# Mitigating PCR Bias in Environmental Zooplankton Metabarcoding Samples

Dante Capone1, Hannah Budroe1, Ramon Gallego2, Stephanie Matthews1, Michelle Nixon3, Jeff Bowman1, Justin Silverman3, Moira Décima1

1. Scripps Institution of Oceanography, 2. Universidad Autónoma de Madrid - Unidad de Genética, 3. Pennsylvania State University

Keyword: Marine zooplankton, DNA metabarcoding, PCR bias, California Current Ecosystem

## Abstract

Zooplankton play a central role in marine pelagic ecosystems through trophic transfer of primary productivity, elemental cycling, and sensitivity to changes in oceanographic conditions. Within the California Current Ecosystem, zooplankton communities covary with a strong cross-shore gradient in productivity and other biogeochemical properties driven by upwelling. Along this gradient, zooplankton communities vary in size-structure and taxonomic diversity. While the emerging use of DNA metabarcoding facilitates measurements of this diversity, methodological challenges such as PCR amplification limit quantitative interpretability. Here, we use DNA metabarcoding of the COI and 18S regions for size-fractionated, net-collected zooplankton samples in conjunction with Zooscan imaging to examine mesozooplankton diversity and community structure related to changes in environmental conditions. Furthermore, we conduct a calibration experiment using pooled DNA comprised of all samples and amplified at varying PCR cycles to measure the amplification efficiencies of key taxa and correct for PCR bias. We find our environmental variability dominated by a cross-shore gradient, which correlates with zooplankton diversity. Diversity is highest offshore for the community and smaller size classes and lower offshore for the largest size class. All samples were dominated by calanoid copepods, however higher taxonomic resolution of metabarcoding allows us to link more specific taxon with environmental conditions. Experimentally derived amplification efficiencies vary with size and relative abundance of key taxa. We demonstrate that for dominant groups like calanoid copepods, PCR bias-correction can enhance the agreement between metabarcoding measures and quantitative Zooscan estimates of abundance. Our results suggest that size-fractionation and constructing sample pools based on similar expected composition could further enhance the efficacy of this method.

## Introduction

DNA metabarcoding has emerged as a transformative tool in biodiversity surveys, offering an efficient, high-resolution approach for characterizing biological communities across a range of environments and taxonomic groups. The process extracts DNA from environmental samples—such as tissue or water—and uses high-throughput sequencing to identify taxonomic groups based on specific genetic markers. Barcodes like the mitochondrial cytochrome c oxidase I (COI) and regions of the nuclear small subunit 18S ribosomal RNA are commonly targeted to differentiate between eukaryotic species. By comparing sequenced barcodes against reference databases, researchers can determine the composition of the community present in a sample (Leray et al., 2013; Zhan et al., 2013; Bucklin et al., 2016). Increasing adoption of molecular methods in ecological monitoring reflects their advantages over traditional morphology-based identification, including reduced labor intensity, the ability to identify cryptic species, and suitability for high-biodiversity systems.

In marine studies, DNA metabarcoding has become particularly valuable for investigating zooplankton, which play critical roles in marine food webs and biogeochemical cycles (Bucklin et al., 2016). Zooplankton community composition is a powerful indicator of ecosystem state, responding to both short-term variability and long-term environmental changes (Taylor, Allen, and Clark, 2002). The large-scale taxonomic identification enabled by metabarcoding, combined with decreasing sequencing costs, allows researchers to investigate patterns of zooplankton diversity across spatial and temporal gradients (Matthews and Ohman, 2023). For example, in the California Current Ecosystem (CCE), metabarcoding has provided detailed insights into relationships between zooplankton communities and environmental conditions, expanding upon earlier ecological understandings derived from net tows and morphological analyses (Matthews and Ohman, 2023; Pitz et al., 2020).

Despite its promise, metabarcoding for assessing marine biodiversity is subject to several methodological limitations that can affect quantitative interpretations (Bucklin et al., 2016; Leray et al., 2013). Various sources of bias impact the reliability of metabarcoding results in re-creating quantitative metrics like biomass. DNA extraction methods and primer bias-where certain primers preferentially amplify specific taxa-can distort the perceived community structure. Sequencing depth and library preparation can also introduce biases, as the choice of sequencing depth affects the detection of rare taxa, and random sampling during library preparation can lead to inconsistent community representation. Additionally, gene copy number variation in eukaryotes can significantly affect biomass estimates; many eukaryotes possess variable numbers of ribosomal RNA gene copies, which means that read counts may not directly reflect the true abundance or biomass of different species. This complicates efforts to relate DNA read counts to quantitative estimates of organismal biomass or ecosystem functions.

Paramount among these biases are those introduced by the polymerase chain reaction (PCR) amplification step, essential for amplifying barcode regions from environmental DNA but prone to differential efficiency among taxa (Suzuki and Giovannoni, 1996). These biases arise from primer mismatches and variations in DNA template abundance, potentially skewing community composition estimates, particularly at later PCR cycles (Shelton et al., 2023).

To address these challenges, researchers often use mock communities to estimate amplification biases in metabarcoding studies. Mock communities consist of DNA from a set of known species and provide a benchmark for assessing how well the amplification process represents community composition. While mock communities offer a reasonable approach to estimating amplification biases, they have significant drawbacks. Constructing an appropriate mock community for a given application can be difficult when there are large numbers of taxa of interest, or the source DNA from important taxa is unavailable, as is often the case in studies of microbial communities. When DNA from key species is unavailable or community composition unknown, creating a representative mock community may be impractical (Silverman et al., 2021).

In such cases, it is possible to modify the PCR protocol for technical replicates to bracket a range of PCR cycles and observe the changes in amplicons at each cycle (Figure 2d). Importantly, this variable PCR cycle calibration approach can be applied to samples of unknown composition.

In the present study, we implement a variable PCR cycle experiment to quantify amplification efficiencies and attempt to mitigate PCR bias in environmental zooplankton samples of unknown composition. By modifying the PCR protocol to use a range of cycle numbers, we observe changes in relative community composition across PCR cycles. To our knowledge, the only other implementation of this method is in its application to the human gut microbiome in the study in which the method was proposed (Silverman et al., 2021). Our study represents the first time that this approach is being used for marine zooplankton samples, highlighting its novel application in this context. By using regression analysis, we estimate the relative amplification rates (α) of different taxa, which, combined with sequence proportions, yields estimates of their initial abundances (β). Although this approach has been employed successfully in some cases with extensive sampling (Silverman et al., 2021), its practical application remains challenging without sufficient replication.

Our goal is to use this method to derive taxon-specific amplification efficiencies, which can be used to correct read abundances and provide a representation of zooplankton community composition that more accurately reflects quantitative metrics such as biomass. To assess whether the PCR bias correction improves upon raw read counts, we compare estimates against the widely adopted Zooscan imaging, which provides based morphologically-derived biomass.

## Materials and Methods

### Oceanographic Zooplankton Sampling/Processing

Zooplankton net tow samples were collected during three quasi-Lagrangian experiments termed ‘cycles’ between July 21st and August 7th, 2021, during the California Current Ecosystem Long-Term Ecological Research (CCE-LTER) process cruise P2107. A Bongo net (0.71-m diameter frame, 202-µm mesh nets and cod ends) was towed obliquely between the surface and ~200m depth (300m wire out) with a wire angle of 45° at a ship speed of 1-2 knots and was recovered at 30 m min-1. Sampling depths were logged with a temperature-depth recorder and volume filtered through the net was measured using a calibrated General Oceanics Flow meter. The starboard side of the Bongo net was split quantitatively, and 3/8ths of the total sample was size-fractionated using a set of sieves (mesh sizes 5mm, 2mm, 1mm, 0.5mm, 0.2mm), was rinsed using filtered seawater onto a 47 mm-diameter, 0.2mm Nitex mesh filter, was placed in a petri and was flash frozen for metabarcoding or other analyses on land (Figure 1). The port side of the Bongo net was preserved in formalin in its entirety and used in subsequent Zooscan imaging analysis.

Cycle 1 and Cycle 2 both tracked cool, salty, dense upwelled water off Point Sur. Cycle 3 was initiated about 800 km offshore near the western boundary of the CCE and provided an ecological and biogeochemical endmember. Additionally, P2107 conducted 2 transects between August 7th and August 11th, 2021. First, an east-west transect spanning the CCE (referred to as CT1) and an alongshore transect (CT2) which included 8 stations extending north-south from San Simeon to the Point Arguello. CT1 spanned a vast range of environmental conditions whereas CT2 was conducted nearshore. Excluding CT1, we selected a subset of the stations sampled within each cycle to examine metabarcoding community composition. For C1-C3 we sequenced a pair of day-night tows and for CT2 we selected 3 of the 8 stations.

### DNA Metabarcoding, Bioinformatics, PCR Calibration Experiment

We used two of the most common DNA barcodes for marine zooplankton: the hypervariable V4 region of the 18S ribosomal RNA marker (Zhan et al. 2013) and the mitochondrial cytochrome c oxidase subunit I gene (COI) (Leray et al. 2013) to resolve zooplankton community composition. The 3 smallest size fractions of zooplankton samples collected above (termed ‘small’, ‘medium’ and ’large’) were subsampled to ⅛ of filter, the tissue was homogenized, and DNA was extracted using the OMEGA E.Z.N.A. Insect Kit or Mollusc Kit following manufacturer protocols, with the specific kit being selected based on visual inspection of zooplankton community composition. Elutions were conducted in duplicate, and DNA concentrations were assessed by Qubit. For each sample, the barcode regions were amplified in triplicate using a 2-step PCR. Amplified products were pooled equimolarly and were sequenced on an Illumina MiSeq with 300bp PE VX chemistry.

Taxonomy was assigned to ASVs for both the COI and 18S marker, using the MetaZooGene database augmented with sequences obtained from local CCE zooplankton voucher specimens, to maximize identification of the zooplankton taxa targeted with the metazoan primers. MetaZooGene global databases and taxonomy files were downloaded on 3 May 2022. Sequences with mismatches between NCBI taxonomic identities and mothur taxonomic identities were removed, and the remaining at each marker were trimmed to the targeting amplicon region using virtual PCR from the insect R package, then were supplemented with sequences obtained from sanger sequencing of positively identified voucher specimens. The combined MetaZooGene and voucher sequences for each marker were used to train amplicon-specific classifiers in QIIME2, with the default parameters of the sk-learn algorithm. ASVs were assigned taxonomy, keeping all classifications with >80% confidence.

A diagram of a dna

Description automatically generated

Figure 1. Schematic of the sequence of data collection and processing steps for our DNA metabarcoding of marine zooplankton. The process follows: plankton net tow, quantitative splitting, Zooscan analysis, size fractionation, sub-sampling and DNA extraction, primer selection, PCR amplification and calibration experiment, sequencing, and taxonomic assignments.

In addition to this more widely implemented metabarcoding pipeline, we included the variable PCR cycle experiment that attempts to mitigate PCR bias in our results. We followed the methodology proposed by Silverman et al. (2021) and combined the calibration experiment with Bayesian Multinomial Logistic-Normal Models using the *R* package *fido* (Figure 2). The log-ratio linear models implemented by *fido* can account for mixtures of more than two taxa with different amplification efficiencies and have been shown to fit PCR bias of microbiome data. These models incorporate the “competition to be counted” in which counting more of one type of thing means there are fewer resources available to count other types of things (Silverman *et al.*, 2022). Compared to other quantitative metabarcoding approaches such as mock communities, the variable PCR method does not depend on *a priori* knowledge of a sample’s community composition. Instead, we included a calibration experiment in which study samples are pooled into a “calibration sample.” This calibration sample ideally contains sequences from every organism in the study.

Next, the calibration sample was split into aliquots which were amplified with PCR at varying cycles. Each of these varying cycle samples was amplified, barcoded, and sequenced in accordance with protocols followed for all other environmental samples. Finally, log-linear models were used to fit the data and determine the PCR bias corrected DNA concentrations present in the sample at cycle 0.

Although this method addresses the significant bias introduced through PCR, it is limited in several ways. First, the approach of pooling all sub-samples of all samples may miss rare taxa and is only applicable to taxa present in all the calibration samples. Another challenge that must be overcome in working towards quantitative metabarcoding of marine metazoans is accounting for gene copy number. Previous work has found that different taxonomic groups possess different numbers of copies for gene markers (Prokopowich, Gregory and Crease, 2003; Wang *et al.*, 2017). Future work remains to determine taxon-specific gene copy numbers if metabarcoding methods hope to produce quantitative measures of abundance and biomass.

For this study, we created pooled samples and amplified them at 3 different PCR cycle numbers (20, 24, and 28) to calculate a taxon-specific amplification efficiency and initial, PCR bias-mitigated relative abundance. We aliquoted 10 µL of DNA into 4 different pooled samples, 3 pools contained different subsets of the total samples (determined by placement in the 96-well plate) and one pool that contained an aliquot from every sample in the study. We term these pools “All pool”,” Pool 1”,” Pool 2”, and “Pool3” (Table 5). Unfortunately, due to PCR failure for the 20-cycle amplification for the COI pools, the model fits for COI ASV’s were limited to 24 and 28 cycle pools in fitting the *fido* model. We use all 4 pools to examine taxon and size-specific amplification efficiencies and use the “All pool” for our methods comparison.

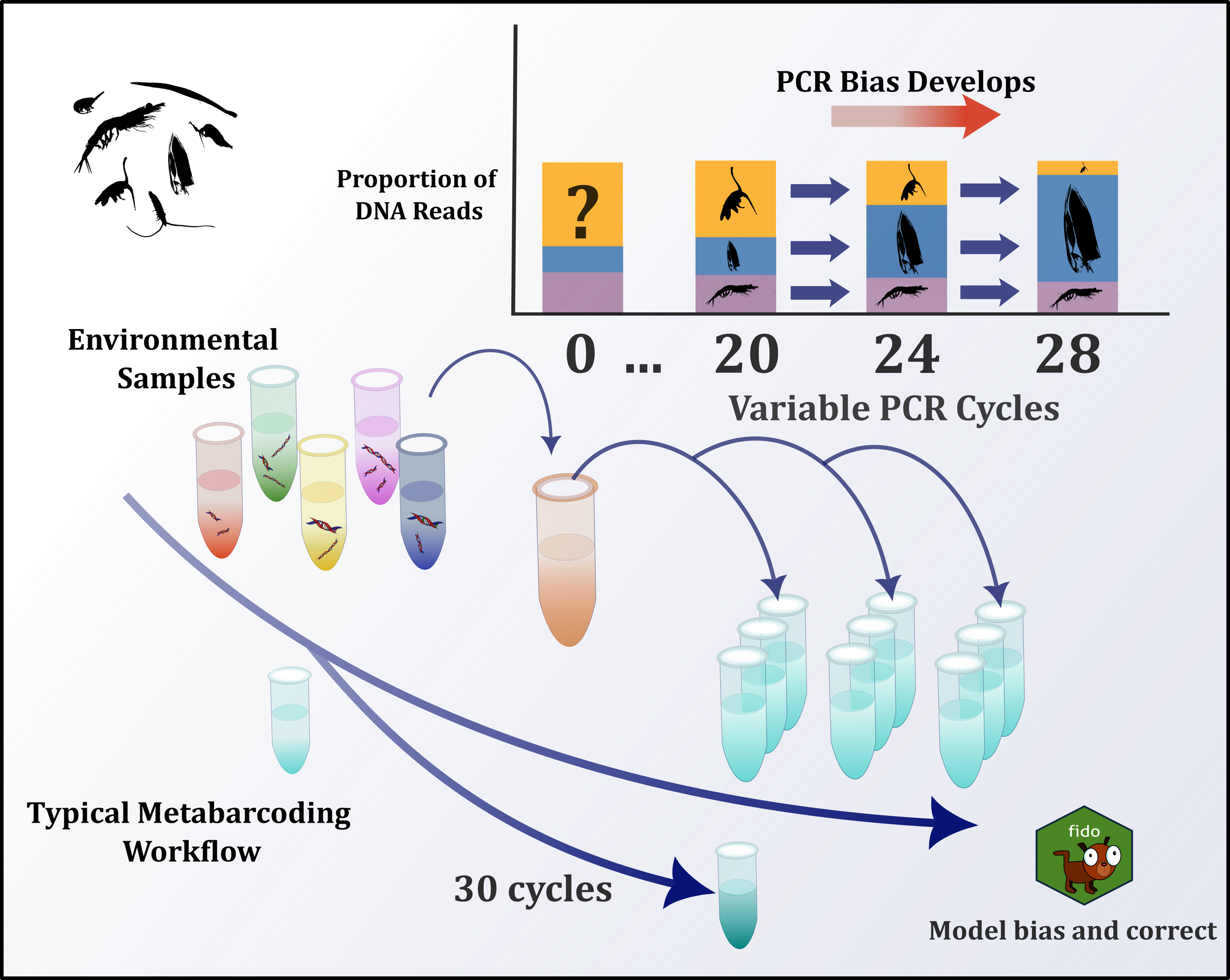


Figure 2. Variable PCR Experiment Workflow. DNA extracted from size-fractionated, net-collected mesozooplankton samples sub-sampled and pooled into a calibration sample. The calibration sample is aliquoted into technical replicates and amplified at variable PCR cycles-for this study 20, 24, and 28 cycles. The R package fido is used to model and mitigate PCR bias induced by taxon-specific amplification efficiencies.

### Dataset Pre-processing

Even after generating ASVs, metabarcoding data poses many opportunities for choice of resolution and specificity when analyzing communities. First, we amplified both the COI and 18S marker regions. COI provides better species identification for the metazoan mesozooplankton taxa of interest but misses key groups such as pelagic tunicates. In contrast, 18S has a wider coverage and is able to capture more taxa but sacrifices specificity due to its lower mutation rate, which limits its ability to discriminate beyond the level of genus or family. Previous work has combined COI and 18S data to account for differences in coverage, however we did not include this approach for these analyses (Steffani et al, 2018).

Prior to quality filtering, we recovered 4941 sequences for COI containing 2180 unique ASVs (Table 1). For 18S we recovered 6918 sequences with 1188 distinct ASVs. We compared taxonomic designations made through BLAST with the Metazoogene database (Bucklin *et al.*, 2021). Based on the coverage, resolution, and relevant filtering to metazoan zooplankton taxa of interest we proceeded with our analysis with Metazoogene assignations. For ASVs that were not identified by Metazoogene, we augmented the taxonomy file using BLAST identifications. This was important for non-zooplankton taxa which although not the target of this study, needed to be accounted for in PCR bias-mitigation.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Total # unique ASVs | Kingdom | Phylum | Class | Order | Family | Genus | Species |
| mtCOI BLAST | 2300 | 1 | 22 | 35 | 67 | 107 | 138 | 152 |
| 18S rRNA BLAST | 1288 | 1 | 14 | 18 | 35 | 49 | 46 | 45 |
| mtCOI (Metazoogene) | 1501 | 1 | 8 | 18 | 16 | 45 | 64 | 78 |
| 18S rRNA (Metazoogene) | 1130 | 1 | 5 | 8 | 14 | 42 | 50 | 50 |

Table 1. The number of ASV’s identified to primary taxonomic levels for COI and 18S using BLAST and Metazoogene. Metazoogene assigns taxonomy and filters to only metazoan zooplankton taxa.

Metabarcoding is limited by taxon-specific amplification biases and masking the detection of rare species (Shelton *et al.*, 2023). Data transformations such as normalization to a median number of read counts can alleviate some of the bias in raw reads, but methods to estimate true proportions of species in a sample that can be correlated to biomass at the species level remain in development. Here we include analysis of raw relative read abundance (RRA) and our PCR bias corrected relative abundances (PCR-RA) only for particular ASVs.

### Additional Pre-processing For fido & Use in Modeling Against Environmental Properties

Prior to the implementation of the R package *fido* and PCR bias correction, we applied additional filtering to examine only abundant and widespread ASV’s that met a specific set of criteria. Previous work using *fido* modeled human gut microbiome data and was analyzed at the genus level and genera that were not seen in at least 30% of samples with at least 3 counts were amalgamated together into a category called ‘other’ for analysis (Silverman *et al.*, 2021). Due to the environmental and size-structured nature of our samples and uneven sample collection across environments, we applied a modified pre-processing filtration.

First, we used the *phyloseq* function ‘*tax\_glom’* to agglomerate ASVs at the family and for 18S and at the genus level for COI. In our analyses we examined whether the taxonomic level of Order might be more suitable for agglomeration for subsequent comparison to the Zooscan data where the images were classified to Order Calanoida. Order agglomeration would also prevent unidentified families from being included in the ‘other’ category. However, we determined that the variation in amplification among families or genera was more significant and proceeded using a higher taxonomic resolution.

Next, we selected only taxa that appeared in all PCR cycles (20, 24, 28) and at least one replicate of each cycle of our variable PCR experiment to ensure that an amplification efficiency (AE) could be adequately calculated for each taxon. Remaining ASV’s that did not meet our criteria were pooled into an ‘other’ category alongside the unidentifiable ASVs. For subsequent methods comparison with Zooscan image data we created a category for ‘unidentified Calanoida’, where ASVs identified to Order Calanoida, but where the 18S family or COI genus levels did not pass the filtering requirements were amalgamated. In some samples for 18S and many for COI, the ‘other’ category comprised a majority of the reads. To avoid biases imposed by samples dominated by ‘other’ reads we applied a threshold that the ‘other’ category comprises no greater than 20% of the total reads for a given sample for 18S. The same threshold of 20% of sample resulted in a majority of the samples failing to pass for COI, so this threshold was relaxed to 55% so that no greater than 25% of samples were discarded.

Separate analyses based on size class were necessary for pre-processing, as larger taxa were abundant in different subsets of samples than smaller taxa and contained distinct biological communities. For each size class we then used fido to predict the PCR bias-mitigated DNA reads and taxon-specific proportions of total reads at cycle 0. We recombined all size classes after this prediction step and proceeded with analyses on these data, from here on referring to them as ‘PCR bias-mitigated’ data.

In addition to its suitability for modeling PCR bias, *fido* can be used to construct Bayesian Multinomial Logistic Normal relationships between taxa and environmental variables. We applied the previously described criteria to examine which environmental properties displayed significant relationships with PCR-bias corrected DNA concentrations of 18S families or COI genera (Table 4). We used the same set of default priors for both primers and all sizes. The resulting mean centered-log ratio values that did not contain 0 in the 95% credible interval were deemed significant, however we present the 50% credible interval for visualizations and to emphasize the central tendency of our posterior results.

### Zooscan Sample Processing and Analysis

Formalin-preserved samples were size fractionated (>1000 µm & 200-1000 µm) and quantitatively sub-sampled to approximately 1500 individuals, which were distributed into the Zooscan imaging tray. A complete image was taken, the background was removed, and individual plankton vignettes were extracted into regions of interest (ROI’s) using ImageJ. Touching plankton were manually separated by the user, then re-processed. Processed samples and metadata were uploaded to Ecotaxa-a web application where users upload and host images of individual plankton and particles, classify them using taxonomic annotation and machine learning algorithms and export processed ecological datasets.

We followed this workflow for formalin-preserved zooplankton samples from the port side net of the oblique bongo tows from which our size-fractionated DNA tissue samples were collected. Due to limitations in taxonomic expertise, we identified images into 35 taxonomic categories, multiples, and an additional 4 non-living categories. Images were classified using deep-learning feature extraction and a fast-to-train classifier using 69 features using a training set comprised of 20% of the total images and the Zooscan Reference dataset (Picheral, Colin and Irisson, 2017). All images were manually validated into one of the 35 categories.

We used estimated spherical diameter (ESD) from each image and relationships derived from Zooscan imaging to convert between ESD and total length (TL) (Cornils *et al.*, 2022). Next we used length-carbon relationships for available categories to calculate carbon-biomass for each organism (Lavaniegos and Ohman, 2007). To compare with size-fractionated metabarcoding data, we used TL (or if unavailable, feret diameter) to group classified plankton images into 0.2-0.5 mm, 0.5-1mm, and 1-2 mm size classes. Taxon abundance and carbon biomass were standardized for aliquot volumes and the filtered volume of seawater measured by the flowmeter during the net tow. These metrics were transformed to calculate an abundance and carbon biomass estimate per m2 by multiplying by the maximum tow depth. Category-based relative abundances and carbon biomass proportions were computed to compare with proportions of DNA reads from raw (RRA) and PCR bias-mitigated (PCR-RA) data. Subsequent analyses focus on Zooscan proportion dry weight carbon biomass (Zoo-PB) to compare with the metabarcoding metrics.

### Statistical Analyses

To determine whether RRA and PCR-RA were correlated with morphological metrics of Zooscan proportion of dry weight carbon biomass (Zoo-PB) and relative abundances (Zoo-RA), we combined the higher-level taxonomic classifications of the metabarcoding data (family for 18S and genus for COI) to the level comparable for Zooscan classifications.

We used data transformations for non-normally distributed data prior to statistical analysis. In modeling taxon abundance on environmental features we log transformed chlorophyll maximum depth, distance from shore, hypoxia depth and nitracline depth. Prior to correlation analyses, the proportional DNA data were transformed to stabilize variance.

## Results

|  |  |
| --- | --- |
| **18S (Families)** | **COI (Genera)** |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | Acanthephyridae | Heterorhabdidae | | Acartiidae | Limacinidae | | Augaptilidae | Lucicutiidae | | Bolinopsidae | **Metridinidae\*** | | **Calanidae\*** | Mysidae | | Candaciidae | Oikopleuridae | | Centropagidae | **Oithonidae** | | Cercopagididae | Oncaeidae | | **Clausocalanidae\*** | **Paracalanidae\*** | | Cliidae | Peraclidae | | Clionidae | Pyrosomatidae | | Corycaeidae | **Rhincalanidae\*** | | Cymbuliidae | Rhopalonematidae | | Diphyidae | **Salpidae** | | Doliolidae | Scolecitrichidae | | **Eucalanidae\*** | Sergestidae | | Euchaetidae | Spinocalanidae | | **Euphausiidae** | Subeucalanidae | | Geryoniidae | Temoridae | |  | |  | |  | |  | |  | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  | | --- | --- | --- | --- | | Abyssorchomene | **Euphausia** | | Notocomplana | | Acartia | Eurythenes | | Oncaea | | Acrocalanus | Eusergestes | | Pantachogon | | Aetideus | Gennadas | | **Paracalanus\*** | | Agalma | Heliconoides | | Paraeuchaeta | | Allosergestes | Heterocarpus | | Paraphronima | | Amphicaryon | Heterorhabdus | | Phacellophora | | Apolemia | Labidocera | | Phronima | | Botrynema | Lensia |  | **Pleuromamma\*** | | **Calanus\*** | Lestrigonus | | Primno | | Calocalanus | Lophothrix | | Pseudevadne | | Candacia | Lucicutia |  | Pseudocalanus | | Charistephane | Marrus |  | Pyrosoma | | Clausocalanus | Mecynocera | | Racovitzanus | | Clio | **Metridia\*** |  | Rhincalanus | | Clione | **Muggiaea** |  | **Rosacea** | | **Ctenocalanus\*** | Nanomia |  | Stylocheiron | | **Ditrichocorycaeus** | Nematobrachion | | Themisto | | Doliolum | Nematoscelis | | Thysanoessa | | Eucalanus | Neocalanus | | Thysanopoda | | Euchaeta | Neosergestes | | Vibilia | | | |  | |  | |

Table 2. Table of all taxa identified in metabarcoding analysis using Metazoogene. For 18S taxa are identified down to family, for COI down to genus. Taxa in bold are included in the PCR bias mitigation process and subsequent analyses. The ‘\*’ identifies the subset of taxa in the order ‘Calanoida.’

Higher taxonomic resolution of family for 18S and genus for COI provided more nuanced insight into cross-shore patterns in taxonomic RRA. We generated treemaps of RRA for the onshore and offshore clusters for both COI and 18S, separating at the order the calanoids and euphausiids from all other taxa for visualization (Figure 7). For COI, we examined diversity differences at the genus and species level, whereas for 18S we compared environments at the family and genus level (the taxonomic resolutions used in PCR bias mitigation).  
 For the onshore environment, COI raw reads were dominated by two genera-*Calanus (*specifically *Calanus pacifius*)and *Euphausia (*specifically *Euphausia pacifica)*, which aligns with the two most dominant orders (Figure 7). Offshore, RRA was spread more evenly between calanoid genera including Eucalanus, Ctenocalanus, Clausocalanus, Candacia, Euchaeta, Mecynocera, Calanus, Acrocalanus and Pleuromamma. Euphausia also represented a significant but smaller portion of the read abundances than in the onshore.

In taxa outside of calanoids and euphausiids, gelatinous taxa were well-represented in both onshore and offshore for COI. Onshore the siphonophore Rosacea was dominant alongside other siphonophore genera including Nanomia, Muggiaea, and Lensia. Offshore the scyphozoan Phacellophora contributed to a considerable proportion of the residual reads, and siphonophores Rosacea, Agalma and Lensia appeared in the top ten species.

For 18S, the calanoids and euphausiids were composed of similar families and genera onshore and offshore, though their relative contributions to total reads varied (Figure 7). In both environments, the genus Metridia comprised a similar majority of the reads. Other calanoids included Pleuromamma, Calanus, Clausocalanus, Ctenocalanus, Candacia, Paracalanus, Rhincalanus, Neocalanus and Eucalanus. Euphausiids had a higher proportion of reads onshore but could not be identified beyond the family level.

Compared to COI, the read abundances of calanoids and euphausiids comprised an overwhelming majority in 18S, so taxa presented here reflect a small fraction of total onshore and offshore reads (Figure 6). Oithonidae, a family of cyclopoid copepods had two representative genera in the residual reads: Dioithona and Oithona, the latter of which was dominant amongst non-calanoids and euphausiids in both environments (Figure 7). Other cyclopoids of the family Oncaeidae and Corycaeidae were present in both environments. In the onshore, pelagic tunicates including salps (Pegea and Cyclosalpa) and appendicularians (Oikopleura) and a family of mysiids, Mysidae, were more represented. In the offshore, pteropods (Limacinidae and Cymbuliidae) and decapods (Acanthephyridae and Sergesitidae) comprised a significant proportion of the reads compared with onshore.

A screenshot of a cell phone

Description automatically generated

Figure 7. Treemaps of taxa from 18S (upper) and COI (lower) with columns associated with offshore (left) and onshore (right) environmental clusters. The upper row of plots for each primer includes the top 10 taxa (genus for 18S and species for COI) from orders Calanoida and Euphausiacea while the bottom row is the top 10 taxa from all other orders. For 18S genus names are printed in italics in the center and family on the upper left. For COI species names are printed in italics in the center and genus in the upper left. The size of each rectangle scales with raw relative read abundance for all samples from each cluster normalized to the median read count. Plots are comprised of proportional read abundances for samples associated with environmental clusters.

### Variable PCR Experiment, Zooscan, and Comparison of Estimates of Zooplankton Abundance

*Amplification efficiencies*

Following the filtering and data pre-processing previously described, we retained 9 of the 40 families for 18S and 12 genera for COI, with the remaining taxa classified into ‘other’ (Table 2). These groupings included 6 families and 6 genera from the Order Calanoida, as well as an ‘unidentified Calanoida’ which we aggregated for our Zooscan methods comparison. The relative abundances of each taxon varied across size, however Metridinidae and Calanidae were dominant in all sizes for 18S, while Calanus, Euphausia and ‘other’ were dominant in COI RRA (Figure S 3).

The results of the variable PCR cycle experiment yielded the first estimates of taxon-specific amplification efficiencies (AE) for environmental marine zooplankton samples. AEs among sizes and taxa were variable, with certain groups showing consistent under or over representation and others where bias changed across sizes or was not significantly different across cycles (Figure 9). In our results, a near-zero log-fold AE meant that a taxa’s relative read abundance closely preserved throughout the PCR process and amplified consistently.

The experiments amplifying the 18S primer revealed trends in AE that varied as a function of size class for many of the families (Figure 9). Calanidae, Metridinidae, and Euphausiidae all displayed positive AEs across the 3 size classes whereas Eucalanidae, Oithonidae and Salpidae were negatively amplified. Clausocalanidae, Paracalanidae, Rhincalanidae and the ‘other’ grouping had amplification efficiencies that were near-zero or included positive and negative CLR values between the size classes.   
 Calanidae and Metridinidae had the most consistent AE across the size classes (Figure 9). Clausocalanidae, Oithonidae and Paracalanidae exhibited decreased AE with size, although for Oithonidae and Paracalanidae the two largest sizes were not significantly different. The remaining 5 families all displayed monotonically increasing AE with size.

Not all families occurred in all pools, which was expected due to the environmental differences between subsets of samples (Table S 2). For example, Salpidae occurred only in the ‘All pool’ and ‘Pool 3’, which was comprised of primarily onshore sites. In contrast, Oithonidae occurred in only the ‘All pool’ and ‘Pool 2', which was an aggregation of primarily offshore sites.

Genera from the COI variable PCR experiment displayed greater variability both across pools and size classes, which may be attributed to using only 2 PCR cycles to compute AEs (Figure 15). Several taxa appeared in the subset PCR pools that did not occur in the ‘All pool’ including Clausocalanus, Ditrichocorycaeus, an unknown genus of Sagittidae (chaetognaths), Eucalanus, and Muggiaea as well as other genera in certain size classes. The genus Calanus had a centered log-ratio amplification that was not significantly different from zero, but which displayed increasing variability between pools with decreasing size. Euphausia and Pleuromamma had the strongest agreement between independent measurements of AE and exhibited opposite trends with more positive or negative AE respectively at the extreme size classes. In contrast with AE determined by 18S, the ‘other’ grouping had an overall positive AE and few genera displayed monotonic relationships between AE and size (Figure 9, Figure 15). In agreement with 18S, the gelatinous taxa (Muggiaea and Rosacea) had negative AE values.

A graph with colorful dots and lines

Description automatically generated

Figure 9. Estimates of PCR amplification efficiency in centered log-ratio values for 18S families. The x-axis is the size fraction, the y-axis is the centered log ratio value for each family related to PCR cycle number and the point color is the sub-pool of the calibration experiment. Error bars represent the 95% credible interval, and the zero line is highlighted in bold,

*Comparison of Raw Relative Read Abundances PCR Bias-Mitigated Relative Read Abundances, and Zooscan Biomass Proportion*

Taxon-specific amplification efficiencies using all calibration pools were used to correct raw reads (RRA) and produce a PCR bias-mitigated relative read abundance (PCR-RA). We compared RRA, PCR-RA and the relative proportion of Zooscan biomass (Zoo-PB) for the most abundant group in our samples: calanoid copepods (Order Calanoida). In pre-processing, reads identified to Order Calanoida but not family or genus were amalgamated into the ‘unidentified Calanoida’ category, which allowed us to compare all calanoid copepod ASVs with image-based abundances (and not just families and genera to which PCR bias correction was applied).

When comparing RRA and PCR-RA with Zoo-PB for 18S, we calculated a median square error (MedSE) in the difference of the log-transformed abundance metrics for each size class. We found that PCR-RA had a lower MedSE across all size classes, which implies that the correction improved the agreement between Zooscan and metabarcoding-based abundances (Figure S 6). As was clear from our description of community composition, RRA overestimates calanoid copepod abundance compared to the other methods for 18S, especially for the 0.2-0.5 mm size class. Applying PCR bias-correction successfully reduced this in most samples, reducing the proportion of calanoid reads to closer to Zoo-PB (Figure 10).

In contrast with 18S, the smallest size class of calanoid copepods for COI was underestimated by RRA relative to Zooscan biomass (Figure 10). In most cases PCR bias-correction increased the relative abundance of calanoids, resulting in a MedSE for COI that was lower for the 0.2-0.5mm and 1-2 mm but slightly higher in the intermediate size class (Figure S 6). For both COI and 18S, MedSE was lowest for the 0.2-0.5mm size class and largest for the 1-2 mm size class (Figure S 6). In comparing the magnitudes of RRA or PCR-RA with Zoo-PB 23-29% of samples for the 18S size classes and for 23-53% COI displayed poorer agreement for PCR-RA than for RRA.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 10. Differences between Zooscan imaging proportion of total carbon biomass (Zoo-PB, green), proportion of PCR bias-mitigated relative read abundance (PCR-RA, blue), and proportion of total raw relative read abundance (RRA, orange) for calanoid copepods ordered by increasing values of PC1, which is associated with the cross-shore environmental gradient. A ‘\*’ indicates samples where there is greater agreeement between Zoo-PB and PCR-RA and ‘X’ denotes samples where there is greater agreement between the Zoo-PB and RRA.

While previous work has shown that biomass correlates better with read abundances than counts of organisms, we still explored whether relative abundance from the Zooscan imaging (Zoo-RA, calculated as the proportion of counts for a given taxa) yielded better agreement with molecular methods than Zoo-PB (Matthews, Goetze and Ohman, 2021). The results of this comparison were highly similar to those of the RRA and PCR-RA comparison with Zoo-PB, however there were marginal improvements in the MedSE for all PCR-RA comparisons for both 18S and COI (Figure S 7, Figure S 8). In contrast, using Zoo-RA yielded a significantly higher MedSE for the RRA comparisons (Figure S 7, Figure S 8). In these comparisons, PCR-RA outperformed RRA across both primers and all size classes. As with the Zoo-PB comparison, the 0.2-0.5 mm size class displayed the smallest MedSE and the 1-2 mm size class the highest MedSE.

Correlations between the different metrics of abundance did not reveal significant relationships, whether we examined proportions, absolute measurements, or centered log-ratio transformed values of calanoid copepod reads or Zooscan biomass or relative abundance. While separating correlation analyses by size was necessary and strengthened correlations, it did not produce significant results.

When comparing RRA and PCR-RA through correlation analysis, we found significant results between the two DNA-based metrics (Figure S 9). Correlations between RRA and PCR-RA were expected because PCR-RA was modeled on RRA values, however the correlation was not one-to-one and varied across primer and size class (Figure S 9). We found the strongest correlation between the 1-2 mm size class for 18S and the 0.2-0.5 mm size class for COI (R=0.86, p<0.0001 and R=0.84, p<0.0001 respectively). For COI, the medium and large size classes did not yield significant correlations (p>0.05) and for 18S, the 0.2-0.5 mm and 0.5-1 mm size classes were weaker (R=0.54, p=0.03, R=0.69, p<0.01 respectively) but still significant. We attribute the weaker correlations for the smaller size classes in 18S to the fact that RRA overestimated calanoid copepods abundance for many of these samples and so RRA was uniformly high.

## **Discussion**

Metabarcoding studies continue to become more widely adopted in marine biodiversity monitoring. Recent publications have leveraged metabarcoding measurements to characterize the diversity of plankton communities and association with environmental measurements such as temperature, oxygen, and chlorophyll-a, and features such as topography and depth (Gallego *et al.*, 2020; Pitz *et al.*, 2020; Govender *et al.*, 2023; Matthews and Ohman, 2023). The collection of physical environmental data is already widespread, with temperature, salinity, chlorophyll-a fluorescence, oxygen, and other measurements monitored continuously and autonomously coastally and in open oceans.

This study revealed cross-shore community differences in the California Current Ecosystem in both environment and biological communities during the conditions observed during P2107 across varied sizes of the zooplankton community. Analysis of our physical and chemical environmental data revealed two distinct clusters-an onshore cluster characterized by denser water higher in nutrients and chlorophyll-a characteristic of upwelling, and an offshore cluster with warmer, more oligotrophic waters. Principal components analysis yielded a first principal component (PC1) that correlated with key environmental measures and explained a majority of the variance of data. The aim of P2107 was to study cross-shore fluxes and changes in plankton community and biogeochemistry. Our finding of an onshore and offshore cluster emerged from the data, and correlations between distance from shore with many biogeochemical measurements corroborates the aim to sample a cross-shore gradient.

Metabarcoding results yielded a diversity of taxa, with >45 species, genera, and families for 18S and >100 families, genera, and species for COI-far exceeding the 35 groups we were able to identify for Zooscan imaging. The zooplankton community identified through metabarcoding exhibited similar cross-shore trends to environmental properties at a community level. Using the raw read abundance data for zooplankton taxa assigned through Metazoogene we found that diversity (calculated as *H’*) correlated significantly with PC1 for 18S. This cross-shore diversity gradient varied between size-fractionated zooplankton communities for both COI and 18S. The largest size (1-2 mm) displayed an increase in diversity towards onshore values of PC1, and the smaller size classes (0.2-0.5mm and 0.5-1 mm) exhibited decreased diversity from offshore to onshore. Our aggregated results for both COI and 18S results agree with those of Matthews and Ohman (2023), where epipelagic richness was lower nearshore.

Using traditional metabarcoding data of read abundances, we characterized mesozooplankton community composition for both COI and 18S across the gradient. Both primers found that calanoid copepods and euphausiids dominated reads, but residual reads in other groups revealed greater diversity that diverged between COI and 18S. For COI, siphonophores, chaetognaths, amphipods and an anomalous patch of scyphozoans were among the remaining reads. For 18S, pelagic tunicates including pyrosomes, salps and doliolids, as well as cyclopoid copepods occurred across the gradient. Relative abundances of these distinct groups, even at lower taxonomic resolutions, provide support for dual-marker analysis and reveal limitations in each primer’s ability to resolve different members mesozooplankton community diversity. When comparing our findings to previous studies of the cross-shore diversity gradient in mesozooplankton, our finding of elevated RRA of siphonophores aligns with the findings of Matthews and Ohman (2023), which was not resolved by the morphological methods of (Lavaniegos and Ohman, 2007). In Matthews and Ohman (2023), Eucalanids were among the most dominant taxa, however in our bongo net data they were less abundant, especially in the 18S community which may be attributed to their peak depth distribution residing below the integrated tow depth for our study.

Modeling the more abundant and ubiquitous families and genera that passed our pre-processing on environmental measurements with fido, we found taxa associated with onshore and offshore conditions, as well as particular environmental relationships such as the association between salps and ammonium and Metridinidae and silicate in 18S and many taxa and NO3 in COI. In our analysis, nutrient measurements including NH4, NO3, and silicate proved to be some of the more important biogeochemical factors associated with zooplankton abundance. Ammonium specifically produced strong associations with a number of both 18S families and COI genera. During P2107, ammonium was observed to be higher in the nearshore cycles and taxa associated with the high NH4 rather than NO3 may be taking advantage of organic matter decomposition or excretion later in the upwelling evolution.

The results of our fido models also revealed that PCR bias-correction of metabarcoding samples for 18S can improve our ability to resolve environmental relationships with DNA-based taxonomic abundances; however, the same strengthening was not observed for the COI models. For the COI fido model, there were more positive than negative associations with the selected environmental variables related to the PC1 environmental gradient. This aligns with the uneven nature of the zooplankton sampling of P2107 where more productive, nearshore waters associated with greater values of the PC1 variables were sampled more heavily. Although the offshore community was more diverse, imbalance in samples and the subsequent PCR bias mitigation pre-processing may have produced a subset of the community more representative of the onshore environment at the higher genus resolution for COI.

### Variable PCR Experiment, Zooscan, and Comparison of Estimates of Zooplankton Abundance

The variable PCR cycle experiments produced amplification efficiencies that showed strong agreement in replicates within a size-fraction, but varied across sizes for many taxa. The taxa that displayed the most consistent amplification efficiency were also those with the highest RRA (Figure 9, Figure 11, Figure S 5). We examined this further and found significant correlations between RRA and amplificaiton efficiency for both 18S and COI (18S: R=0.65, p<0.001, COI: 0.61, p<0.001) (Figure 11). These findings imply that more abundant taxa are preferentially amplified by PCR and that amplification efficiency for a taxon tends to stabilize when that taxon dominates.

The pattern for the relationship between AE and RRA was an loosely exponential, which aligns with the expectation that PCR bias does not linearly amplify the more abundant taxa across multiple PCR cycles. However, particular taxa or sizes of taxa which deviated from the expected curve could reveal taxon and primer-specific, rather than abundance-based biases in amplification. For example, in 18S the large ‘other’group and in COI larger Pyroteuthis and medium Calanus displayed higher than expected amplification efficiencies.

Comparing the two primers, 18S adheres much more closely to an expected exponential function, whereas COI displays divergent patterns at positive AE values. Many more COI genera had higher than expected total RRA compared to 18S. In both cases the highest observed identifiable AE values were crustaceans and the lowest were gelatinous taxa (as well as a non-zooplankton *Pyroteuthis* genus). As noted previously, sizes of a given family or genus roughly had similar AE values. From our comparison, amplification efficency for 18S families appears more predictable from gene copy reads alone than for COI genera. We recognize that COI AEs as calculated here contain more uncertainty from a 2 PCR cycle calculation and that more COI genera were included in our analysis thatn 18S families. Future work should explore whether this bias is primer-specific, taxon specific, or a function of the taxonomic resolution optimal for the given primer. A graph of different colored dots

Description automatically generated with medium confidence

Figure 11. Amplification Efficiency (AE) vs. the Sum of relative read abundances (RRA) for selected families (18S) and genera (COI). Color is associated with taxonomic group and listed by increasing mean AE and shape is associated with size fraction.

The results of our methodological comparison indicate that correcting for PCR bias using the variable PCR cycle method as implemented here can improve agreement between molecular and morphological measures of relative proportions of DNA and carbon biomass or species counts for abundant taxa. In the case of COI, we were limited by only 2 PCR cycles to compute an amplification efficiency, however the correction still improved methological agreement based on differences in median squared error.

Our study is the first to implement this method using metazoan plankton communities and the collection, processing and composition of these data come many complexities that were not present in this method’s inception using microbiome data (Silverman *et al.*, 2021). First, our community was comprised of metazoans, which are multicellular, vary in gene copy number between species, and include a diversity of taxanomic groups with different primer affinities. Second, the nature of the sampling, sub-sampling and environment resulted in communities with distinct differences in size structure and composition related to a gradient in water mass properties. Rare taxa that occurred in few samples or environments could not be detected in the ‘competition to be counted’ of the pooled calibration sample or pass the pre-processing filtering. By manually size fractionating our sample upon collection, we were able to resolve some of the bias that results from traditional approaches that aggregate all sizes.

Unfortunately, calibration samples were not pooled by size, so interactions between sizes within the calibration samples remain unquantified. Diversity in the smallest size class was highest offshore, which also had fewer representative samples and which may have further limited the diversity captured in the pooled calibration sample. In contrast, the larger size classes were more likely to contain contamination from smaller zooplankton that were not completely size fractionated, were in guts, or stuck together. Regardless, the smallest size class was most effective in reproducing the proportions of biomass and relative abundances from the Zooscan.

In the fido models of environmental variables we noted that uneven sampling across the gradient may be especially impactful when working with a higher taxanomic resolution barcode such as COI where a genus has a narrower niche than a family. To further explore the relationship between sample composition and environment and the biases imposed within our PCR bias correction experiment, we computed the Bray-Curtis dissimilarity between each sample and the calibration pool samples for the pre-processed fido taxa (Figure S 10). We found that COI had greater dissimilarity than 18S and that dissimilarity was highest in the intermediate environment for all sizes. The 0.2-05 mm size class was more similar to the calibration samples offshore, whereas the 0.5-1 mm and 1-2 mm size classes were more similar onshore. This aligns with the expectation that the diversity captured in the middle of the environmental gradient where there was the least replication may have been absorbed into the ‘other’ category or lost during PCR bias-mitigation.

Bray-Curtis dissimiarilty was more variable for 18S, especially on the onshore and offshore extremes. The highest dissimilarity for 18S was observed nearshore for the smaller size classes and towards the offshore for the 1-2 mm size class (Figure S 10). In contrast with COI, the lowest dissimilarity for 18S was observed at a set of the intermediate transect samples. Lower dissimilarity bewteen the calibration sample and transect aligns with the concept that the calibration sample is a ‘mean’ sample, representative of all environments and taxa sampled. In this way, transect samples that mix onshore and offshore taxa might physically generate a ‘mean’ sample. Because COI displayed the opposite pattern, the pre-processing of agglomeration at family vs. genus for 18S may have allowed the diverse taxa lost in the COI samples to be absorbed into a 18S family.

Primer choice, size-fraction, and other grouping variables imbedded naturally and methodologically in our data such as the sampling environment (cycle or cluster) might exhibit more influence on the comparison between Zoo-PB or Zoo-RA and PCR-RA than with RRA. This may be a product of the variable PCR experiment in itself, where pooling environmentally and taxonomically diverse samples creates an artificial community that does not accurately reflect all samples equally. In future implementations of this method we advise the following:

1. **Establish optimal poolings using ­*a-priori­* data on environment or biological communities.** Use available environmental data to generate hypotheses as how to group your samples in creating a pooled calibration sample, similar to our cluster analysis. If resources allow, first sequence your collected samples and then use multivariate analysis such as NMDS to generate clusters of similar communities based on taxanomic composition.
2. **Run more PCR cycles.** In our experiment we used 3 variable PCR cycles (20, 24, 28) which left us with only 2 points in the case of COI where one experiment failed. More cycles would buffer against PCR failure and allow for a more robust model of amplifcation efficiency.
3. **Incorporate scale-dependence into PCR bias correction model.** Zooplankton metabarcoding studies that often normalize based on a semi-arbitrary sequencing depth could benefit from scale-reliant normalization. For example, inclusion of scales such as volume of water filtered, sub-sample sizes, manually measured dry weight carbon biomass or into emerging models such as ALDEx2 might better constrain uncertainty on the relationship between metabarcoding and morphological metrics (Nixon *et al.*, 2024; Nixon, Gloor and Silverman, 2024).

It is worth noting that while this study aimed to mitigate PCR bias in the metabarcoding measurements, there was also likely significant error present in the length-carbon calculations of zooplankton dryweight carbon biomass from the Zooscan. The length-carbon regressions for calanoids are well determined, however in computing calanoid biomass proporiton for the community, error associated with other groups was introduced. As with DNA metabarcoding, the Zooscan image generation workflow requires many steps involving manipulation, and subsampling with error we did not fully quantify here. While agreement with morphological and imaging methods is important when adopting metabarcoding approaches, unexplained variability between them methods may not necessarily be entirely attributed to metabarcoding. There remains a need to refine image-based prediction through expert tanomist annotation, more robust biomass models based on more flexible image parameters, and comparing outputs with not just metabarcoding, but also manual measurements.

# **Conclusions**

Based on the results of our metbarcoding analysis of P2107 and implementation of PCR bias-mitigation through variable PCR cycle experiments and subsequent methodological comparison, we conclude the following: 1) Metabarcoding resolves broad and established enviromental patterns in the CCE when compared with co-located physical and chemical measurements. 2) The high taxanomic resolution of metabrcoding identifies taxa associated with particular environments, including groups distinct to each primer. 3) Manual size fractionation of samples allows for not only more nuanced ecological interpretations, but also can influence the effects of PCR amplification and bias. 4) Utilization of *a-*­*priori* knowledge of data to pool samples for the calibration experiment based on similarities in expected composition can enhance agreement between molecular and morphological metrics of relative abundance. These prior groupings can be determined using environmental data (which we show here correlates broadly with zoopplankton community) or historical records of biogeography. Varying PCR cycle experiments may even be conducted after sequencing and analysis of the environmental samples, which could allow for a data-driven clustering to determine samples to pool. 5) PCR bias mitigation through the variable PCR cycle experiment can be applied to abundant groups in environmentally collected samples and these corrected relative abundances show stronger agreement with independent measures of relative abundance.

# Author Contributions

# Acknowledgements

We thank the *California Current Ecosystem* Long-Term Ecological Research program, the crew, and volunteers aboard the R/V Revelle for zooplankton sample and oceanographic data collection. A. McSorley and A. Rosenbaum assisted with Zooscan imaging. The SIO Pelagic Invertebrate Collection provided access to zooplankton samples for Zooscan imaging. Support was provided by the US NSF Graduate Research Fellowship to D. Capone, NSF to the *California Current Ecosystem* LTER site and REU program for H. Budroe.

# Conflict of Interest

The authors have no conflicts of interest to disclose.

# Data Accessibility Statement

Raw sequence reads and the code used to perform analyses have been deposited; additionally, the raw sequence data have been uploaded to NCBI. Code for the analysis is available on Github: <https://github.com/DanteCapone/CCE_Zooplankton_Metabarcoding_Pub> .

# Benefit-Sharing Statement

## References

Brinton, E. and Townsend, A. (2003) ‘Decadal variability in abundances of the dominant euphausiid species in southern sectors of the California Current’, *Deep Sea Research Part II: Topical Studies in Oceanography*, 50(14), pp. 2449–2472. Available at: https://doi.org/10.1016/S0967-0645(03)00126-7.

Bucklin, A. *et al.* (2016) ‘Metabarcoding of marine zooplankton: prospects, progress and pitfalls’, *Journal of Plankton Research*, 38(3), pp. 393–400. Available at: https://doi.org/10.1093/plankt/fbw023.

Bucklin, A. *et al.* (2021) ‘Toward a global reference database of COI barcodes for marine zooplankton’, *Marine Biology*, 168(6), p. 78. Available at: https://doi.org/10.1007/s00227-021-03887-y.

Carvalho, D.C. (2022) ‘Ichthyoplankton DNA metabarcoding: Challenges and perspectives’, *Molecular Ecology*, 31(6), pp. 1612–1614. Available at: https://doi.org/10.1111/mec.16387.

Collier, J.L. and Palenik, B. (2003) ‘Phycoerythrin-containing picoplankton in the Southern California Bight’, *Deep Sea Research Part II: Topical Studies in Oceanography*, 50(14), pp. 2405–2422. Available at: https://doi.org/10.1016/S0967-0645(03)00127-9.

Cornils, A. *et al.* (2022) ‘Testing the usefulness of optical data for zooplankton long-term monitoring: Taxonomic composition, abundance, biomass, and size spectra from ZooScan image analysis’, *Limnology and Oceanography: Methods*, 20(7), pp. 428–450. Available at: https://doi.org/10.1002/lom3.10495.

Décima, M., Landry, M.R. and Popp, B.N. (2013) ‘Environmental perturbation effects on baseline δ15N values and zooplankton trophic flexibility in the southern California Current Ecosystem’, *Limnology and Oceanography*, 58(2), pp. 624–634. Available at: https://doi.org/10.4319/lo.2013.58.2.0624.

Di Lorenzo, E. and Ohman, M.D. (2013) ‘A double-integration hypothesis to explain ocean ecosystem response to climate forcing.’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(7), pp. 2496–2499. Available at: https://doi.org/10.1073/pnas.1218022110.

Gallego, R. *et al.* (2020) ‘Environmental DNA metabarcoding reveals winners and losers of global change in coastal waters’, *Proceedings of the Royal Society B: Biological Sciences*, 287(1940), p. 20202424. Available at: https://doi.org/10.1098/rspb.2020.2424.

Gorsky, G. *et al.* (2010) ‘Digital zooplankton image analysis using the ZooScan integrated system’, *Journal of Plankton Research*, 32(3), pp. 285–303. Available at: https://doi.org/10.1093/plankt/fbp124.

Govender, A. *et al.* (2023) ‘Metabarcoding analysis of marine zooplankton confirms the ecological role of a sheltered bight along an exposed continental shelf’, *Molecular Ecology*, 32(23), pp. 6210–6222. Available at: https://doi.org/10.1111/mec.16567.

Hayward, T.L. and Venrick, E.L. (1998) ‘Nearsurface pattern in the California Current: coupling between physical and biological structure’, *Deep Sea Research Part II: Topical Studies in Oceanography*, 45(8–9), pp. 1617–1638. Available at: https://doi.org/10.1016/S0967-0645(98)80010-6.

James, C.C. *et al.* (2022) ‘Influence of nutrient supply on plankton microbiome biodiversity and distribution in a coastal upwelling region’, *Nature Communications*, 13(1), p. 2448. Available at: https://doi.org/10.1038/s41467-022-30139-4.

Jungbluth, M.J. *et al.* (2022) ‘Species-specific biomass estimation from gene copy number in metazoan plankton’, *Limnology and Oceanography: Methods*, 20(6), pp. 305–319. Available at: https://doi.org/10.1002/lom3.10487.

Lavaniegos, B.E. and Ohman, M.D. (2007) ‘Coherence of long-term variations of zooplankton in two sectors of the California Current System’, *Progress in Oceanography*, 75(1), pp. 42–69. Available at: https://doi.org/10.1016/j.pocean.2007.07.002.

Leray, M. *et al.* (2013) ‘A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents’, *Frontiers in Zoology*, 10(1), p. 34. Available at: https://doi.org/10.1186/1742-9994-10-34.

Lilly, L.E. and Ohman, M.D. (2018) ‘CCE IV: El Niño-related zooplankton variability in the southern California Current System’, *Deep Sea Research Part I: Oceanographic Research Papers*, 140, pp. 36–51. Available at: https://doi.org/10.1016/j.dsr.2018.07.015.

Matthews, S.A., Goetze, E. and Ohman, M.D. (2021) ‘Recommendations for interpreting zooplankton metabarcoding and integrating molecular methods with morphological analyses’, *ICES Journal of Marine Science*, 78(9), pp. 3387–3396. Available at: https://doi.org/10.1093/icesjms/fsab107.

Matthews, S.A. and Ohman, M.D. (2023) ‘A space-for-time framework for forecasting the effects of ocean stratification on zooplankton vertical habitat use and trait composition’, *Limnology and Oceanography*, 68(12), pp. 2688–2702. Available at: https://doi.org/10.1002/lno.12450.

Nixon, M.P. *et al.* (2024) ‘Scale Reliant Inference’. arXiv. Available at: http://arxiv.org/abs/2201.03616 (Accessed: 6 June 2024).

Nixon, M.P., Gloor, G.B. and Silverman, J.D. (2024) ‘Beyond Normalization: Incorporating Scale Uncertainty in Microbiome and Gene Expression Analysis’, *bioRxiv*, p. 2024.04.01.587602. Available at: https://doi.org/10.1101/2024.04.01.587602.

Picheral, M., Colin, S. and Irisson, J.-O. (2017) ‘EcoTaxa, a tool for the taxonomic classification of images: http://ecotaxa.obs-vlfr.fr’.

Pitz, K.J. *et al.* (2020) ‘Zooplankton biogeographic boundaries in the California Current System as determined from metabarcoding’, *PLOS ONE*, 15(6), p. e0235159. Available at: https://doi.org/10.1371/journal.pone.0235159.

Prokopowich, C.D., Gregory, T.R. and Crease, T.J. (2003) ‘The correlation between rDNA copy number and genome size in eukaryotes’, *Genome*, 46(1), pp. 48–50. Available at: https://doi.org/10.1139/g02-103.

Rebstock, G.A. (2002) ‘Climatic regime shifts and decadal-scale variability in calanoid copepod populations off southern California’, *Global Change Biology*, 8(1), pp. 71–89. Available at: https://doi.org/10.1046/j.1365-2486.2002.00456.x.

Shelton, A.O. *et al.* (2023) ‘Toward quantitative metabarcoding’, *Ecology*, 104(2), p. e3906. Available at: https://doi.org/10.1002/ecy.3906.

Silverman, J.D. *et al.* (2021) ‘Measuring and mitigating PCR bias in microbiota datasets’, *PLOS Computational Biology*, 17(7), p. e1009113. Available at: https://doi.org/10.1371/journal.pcbi.1009113.

Silverman, J.D. *et al.* (2022) ‘Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes’, *Journal of Machine Learning Research*, 23(7), pp. 1–42.

Suzuki, M.T. and Giovannoni, S.J. (1996) ‘Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR’, *Applied and Environmental Microbiology*, 62(2), pp. 625–630. Available at: https://doi.org/10.1128/aem.62.2.625-630.1996.

Taylor, A.H., Allen, J.I. and Clark, P.A. (2002) ‘Extraction of a weak climatic signal by an ecosystem’, *Nature*, 416(6881), pp. 629–632. Available at: https://doi.org/10.1038/416629a.

Wang, C. *et al.* (2017) ‘Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error’, *Proceedings of the Royal Society B: Biological Sciences*, 284(1859), p. 20170425. Available at: https://doi.org/10.1098/rspb.2017.0425.

Zhan, A. *et al.* (2013) ‘High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities’, *Methods in Ecology and Evolution*, 4(6), pp. 558–565. Available at: https://doi.org/10.1111/2041-210X.12037.

|  |  |
| --- | --- |
| **Calibration Sample Pool** | **Samples Represented** |
| All | *All samples* |
| Pool 1 | **Cycle 1**: *C1.T7.H9\_S1, C1.T7.H9\_S2, C1.T7.H9\_S3, C1.T8.H10\_S1, C1.T8.H10\_S2, C1.T8.H10\_S3*  **Cycle 2:** *C2.T8.H18\_S1, C2.T8.H18\_S2, C2.T8.H18\_S3, C2.T9.H19\_S1, C2.T9.H19\_S2, C2.T9.H19\_S3*  **Cycle 3:** *C3.T6.H25\_S1, C3.T6.H25\_S2, C3.T6.H25\_S3, C3.T7.H26\_S1, C3.T7.H26\_S2* |
| Pool 2 | **Cycle 3:** *C3.T7.H26\_S3*  **Transect 1:** *CT1.T1.H28\_S1, CT1.T1.H28\_S2, CT1.T1.H28\_S3, CT1.T2.H29\_S1, CT1.T2.H29\_S2, CT1.T2.H29\_S3, CT1.T3.H30\_S1, CT1.T3.H30\_S2, CT1.T3.H30\_S3, CT1.T4.H31\_S1, CT1.T4.H31\_S2, CT1.T4.H31\_S3, CT1.T5.H32\_S1, CT1.T5.H32\_S2, CT1.T5.H32\_S3, CT1.T6.H33\_S1* |
| Pool 3 | **Transect 1:** *CT1.T6.H33\_S2, CT1.T6.H33\_S3, CT1.T7.H34\_S1, CT1.T7.H34\_S2, CT1.T7.H34\_S3, CT1.T8.H35\_S1, CT1.T8.H35\_S2, CT1.T8.H35\_S3*  **Transect 2:** *CT2.T1.H36\_S1, CT2.T1.H36\_S2, CT2.T1.H36\_S3, CT2.T4.H39\_S1, CT2.T4.H39\_S2, CT2.T4.H39\_S3, CT2.T8.H43\_S1, CT2.T8.H43\_S2, CT2.T8.H43\_S3* |

Table S 2. Variable PCR Experiment Calibration Sample Pools. The 4 pools used to calculate PCR bias and amplification efficiency. Each pool was run at 20, 24 and 28 PCR cycles and contained samples from each size class. Pool 1 contained taxa from the 3 cycles, Pool 2 was comprised of samples from primarily offshore sites and Pool 3 was comprised of samples from primarily onshore sites.

A screenshot of a music score

Description automatically generated

Figure S 5. Amplification efficiencies for COI genera. The x-axis is the size fraction, the y-axis is the centered log-ratio value for each family related to PCR cycle number and the point color is the sub-pool of the calibration experiment. Error bars represent the 95% credible interval, and the zero line is highlighted in bold.

A screenshot of a graph

Description automatically generated

Figure S 6. Differences in square error between RRA-Zoo-PB and PCR-RA-Zoo-PB comparisons for calanoid copepods plotted as a function of sample and ordered by increasing values of PC1. Median Square Error is displayed on the lower left for 18S (left) and COI (right). A red/positive anomaly implies that PCR-RA produced smaller error for a sample, whereas a blue/negative anomaly implies that RRA-Zoo-PB comparison error was smaller.

A graph of different colored bars

Description automatically generated with medium confidence

Figure S 7. Differences between Zooscan imaging relative abundance (Zoo-RA), proportion of PCR bias-mitigated relative read abundance (PCR-RA, blue), and proportion of total raw relative read abundance (RRA, orange) for calanoid copepods ordered by increasing values of PC1, which is associated with the cross-shore environmental gradient. A ‘\*’ indicates samples where there is greater agreeement between Zoo-RA and PCR-RA and ‘X’ denotes samples where there is greater agreement between the Zoo-RA and RRA.

A graph of different numbers and a number of letters

Description automatically generated with medium confidence

Figure S 8. Differences in square error between RRA-Zoo-RA and PCR-RA-Zoo-RA comparisons for calanoid copepods plotted as a function of sample and ordered by increasing values of PC1. Median Square Error is displayed on the lower left for 18S (left) and COI (right). A red/positive anomaly implies that PCR-RA produced smaller error for a sample, whereas a blue/negative anomaly implies that RRA-Zoo-RA comparison error was smaller.

A diagram of different colored shapes

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated

Figure S 9 Scatter plots between PCR bias-mitigated relative abundance and raw reads relative abundance for 18S (upper) and COI (lower). The proportion data have arcsine-square root transformed to stabilize the variance. The facets of each plot and point colors are associated with the 3 zooplankton size classes and the point shape is the cycle. Pearson’s R and associated p-value with Benjamini-Hochberg correction applied are displayed in the upper right.

A graph of different colored dots

Description automatically generated

Figure S 10. Bray-Curtis Dissimilarity Index computed between variable PCR cycle calibration samples (cycles 20, 24, 28) and individual environmental samples for 18S (upper) and COI (lower). Color is associated with size fraction (red=0.2-0.5mm, green=0.5-1mm, blue=1-2mm), shape is the PCR Cycle Pool, and replicate points are technical replicates.

# Figure Outline

## Questions

1. **How does PCR bias correction impact relative abundances of the community?**
   1. Fig 1. Figure illustrating the selection criteria of taxa from raw to those for the PCR bias pipeline
   2. Fig 2. Stacked barplot comparing PCR to raw across sizes
2. **Which taxa are most impacted by PCR Bias?**
   1. Fig 3. Amplification
   2. Fig 4.
3. **How does PCR corrected reads compare to Raw reads when compared to Zooscan imaging?**
4. A screenshot of a computer

   Description automatically generated