F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, *9*(1), 214. https://doi.org/10.1186/1471-2164-9-214

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Mthethwa-Hlongwa, N. P., Amoah, I. D., Gomez, A., Davison, S., Reddy, P., Bux, F., & Kumari, S. (2024). Profiling pathogenic protozoan and their functional pathways in wastewater using 18S rRNA and shotgun metagenomics. *Science of the Total Environment*, *912*, 169602. https://doi.org/10.1016/j.scitotenv.2023.169602

Murchie, T. J., Kuch, M., Duggan, A. T., Ledger, M. L., Roche, K., Klunk, J., Karpinski, E., Hackenberger, D., Sadoway, T., MacPhee, R., Froese, D., & Poinar, H. (2021). Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. *Quaternary Research*, *99*, 305–328. https://doi.org/10.1017/qua.2020.59

Murchie, T. J., Monteath, A. J., Mahony, M. E., Long, G. S., Cocker, S., Sadoway, T., Karpinski, E., Zazula, G., MacPhee, R. D. E., Froese, D., & Poinar, H. N. (2021). Collapse of the mammoth-steppe in central Yukon as revealed by ancient environmental DNA. *Nature Communications*, *12*(1), 7120. https://doi.org/10.1038/s41467-021-27439-6

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E., & Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, *18*(5), 927–939. https://doi.org/10.1111/1755-0998.12895

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., & Förster, D. W. (2016). Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources*, *16*(1), 42–55. https://doi.org/10.1111/1755-0998.12420

Parducci, L., Alsos, I. G., Unneberg, P., Pedersen, M. W., Han, L., Lammers, Y., Salonen, J. S., Väliranta, M. M., Slotte, T., & Wohlfarth, B. (2019). Shotgun environmental DNA, pollen, and macrofossil analysis of Lateglacial Lake sediments from southern Sweden. *Frontiers in Ecology and Evolution*, *7*, 15. https://doi.org/10.3389/fevo.2019.00189

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., Mendoza, M. L. Z., Beaudoin, A. B., Zutter, C., Larsen, N. K.,

Potter, B. A., Nielsen, R., Rainville, R. A., Orlando, L., Meltzer, D. J., Kjær, K. H., & Willerslev, E. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature*, *537*(7618), 45–49. https://doi.org/10.1038/nature19085

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A., Bowie, R. C. K., & Moritz, C. (2014). Sequence capture using PCR -generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources*, *14*(5), 1000–1010. https://doi.org/10.1111/1755-0998.12249

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. https://doi.org/10.1093/bioinformatics/bti079

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. https://www.r-project.org/

Renaud, G., Hanghøj, K., Willerslev, E., & Orlando, L. (2017). Gargammel: A sequence simulator for ancient DNA. *Bioinformatics*, *33*(4), 577–579. https://doi.org/10.1093/bioinformatics/btw670

Richter, D. J., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G., Maillet, N., Henry, N., Benoit, G., Da Silva, O., Delmont, T. O., Fernàndez-Guerra, A., Suweis, S., Narci, R., Berney, C., Eveillard, D., Gavory, F., Guidi, L., Labadie, K., … Jaillon, O. (2022). Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *eLife*, *11*, e78129. https://doi.org/10.7554/eLife.78129

Rodriguez-Martinez, S., Klaminder, J., Morlock, M. A., Dalén, L., & Huang, D. Y.-T. (2022). The topological nature of tag jumping in environmental DNA metabarcoding studies. *Molecular Ecology Resources*, *23*, 621–631. https://doi.org/10.1111/1755-0998.13745

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Schulte, L., Meucci, S., Stoof-Leichsenring, K. R., Heitkam, T., Schmidt, N., von Hippel, B., Andreev, A. A., Diekmann, B., Biskaborn, B. K., Wagner, B., Melles, M., Pestryakova, L. A., Alsos, I. G., Clarke, C., Krutovsky, K. V., & Herzschuh, U. (2022). Larix species range dynamics in Siberia since the last glacial captured from sedimentary ancient DNA. *Communications Biology*, *5*(1), 570. https://doi.org/10.1038/s42003-022-03455-0

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Slon, V., Glocke, I., Barkai, R., Gopher, A., Hershkovitz, I., & Meyer, M. (2016). Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiagnostic remains, and its application to middle Pleistocene specimens from Qesem cave (Israel). *Quaternary International*, *398*, 210–218. https://doi.org/10.1016/j.quaint.2015.03.039

Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., Stewart, J. R., Derevianko, A. P., Jacobs, Z., Li, B., Roberts, R. G., Shunkov, M. V., de Lumley, H., Perrenoud, C., Gušić, I., … Meyer, M. (2017). Neandertal and denisovan DNA from pleistocene sediments. *Science*, *356*(6338), 605–608. https://doi.org/10.1126/science.aam9695

Suchan, T., Chauvey, L., Poullet, M., Tonasso-Calvière, L., Schiavinato, S., Clavel, P., Clavel, B., Lepetz, S., Seguin-Orlando, A., & Orlando, L. (2022). Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA. *Molecular Ecology Resources*, *22*(6), 2262–2274. https://doi.org/10.1111/1755-0998.13619

Suchan, T., Kusliy, M. A., Khan, N., Chauvey, L., Tonasso-Calvière, L., Schiavinato, S., Southon, J., Keller, M., Kitagawa, K., Krause, J., Bessudnov, A. N., Bessudnov, A. A., Graphodatsky, A. S., Valenzuela-Lamas, S., Wilczyński, J., Pospuła, S., Tunia, K., Nowak, M., Moskal-delHoyo, M., … Orlando, L. (2022). Performance and automation of ancient DNA capture with RNA hyRAD probes. *Molecular Ecology Resources*, *22*(3), 891–907. https://doi.org/10.1111/1755-0998.13518

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring* (Illustrated ed.). Oxford University Press.

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Vernot, B., Zavala, E. I., Gómez-Olivencia, A., Jacobs, Z., Slon, V., Mafessoni, F., Romagné, F., Pearson, A., Petr, M., Sala, N., Pablos, A., Aranburu, A., de Castro, J. M. B., Carbonell, E., Li, B., Krajcarz, M. T., Krivoshapkin, A. I., Kolobova, K. A., Kozlikin, M. B., … Meyer, M. (2021). Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. *Science*, *372*(6542), eabf1667. https://doi.org/10.1126/science.abf1667

Wang, Y., Korneliussen, T. S., Holman, L. E., Manica, A., & Pedersen, M. W. (2022). ngsLCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data. *Methods in Ecology and Evolution*, *13*(12), 2699–2708. https://doi.org/10.1111/2041-210X.14006

Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., … Willerslev, E. (2021). Late quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, *600*, 86–92. https://doi.org/10.1038/s41586-021-04016-x

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis [computer software]*. Springer.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina paired-end reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

SUPPORTING INFORMATION

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

How to cite this article: Nota, K., Orlando, L., Marchesini, A., Girardi, M., Bertilsson, S., Vernesi, C., & Parducci, L. (2024). Enriching barcoding markers in environmental samples utilizing a phylogenetic probe design: Insights from mock communities. *Environmental DNA*, *6*, e593. https://doi.org/10.1002/edn3.593