


# Quantitative assessment of reef foraminifera community from metabarcoding data

Elsa B. Girard<sup>1,2</sup>  | Emilie A. Didaskalou<sup>3</sup> | Andi M. A. Pratama<sup>4</sup> | Carolina Rattner<sup>1</sup> | Raphaël Morard<sup>5</sup> | Willem Renema<sup>1,2</sup>

<sup>1</sup>Naturalis Biodiversity Center, Leiden, The Netherlands

<sup>2</sup>IBED, University of Amsterdam, Amsterdam, The Netherlands

<sup>3</sup>CML, University of Leiden, Leiden, The Netherlands

<sup>4</sup>Marine Science Department, Faculty of Marine Science and Fisheries, Hasanuddin University, Makassar, Indonesia

<sup>5</sup>MARUM, University of Bremen, Bremen, Germany

## Correspondence

Elsa B. Girard, Naturalis Biodiversity Center, Darwinweg 2, Leiden, The Netherlands.

Email: [elsa.girard@naturalis.nl](mailto:elsa.girard@naturalis.nl)

## Funding information

Cluster of Excellence "The Ocean Floor—Earth's Uncharted Interface", Grant/Award Number: 390741603; H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 813360

Handling Editor: Pierre Taberlet

## Abstract

Describing living community compositions is essential to monitor ecosystems in a rapidly changing world, but it is challenging to produce fast and accurate depiction of ecosystems due to methodological limitations. Morphological methods provide absolute abundances with limited throughput, whereas metabarcoding provides relative abundances of genes that may not correctly represent living communities from environmental DNA assessed with morphological methods. However, it has the potential to deliver fast descriptions of living communities provided that it is interpreted with validated species-specific calibrations and reference databases. Here, we developed a quantitative approach to retrieve from metabarcoding data the assemblages of living large benthic foraminifera (LBF), photosymbiotic calcifying protists, from Indonesian coral reefs that are under increasing anthropogenic pressure. To depict the diversity, we calculated taxon-specific correction factors to reduce biological biases by comparing surface area, biovolume and calcite volume, and the number of mitochondrial gene copies in seven common LBF species. To validate the approach, we compared calibrated datasets of morphological communities from mock samples with bulk reef sediment; both sample types were metabarcoded. The calibration of the data significantly improved the estimations of genus relative abundance, with a difference of  $\pm 5\%$  on average, allowing for comparison of past morphological datasets with future molecular ones. Our results also highlight the application of our quantitative approach to support reef monitoring operations by capturing fine-scale processes, such as seasonal and pollution-driven dynamics, that require high-throughput sampling treatment.

## KEYWORDS

COI, coral reef, foraminifera, monitoring, quantitative metabarcoding

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

Describing living community composition is essential to monitor ecosystems in a rapidly changing world, but it is a challenging task. It is especially important for well-studied ecosystems such as coral reefs, on which hundreds of organisms (including humans) rely (Hughes et al., 2017). Coral cover has been globally declining in response to human activities in the past decades, thereby increasing the urgency for improving existing monitoring tools to rapidly identify local causes to reef decline (Downs et al., 2005; Madin & Madin, 2015). One of the established bioindicator taxon groups of reef environmental conditions associated with coral growth and water quality are large benthic foraminifera (LBF) (Hallock et al., 2003; Humphreys et al., 2022; Prazeres et al., 2020; Renema, 2018; Uthicke & Nobes, 2008). LBF are calcifying protists that are generally larger than 0.5 mm in diameter, but rarely bigger than a centimetre ubiquitous in shallow coral reefs (Renema, 2018). Contrary to other, usually smaller, benthic foraminifera, LBF have a characteristic symbiosis with photosynthetic microalgae, similar to corals, next to a high diversity of endobiotic prokaryotes (Prazeres & Renema, 2019). They have a short community turnover rate (some months to a year), much shorter than coral colonies (many years) (Hallock & Reymond, 2022); hence, LBF community composition changes at the rhythm of changing environmental conditions. They can also make up to 70% of the inter-reef sediment and produce on average 5% and exceptionally up to 55% of the carbonate reef sediment (Dawson et al., 2014; Doo et al., 2017; Narayan et al., 2022; Renema, 2018, and references therein). LBF are therefore ecosystem engineers in reef environments.

Continuous progress is being made to improve taxonomy, species identification and detection, as well as community composition based on genetic information (Taberlet et al., 2012), especially for rare taxa or small organisms, like foraminifera (Pawlowski et al., 2016; Skelton et al., 2022 and references therein). Metabarcoding is a molecular tool that is used for community assessment (Gielings et al., 2021; Hassan et al., 2022; Miya, 2022; Taberlet et al., 2012) and biomonitoring (Cordier et al., 2021). It is a time-effective approach compared to morphological methods, which can be time-consuming and require taxonomic expertise (Miller et al., 2011). However, quantitatively estimating community composition (i.e., estimation of relative abundance and proportional biovolume) from molecular datasets is not straightforward and limited due to significant technical biases, such as those inherent to DNA extraction, PCR amplification and primer choice (Ficetola et al., 2016; Moinard et al., 2023; Shelton et al., 2023; Taberlet et al., 2012), in addition to environmental biases, such as DNA degradation, currents and sediment dynamics. Steps towards the resolution of some technical biases are ongoing with, for example, the development of corrections that can be implemented retroactively on already sequenced datasets, as described by Moinard et al. (2023) and Silverman et al. (2021) to overcome PCR-induced biases. Besides technical biases, biological biases are equally problematic because of differential gene copy numbers that can unpredictably fluctuate between

closely related species (Lamb et al., 2019; Pawluczyk et al., 2015). Such biological differences directly influence the relative number of sequence reads and can result in spurious proportional values upon estimating the community composition (Weber & Pawlowski, 2013). One way to remedy the later issue is by performing taxon-specific calibration in the form of correction factors (Lamb et al., 2019; Piñol et al., 2019; Shelton et al., 2023). These correction factors permit a translation of the number of reads into proportional biomass, biovolume or relative abundance estimates closer to reality. This approach has the potential to allow for more informative environmental monitoring, compared to uncorrected metabarcoding outputs, by rapidly producing outputs similar to specimen counting with higher taxonomic accuracy, although accuracy is dependent on the quality of the reference database (e.g., Ershova et al., 2023; Ratcliffe et al., 2021).

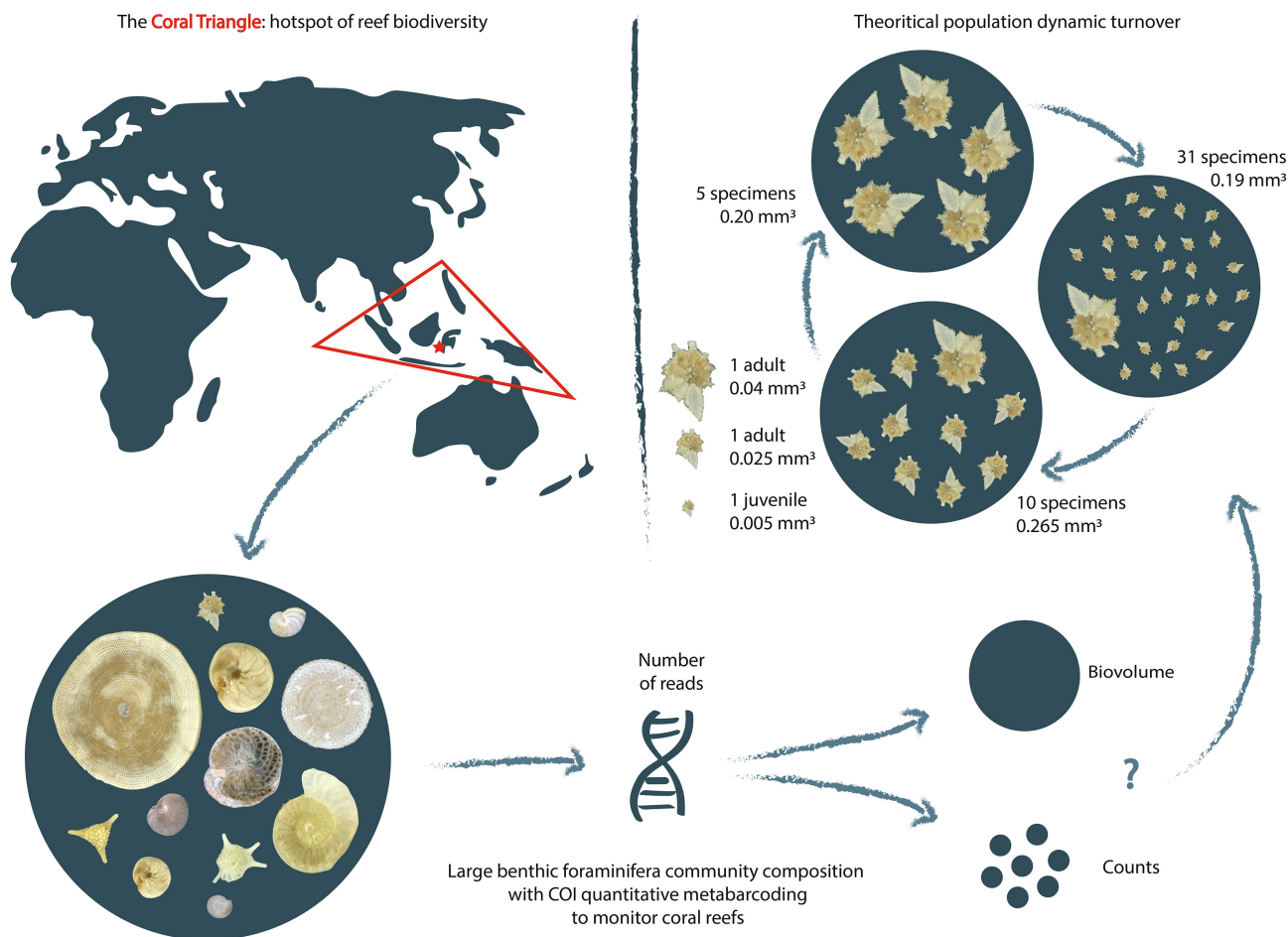
In foraminifera, the vast majority of the metabarcoding studies have used the nuclear marker SSU rDNA (e.g., Barrenechea Angeles et al., 2024; Brinkmann et al., 2023; Eqbal et al., 2022; Pawlowski et al., 2016). Discrepancies arise when comparing the proportion of reads and the number of specimens. The incompleteness of the reference database for the study sites and other biological biases are responsible for most of those discrepancies (Cavaliere et al., 2021; Frontalini et al., 2020, 2018). Biological biases may include variable numbers of nuclei from different reproduction strategies (Weber & Pawlowski, 2013; Zhao et al., 2019), variable number of gene copies in the genome (Milivojević et al., 2021; Weber & Pawlowski, 2013) and hybridization events (Pillet et al., 2012). Considering the above-mentioned biases for the nuclear marker, it might be more realistic to expect a correlation between the proportion of reads and proportional biomass or biovolume rather than the species relative abundance, assuming that gene copy densities are tightly bound with specimen size. However, it has been shown that the number of gene copies of common foraminiferal nuclear regions varied independently of cell size (Milivojević et al., 2021) as well as the number of nuclei within the cell (Weber & Pawlowski, 2013). Recently, a new mitochondrial marker was developed, the cytochrome c oxidase subunit I (COI) located at the Leray-region (Leray et al., 2013; Macher, Wideman, et al., 2021). This marker is a conserved coding region that has the potential to solve many of the issues encountered with the SSU rDNA nuclear marker (Girard, Langerak, et al., 2022). To our knowledge, no similar study to date has investigated the relationship between mitochondrial gene copy number and individual size, but we hypothesise mitochondria to be more abundant in larger cells since they provide energy to the cell and participate in regulation of cell growth, among other functions (Friedman & Nunnari, 2014; Wu et al., 2013).

We aim to develop an efficient tool to quantitatively assess foraminiferal community composition using the mitochondrial marker by correcting for biological biases (Girard, Macher, et al., 2022; Macher, Wideman, et al., 2021). This tool is meant to monitor, among others, coral reefs from the Coral Triangle, a hotspot of marine biodiversity under increasing anthropogenic pressure. LBF communities are highly diverse in the Coral Triangle, with 21 genera and more than

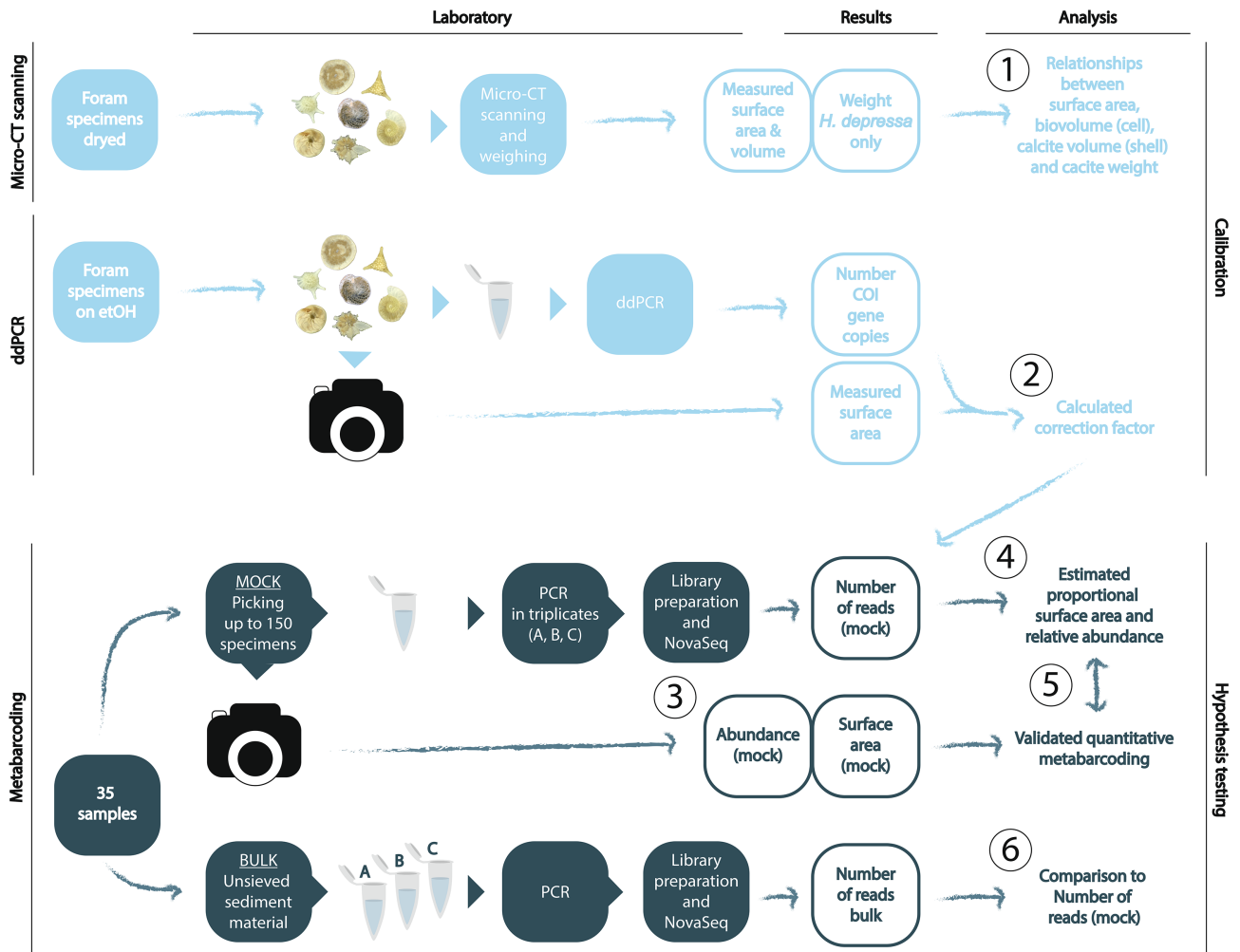
40 species (Förderer et al., 2018). The Spermonde Archipelago, our region of interest, located in the middle of the Coral Triangle, hosts at least 26 LBF species, of which 17 species (11 genera) are dominant (Girard, Estradivari, et al., 2022; Renema, 2018) and upon which we based our study. At present, molecular and morphological communities are often seen as two different entities (Eqbal et al., 2022; Frontalini et al., 2018). Our work focuses on making those two communities (molecular and morphological) comparable and relatable to each other by correcting metabarcoding output data from mock samples using a quantitative metabarcoding approach (Figure 1). On the basis of the life cycle of foraminifera and their population dynamics and turnover rates (Hallock & Reymond, 2022), we hypothesise that LBF assemblage composition can be estimated from the number of mitochondrial reads in correlation to the taxa proportional biovolume in a sample rather than relative abundance of specimen counts. We assessed this hypothesis by using genus-specific calibration curves on our metabarcoding data to quantitatively estimate generic composition in sediment samples.

## 2 | MATERIALS AND METHODS

To develop a quantitative approach for metabarcoding data, we used an integrated approach that included several quantitative techniques for calibration and hypothesis testing (steps 1–2 and 3–6 in Figure 2, respectively). Step 1: We extracted morphometrics from 3D reconstructions of seven large benthic foraminifera species (LBF) to study the correlation between the surface area, the biovolume (unicellular volume of the specimen) and calcite volume (volume of an empty shell) and calcite weight. Step 2: We used droplet digital PCR on single specimens (single-cell) to study the relationship between the number of mitochondrial gene copies and the surface area, from which we calculated a genus-specific correction factor. Step 3: We measured the proportional surface area and relative abundance in the mock assemblages at the genus level. Step 4: We applied the correction factors to the number of reads from the mock assemblages to estimate the proportional surface area and relative abundance. Step 5: We tested



**FIGURE 1** Sediment metabarcoding of large benthic foraminifera (LBF) to quantitatively assess community composition for reef monitoring: An overview of the study. From a sediment sample, we compare and relate the relative number of sequence reads from DNA metabarcoding to morphological assessment (relative abundance and size), using a quantitative metabarcoding approach. The life cycle of foraminifera and theoretical population dynamics turnover of LBF (here *Neorotalia gaimardi*), after Hallock and Reymond (2022), suggest that the biovolume is likely better represented by the number of mitochondrial reads than the specimen counts, on which our hypothesis is based.



**FIGURE 2** Summary of the method workflow performed in this study (steps 1 to 6). Filled squares are methods and empty squares are data.

the accuracy of the quantitative metabarcoding method by comparing the genus relative abundance, the proportional surface area and the relative pre-corrected and post-corrected number of reads obtained from the mock assemblages. Step 6: We assessed the similarity between metabarcoding outputs of the mock and bulk assemblages to see whether bulk samples provide comparable results to the mock samples. The mock assemblages were created with specimens picked from a subsample of the bulk sediment, morphologically identified (step 3), pooled and metabarcoded (step 4). They served as mock communities to assess the accuracy of the developed approach (step 5) and during comparisons with bulk assemblages (step 6).

## 2.1 | Sample collection

Coral rubble, algae and/or sand samples were collected in cotton bags (18 x 32 cm) by scuba diving or snorkelling in the Spermonde Archipelago (South Sulawesi, Indonesia) in 2012 and 2013 for morphometrics measurements and in 2022 for calibration and

hypothesis testing (see [Table S1](#) for sampling details). For the LBF assemblage composition, we visited three islands (Pajenekang, Badi and Lumulumu), at which samples ( $n=35$ ) were collected along a depth gradient from the reef flat to the base of the reef slope. The coral rubble and algae were brushed to detach the foraminiferal community from its substrate. The sand and brushed material were split into two 8-mL falcon tubes (creating two subsamples). One subsample was used to morphologically assess the community and do community DNA on the pooled specimens (referred to as “mock” samples throughout); the second subsample was used for bulk-DNA analysis of unsieved sediment (referred to as “bulk” samples throughout). Both mock and bulk samples were used for metabarcoding. The samples were stored in ethanol 96% in a freezer until further steps. For the calibration and the morphometric measurements, specimens from seven species (*Amphisorus* SpL, *Amphistegina lessonii*, *Baculogypsinoidea spinosus*, *Calcarina spengleri*, *Heterostegina depressa*, *Neorotalia gaimardi* and *Operculina ammonoides*) were selected and picked from separate sediment samples (see [Table S1](#) for sampling details). We chose those seven species because (1) they represent the most important

taxonomic groups; (2) they represent different morphologies; (3) they have a smooth shell, which reduces the risk of co-amplification of extracellular DNA and facilitates morphometric measurements; and (4) they are the most abundant species in the study site. LBF are an informal grouping in multiple higher classified groups. Our species represent the porcelaneous Soritacea (*Amphisorus*) and the lamellar perforate families Calcarinidae (*Neorotalia*, *Calcarina*, *Baculogypsinoidea*), Nummulitidae (*Operculina*, *Nummulites*, *Heterostegina*) and Amphisteginidae (*Amphistegina*). These represent a coin-shaped annular genus (*Amphisorus*), trochospiral taxa (*Elphidium*, *Neorotalia*, *Calcarina*, *Amphistegina*), and planispiral coiling (*Operculina*, *Nummulites*, *Peneroplis*). In total, these represent about half of the genera reported from the Coral Triangle (see e.g., Renema (2018) and Förderer et al. (2018)). With those 11 genera, we cover 10 of the 11 most abundant species in the Spermonde Archipelago, based on previous studies (Girard, Estradivari, et al., 2022; Renema, 2018). All species except *H. depressa* were sampled in the Spermonde Archipelago. Because no fieldwork in Indonesia was possible between March 2020 and June 2022, *H. depressa* specimens were collected from the Indo-Pacific aquarium at Burger's zoo (Arnhem, the Netherlands) in April 2022 and kept alive until DNA extraction at MARUM (Bremen) to conduct preliminary work.

## 2.2 | Morphometric measurements using micro-CT scanning

All specimens ( $n=193$ ) were scanned using a Micro X-ray computed tomography scanner (Micro-CT) with a voxel resolution of 1.7909–7.5853  $\mu\text{m}$  (mean = 2.5  $\mu\text{m}$ ) at 80 kV (Zeiss Xradia 520 Versa, Germany) at Naturalis Biodiversity Center, the Netherlands (NBC) to quantify their volume. Using AVIZO Lite 3D software (version 2020.3.1, ThermoFisher Scientific, Waltham, MA, United States), we (1) took a screenshot of the reconstructed specimen as if lying flat under a microscope (surface area), (2) selected the shell by adjusting the contrast ratio (shell volume), (3) filled the empty space inside the shell, (4) shrunk and expended the inner volume to isolate the chambers (cell volume). Using ImageJ 1.53 (Wayne Rasband and contributors, National Institutes of Health, United States of America; <http://imagej.nih.gov/ij>), we measured the surface area of all specimens from the screenshot taken in step 1. Relationships between surface area, shell volume (calcite volume) and cell volume (biovolume) of the scanned specimens were non-linear and therefore calculated using a power law. The power law coefficients were determined from the linear regression of log-transformed variables. Additionally, the specimens of *Heterostegina depressa* were used as a model organism to describe the relationship between calcite volume and weight. Those specimens were weighed on a water-based microbalance reaching a precision of 0.001 mg (Sartorius GPC26-CW Precision Weigh Cell, Germany). A linear model was used to describe the relationship between calcite volume and weight.

## 2.3 | Quantification of COI gene copy number

All specimens ( $n=271$ ; 33 *Amphisorus* SpL, 24 *A. lessonii*, 32 *B. spinosus*, 35 *C. spengleri*, 82 *H. depressa*, 30 *N. gaimardi* and 35 *O. ammonoides*) were photographed using a microscope-mounted camera (Leica Microsystems, Wetzlar, Germany) and their smooth shell was brushed to remove extracellular DNA and other potential organisms living on the shell before DNA extraction. DNA was extracted from each specimen using the QIAamp DNA Micro Kit Tissue, using a modified protocol to enhance DNA retrieval (QIAGEN GmbH, Hilden, Germany). The specimens were transferred into individual 1.5 mL tubes and let to dry for 5 min and then crushed using a sterile pestle. The samples were lysed overnight into the AL buffer and carrier RNA to increase DNA yield (volume of reagents following the manufacturer's recommendation). The subsequent steps of the protocol were carried out following the manufacturer's instruction. The number of mitochondrial gene copies was quantified using the Droplet Digital PCR system (ddPCR) with the EvaGreen assay (Bio-Rad Laboratories, Inc.) using cytochrome c oxidase subunit 1 (COI) marker specific to Foraminifera (forward primer Foram\_COI\_fwd1 5'-GWGGWGTTAATGCTGGTYGAAC-3'; reverse primer Foram\_COI\_rev 5'-RWRCTTCWGGATGWCTAAGARATC-3') (Macher, Wideman, et al., 2021). With the precautions and protocol followed before DNA extraction and the choice of primers, we consider it unlikely that the number of COI gene copies measured during the ddPCR experiment was resulting from the co-amplification from extracellular DNA, smaller foraminifera, squatter or other eukaryotes (Girard, Macher, et al., 2022), and instead is reflecting the specimen's gene copy numbers.

The ddPCR reaction mix (22  $\mu\text{L}$ ) consisted of 11  $\mu\text{L}$  of QX200™ ddPCR™ EvaGreen Supermix (Bio-Rad Laboratories, Inc.), 1  $\mu\text{L}$  of each primer (10  $\mu\text{M}$ ), 7  $\mu\text{L}$  of RNase and DNase free-water and 2  $\mu\text{L}$  of DNA template diluted 1:100, with the exception of a few samples that required a 1:10 dilution to amplify. The QX200 Droplet Generator (Bio-Rad Laboratories, Inc.) was used to partition 20  $\mu\text{L}$  of the PCR reaction mix into droplets and samples were further amplified using a T100 Touch thermal cycler (Bio-Rad Laboratories, Inc.). Initial denaturation was performed at 95°C for 5 min, followed by 40 cycles at 95°C for 30s and at 54°C for 1 min, then a signal stabilization at 4°C for 5 min and at 90°C for 5 min and finally an infinite hold at 4°C. After amplification, droplets were analysed using the QX200 Droplet Reader (Bio-Rad Laboratories, Inc.). Threshold values for positive droplets were determined using the QuantaSoft software (version 1.7; Bio-Rad Laboratories, Inc.). The threshold for a positive signal was set based on a sample that showed good band separation, and droplets above that threshold were counted as positive events. For low DNA concentrations, count estimates for each sample were compared to the maximum confidence interval (95%) of the negative controls to determine if they were statistically different from zero. We judged technical replications not necessary after testing the variability between 3 replicates for 7 specimens (see Table S2). The range of deviation

between the replicates depended on the gene copy densities. For high numbers of gene copies ( $10^6$ – $10^7$ ), deviation between replicates was low (<10%); for specimens with very low numbers of gene copies ( $10^3$ ), the variability was much higher (>25%). Since the number of gene copies between replicates was within the same order of magnitude ( $10^x$ ), we considered that we reached satisfactory biological replication through the elevated number of samples processed in our work ( $n=271$ ).

## 2.4 | Morphological assessment of the mock samples

For the community morphological assessment of the mock samples, the sediment material was randomly spread on a petri dish. Under a stereo microscope, the petri dish was searched section by section and all LBF seen were picked out. For samples enriched in LBF, a maximum of 150 specimens were picked. This protocol was followed to mimic traditional methods for morphological assessment of living LBF assemblages. To isolate the living community from the sediment samples, only LBF that showed coloured endosymbionts were selected, which indicated that they were living at the time of sample collection. The specimens were morphologically identified based on the description from Macher, Prazeres, et al. (2021) and Renema (2018). However, many specimens could only be identified with certainty to the genus level, especially in the calcarinid, soritid and peneroplid groups. Photos grouping all specimens of a genus per sample were taken. Finally, all the specimens from a sample were pooled together for DNA extraction. The pooled specimens were not brushed before DNA extraction and it is therefore likely that smaller foraminifera and extracellular DNA co-amplified.

## 2.5 | DNA extraction and library preparation of mock and bulk samples

DNA extraction of the mock and bulk samples was performed using the NucleoSpin® Soil (Macherey-Nagel, Germany), using a modified protocol to enhance DNA retrieval. To improve the lysis, the samples were first dried overnight and crushed with a clean porcelain mortar and pestle. The powder resulting from the mock specimen pool was extracted at once. The powder from each bulk sample was divided equally in triplicates (A, B, C), using all or up to 500mg of material per replicate (maximum material weight according to the extraction protocol). To enhance the digestion of eukaryotic cell walls, we performed a chemical lysis step originally not included in the manufacturer protocol, in which only a mechanical lysis is the default. For this extra step, we added 50µL of Proteinase-K before the bead-beating step and incubated the samples at 37°C overnight in a thermomix after the bead-beating step. We followed the rest of the protocol as stated by the manufacturer. Since the mock samples were extracted only once, amplification in triplicates was performed on those samples (A, B, C) (Figure 2). We performed library preparation for a NovaSeq 6000 (Illumina, United States of America) sequencing run

using IDT10 tails and indexes (Integrated DNA Technologies, Inc., United States of America). The target region was the mitochondrial COI as used for the gene copy quantification (see method section above). The initial amplification and library preparation followed the steps of Girard, Macher, et al. (2022), with an initial amplification of 35 cycles instead of 40 to reduce potential amplification biases. The NovaSeq 6000 sequencing (250 paired-end reads) run was performed at BaseClear B.V. (Leiden, the Netherlands).

## 2.6 | Molecular data processing

We treated the demultiplexed data (referred to as 'raw data') using the VSEARCH-based software APSCALE (Advanced Pipeline for Simple yet Comprehensive AnalyEs) resulting in an exact sequence variant (ESV) table (Buchner et al., 2022). During the treatment, the following steps were performed with specific settings to the target marker (see Table S3 for details on algorithms, versions and settings): (1) sequence pairing and merging, (2) primer trimming, (3) sequence filtering based in length, (4) dereplication, (5) denoising into ESVs, also known as amplicon sequence variants (Callahan et al., 2017), and (6) quality filtering and chimeras removal. Additional details on the programs, algorithms and commands used at every step of the raw data processing up to the ESV table are described in Buchner et al. (2022). Samples with fewer than 1000 reads were disregarded. We checked the quality of the ESV table by filtering out ESVs with less than 0.1% of the total read number in that same sample (98.25% of reads retained), to correct for cross-contamination and tag switching (Cock et al., 2023; Di Muri et al., 2020). We assigned the ESVs to species level (at 99.4% ID) using megaBLAST (Version 2.13.0) (Morgulis et al., 2008) against a custom mitochondrial reference database for LBF from the Spermonde Archipelago region (identity threshold for ESV assignments and sequences were published by Girard, Macher, et al. (2022), see also Table S4). Because we are only interested in known large benthic foraminifera living in the Spermonde Archipelago (Girard, Estradivari, et al., 2022; Renema, 2018), any sequences not classified to species level or classified to other taxa than LBF were disregarded. This step ensured the removal of non-target smaller foraminifera that (potentially) co-amplified in the mock and bulk samples. Finally, we considered an ESV present in a sample only when it had been sequenced in at least two of the three biological and technical replicates to reduce the effect of index hopping (Costello et al., 2018; Farouni et al., 2020). To test the method reliably, we decided to merge the data to genus level for further analyses, because some specimens could not be taxonomically assigned to the species level with certainty.

## 2.7 | Comparing morphological and molecular data

We used the surface area as a variable indicative of foraminifera biovolume. The surface area of foraminifera specimens, which were used to quantify the number of mitochondrial gene copies, was measured using ImageJ. To define taxon-specific calibration coefficients, we determined the relationship between the number of gene copies

(millions) and the surface area ( $\text{mm}^2$ ) for each species. We tested a linear (Equation 1) and a logarithmic (Equation 2) model as follows:

$$\begin{aligned} \text{Number of gene copies (millions)} = \\ \text{Gene copy density (millions/mm}^2\text{)} * \text{surface area (mm}^2\text{)} \end{aligned} \quad (1)$$

$$\log(\text{Number of gene copies (millions)}) = a * \log(\text{surface area (mm}^2\text{)}) + b \quad (2)$$

For the linear model (Equation 1), the gene copy density was used for the model coefficient, which corresponds to the slope of the linear regression applied to non-transformed values of the number of gene copies and the surface area. The intercept was forced to zero, due to biological and physical limitations: a specimen with a surface area of  $0\text{mm}^2$  can only have 0 million gene copies. For the logarithmic model (Equation 2), the model coefficients  $a$  and  $b$  correspond to the slope and intercept, respectively, of the linear regression applied to the logarithmic values of the number of gene copies and the surface area. In this case, the intercept could not be forced to zero since the logarithmic value of 0 is undefined, and the value of zero can never be reached. To calculate robust model coefficients (*Gene copy density*,  $a$  and  $b$ ), we used bootstrapping, by subsampling the dataset to 30 specimens per species with 999 permutations (for species  $n > 30$ ). The mean of every coefficient was calculated. The average coefficients were further used as genus-specific biological correction factors applied to the number of sequence reads from the mock samples to validate the approach (see formulas in Table S5). We calibrated the mock data based on seven genera (see the method section 'Sample collection' for details on the choice of those seven genera). Four remaining genera (*Elphidium*, *Nummulites*, *Peneroplis* and *Sorites*) had no calibration coefficient. For those, we applied the calculated coefficients of the phylogenetically closest genus in our dataset out of the seven analysed (for the genus *Elphidium* we used the correction factors applied to the genus *Calcarina*; for *Nummulites* we used *Heterostegina*; for *Peneroplis* and *Sorites* we used *Amphisorus*). The equations to correct the number of reads for each genus with the linear model (Equations 1 and 3) and the logarithmic model (Equations 2 and 4) were applied as follows:

$$\text{Linearly postcorrected number of reads} = \frac{\text{Precorrected number of reads}}{\text{Gene copy density}} \quad (3)$$

$$\text{Logarithmically postcorrected number of reads} = \exp\left(\frac{\log(\text{Precorrected number of reads}) - b}{a}\right) \quad (4)$$

These results were compared to the proportional surface area and relative abundance for each genus present in a mock sample. The surface area was measured from group photos taken before DNA extraction using ImageJ. If a genus was absent from the mock fraction assessed morphologically and yielded a small number of reads, this number was assumed to come from remaining traces of index hopping or co-amplification of extracellular DNA and the read number was put to zero. For the comparisons, we used the relative and proportional values for the four data types (genus abundance,

surface area, pre- and post-corrected number of reads), which were calculated as follows:

$$\text{Genus relative value for a data type} = \frac{\text{Genus value for a data type}}{\text{sum of all values for a data type in that sample}} \quad (5)$$

## 2.8 | Statistical analysis

To decide which correction factors to use, we assessed the fit of the linear and logarithmic models on the ddPCR data by calculating the standard deviation of the mean after bootstrapping, the mean of the standard error and the mean of  $p$ -value for every coefficient. The rest of the statistical analyses was performed in comparisons to the post-corrected number of reads from the best fitting model. We compared the difference between the relative values of all combinations of data types from the mock samples (obtained with Equation 5) using pairwise  $t$ -tests to assess whether any of those data type combinations are significantly different from each other. For example, genus relative abundance values were tested against genus pre- and post-corrected relative number of reads and genus proportional surface area. Pairwise  $t$ -tests were performed in R, using the function `pairwise_t_test()` (R package *rstatix* version 0.7.2 (Kassambara, 2021)) with the  $p$ -value adjusted method 'bonferroni'.

We then assessed which of the two estimates derived from the mock metabarcoding output (pre- and post-corrected number of reads) were significantly more similar to the measured data types (relative abundance and proportional surface area). This allowed us to test whether the post-corrected relative number of reads provided better estimates of the genus relative abundance or the proportional surface area compared to the pre-corrected values. To do so, Euclidean distance matrices were calculated between all combinations of two data types. Means and variances of the calculated matrices were compared in pairs with the Welch Two Sample  $t$ -test and  $F$ -test performed in R, using the in-built functions `t.test()` and `var.test()`. The smaller the mean distance and the lower the variance, the more similar those two data types are to each other. If the null hypothesis was rejected, the alternative hypothesis stated that the value is significantly lower.

Furthermore, we verified for eventual biases, such as co-amplification of extracellular DNA, between the mock (sorted) and bulk (unsorted) samples to test whether manually sorted pools of specimens are comparable to the bulk sediment. The aim was to assess that replicates and metabarcoding outputs from mock and bulk samples within a sampling site are more similar to each other than between sampling sites. To do so, we compared the difference between the pre-corrected relative number of reads from bulk and mock samples. Additionally, we performed a non-metric multidimensional scaling plot (NMDS) and an Analysis of Similarity (ANOSIM) with the grouping for sampling sites and sample types (mock, bulk), using the functions `metaMDS()` with Bray-Curtis distances and `anosim()` (R package *vegan* version 2.6-4 (Oksanen et al., 2022)), respectively. An ANOSIM  $p$ -value  $< .01$  signifies that the groups compared are significantly different from each other, and the correlation coefficient  $R$  near 1 signifies that the groups

are spatially isolated in the ordination with little to no overlap and  $R$  near 0 means that the communities are very similar. The sum of the number of reads (uncorrected) at the genus level was compared between all pairs of replicates using pairwise  $t$ -tests, to assess whether any of the replicates were significantly different from each other. The tests were performed following the same function as stated above.

### 3 | RESULTS

#### 3.1 | Morphometrics and calculated correction factor in LBF

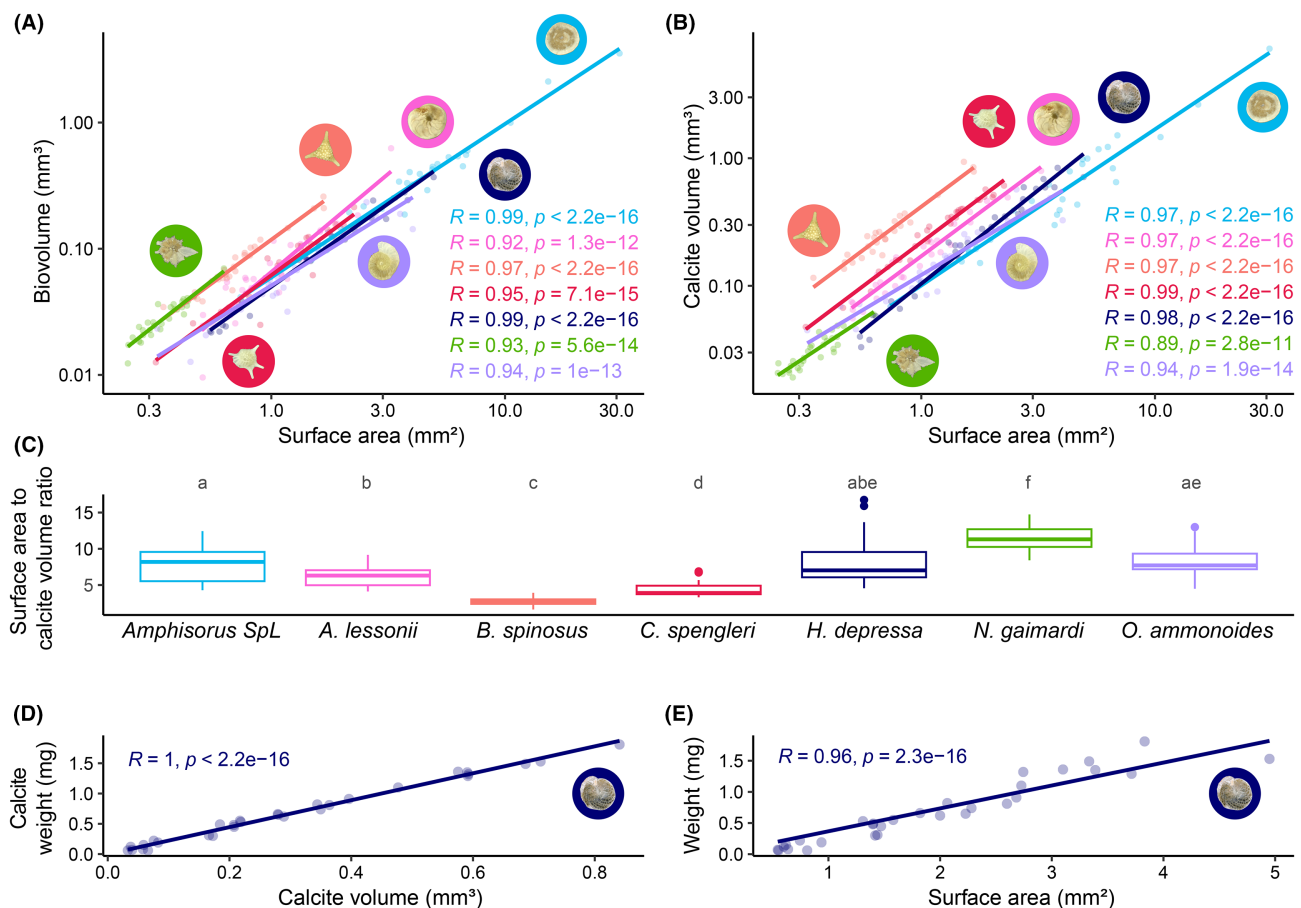
The relation of surface area ( $\text{mm}^2$ ) to biovolume ( $\text{mm}^3$ ) and calcite volume ( $\text{mm}^3$ ) for the seven species was equally strong (Figure 3A,B). The surface area to calcite volume ratio was significantly different between most species, except for *Amphisorus* SpL, *H. depressa* and *O. ammonoides* (Figure 3C). Additionally, the correlation between the surface area, calcite volume and calcite weight was almost perfect in *H. depressa* with  $R$  values  $>.96$  and  $p$ -values  $<.001$  (Figure 3D,E).

Those results showed that surface area is a strong proxy for biovolume and calcite volume, and further used as is.

We observed that the density (to surface area) of COI gene copy number varies between species, with *H. depressa* having the highest density at 5 million copies/ $\text{mm}^2$  on average and *N. gaimardi* and *O. ammonoides* the lowest at about 0.2 million copies/ $\text{mm}^2$  (Figure 4A). The fitted linear and logarithmic models show a positive relationship between the number of gene copies and the surface area (Figure 4B), most of them being highly significant (mean  $p$ -values  $<.01$ ; Table 1). Compared to the logarithmic model, the linear model resolved a larger proportion of the variation in COI gene copy number by surface area (40%–70% vs. 25%–65%) in more species (5 vs. 3); hence, in the remainder of the analysis, we used the correction factors based on the linear model (Table 1).

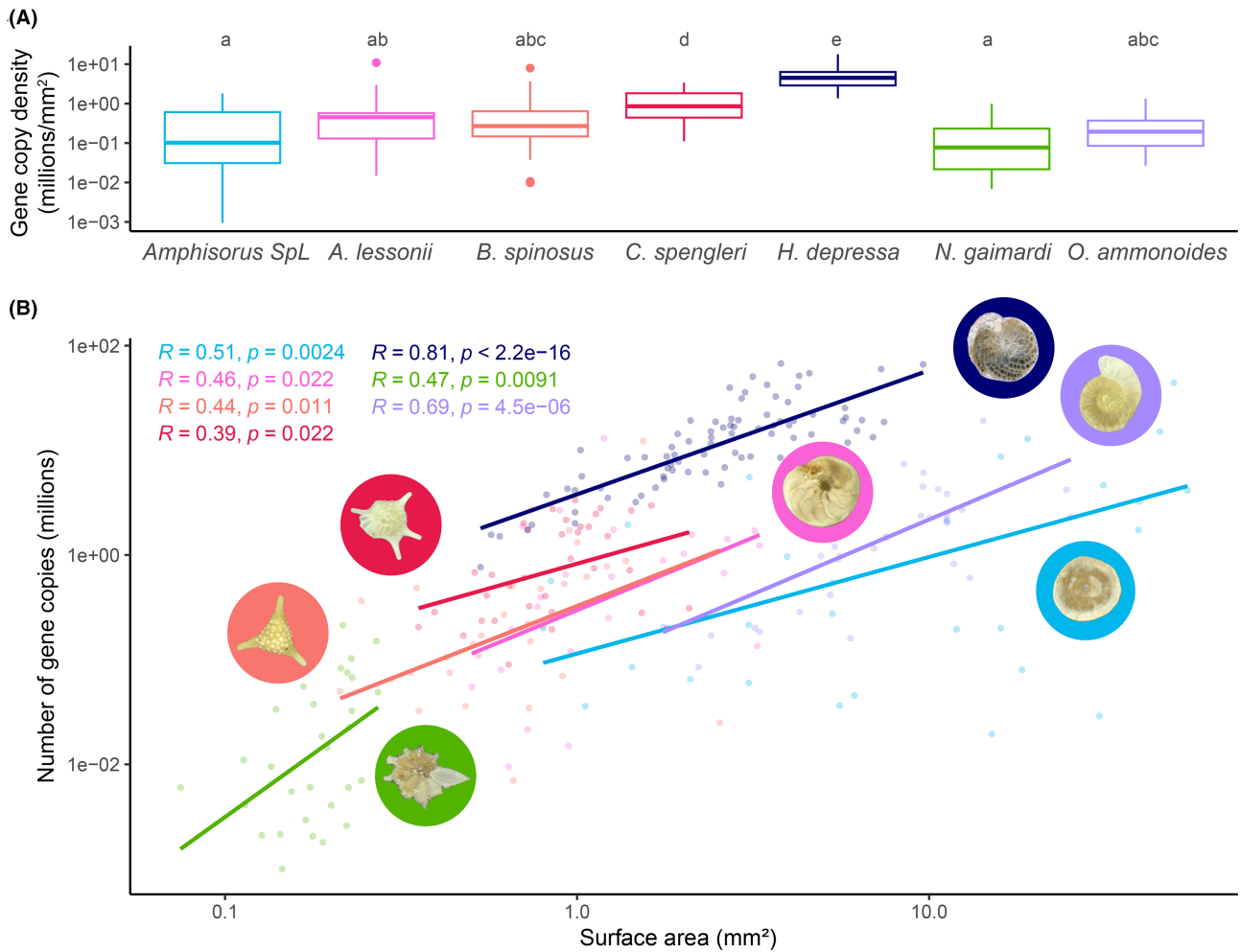
#### 3.2 | Accuracy of the metabarcoding data for the quantification of LBF relative abundance

The metabarcoding output (mock and bulk samples combined) totalled 532 ESVs (21,145,944 reads), of which 493 ESVs (92.7% of



**FIGURE 3** Logarithmic relations surface area ( $\text{mm}^2$ ) to biovolume ( $\text{mm}^3$ ) (A) and calcite volume ( $\text{mm}^3$ ) (B) for the seven species. Note that the values are not log-transformed, only the axes are displayed along a logarithmic scale. Surface area to calcite volume ratio (C) shows that the test shape is distinct between the different species (significance displayed with letters). Linear relationship surface area ( $\text{mm}^2$ ) (D) and the calcite volume ( $\text{mm}^3$ ) (E) to the calcite weight (mg) for the LBF species *Heterostegina depressa*. The  $R$  and  $p$ -values qualifying the relationships are displayed in the facets A, B, D and E. Significance is evaluated at  $p$ -value  $<.01$  in C.





**FIGURE 4** The gene copy density (A) range and median is displayed for the seven species. Significant differences ( $p$ -value < .01) in gene copy density between species are shown with letters above the boxplots. Relationship (B) between the number of gene copies (millions) and the specimen surface area (mm<sup>2</sup>) resulting from the ddPCR analysis, including all specimens. The  $R$  and  $p$ -values for each species are shown on the top left corner. Note: The axes are displayed along a logarithmic scale to highlight the variations between species with low number of gene copies (and densities).

ESVs, 99.2% of reads) were assigned to phylum Foraminifera, and 28 ESVs (5.3% of ESVs) were assigned to LBF at the species level, representing 84.4% of the total number of reads. Some differences appeared between the mock and the bulk samples, for example, 90.5% of the reads in the mock samples were assigned to LBF species contrary to 72.1% in the bulk samples (Table 2, see also results from sample processing and quality control in Table S3). During the morphological assessment, 11 genera were identified, all of which were detected in the molecular mock samples data. The use of correction factors, which accounted for the differences in gene copy density between genera during the corrections of the number of reads from the mock samples (Equations 3 and 4), did not significantly improve the accuracy of the mock metabarcoding output in estimating the proportional surface area taken by a genus (Welch Two Sample  $T$ -test:  $p$ -value = .3315,  $F$ -test:  $p$ -value = .5196) (Figures 5 and 6a). However, the corrections significantly improved the estimation of relative genus abundance by reducing the distance between the relative number of reads and the relative abundance within genera

and by increasing their correlation (Welch Two Sample  $T$ -test:  $p$ -value = .0171,  $F$ -test:  $p$ -value = .0134) (Figures 5 and 6b, Table S6).

Altogether, the four data types from the mock samples (proportional surface area, relative abundance, relative pre- and post-corrected number of reads) resulted in an almost identical assemblage composition with strong similarities between values of different data types at the same sampling site, which were supported by the low  $R$  value (ANOSIM, data types:  $p$ -value = .025 and  $R$  = .019, Figure 6) (see also pairwise  $t$ -test results in Table S7). The difference between the relative post-corrected number of reads and the proportional surface area only superficially improved to  $\pm 10\%$  on average; the difference between the proportion of relative post-corrected number of reads and the relative abundance was reduced to  $\pm 5\%$  on average (Figure 5). Despite the corrections, the proportional surface area of the soritids *Amphisorus* and *Sorites* remained generally underestimated, which is not the case for the relative abundance. For the rare genus *Baculogypsinooides*, found only on the mid-slope of Pajenekang, no improvements were observed.

TABLE 1 Model coefficients ('gene copy density', 'a', 'b') calculated after bootstrapping ( $n=30$ ). Linear model (Equation 1) and logarithmic model (Equation 2).

Genus	Linear model (Equation 1)				Logarithmic model (Equation 2)					
	Mean gene copy density	Sd gene copy density	Mean $R^2$	Mean $p$ -value	Mean a	Sd a	Mean b	Sd b	Mean $R^2$	Mean $p$ -value
<i>Amphisorus</i>	0.372	0.039	.530	.000**	0.924	0.082	-2.172	0.189	.262	.005**
<i>Amphistegina</i>	0.659	NA <sup>†</sup>	.157	.050*	1.386	NA <sup>†</sup>	-1.214	NA <sup>†</sup>	.215	.022*
<i>Baculogypsinoides</i>	1.231	0.222	.230	.007**	1.313	0.203	-1.120	0.145	.197	.019*
<i>Calcarina</i>	1.135	0.091	.592	.000**	0.949	0.199	-0.187	0.086	.154	.049*
<i>Heterostegina</i>	5.435	0.715	.717	.000**	1.184	0.142	1.341	0.131	.646	.000**
<i>Neorotalia</i>	0.203	NA <sup>†</sup>	.428	.000**	2.412	NA <sup>†</sup>	-0.210	NA <sup>†</sup>	.219	.009**
<i>Operculina</i>	0.286	0.034	.461	.000**	1.423	0.112	-2.493	0.225	.475	.000**

Note: Sd' stands for standard deviation from the mean. See Table S5 for additional information.

<sup>†</sup>NA, no standard deviation from the mean for *A. lessonii* ( $n=24$ ) and *N. gaimardi* ( $n=30$ ), because bootstrapping resampled the whole dataset for those two species.

\*Significance at  $p$ -value < .05. \*\*High significance at  $p$ -value < .01.

Sample type		Number of ESVs (%)	Proportion of reads (%)
Mock samples	Total sequences after quality control	117 (100%)	100
	Total foraminifera sequences (>75% ID)	113 (96.6%)	99.9
	Foraminifera assigned to species level (>94.4% ID)	30 (25.6%)	90.6
	Species assigned to LBF	27 (23.1%)	90.5
Bulk samples	Total sequences after quality control	495 (100%)	100
	Total foraminifera sequences (>75% ID)	457 (92.3%)	97.5
	Foraminifera assigned to species level (>94.4% ID)	54 (10.9%)	77.3
	Species assigned to LBF	26 (5.3%)	72.1

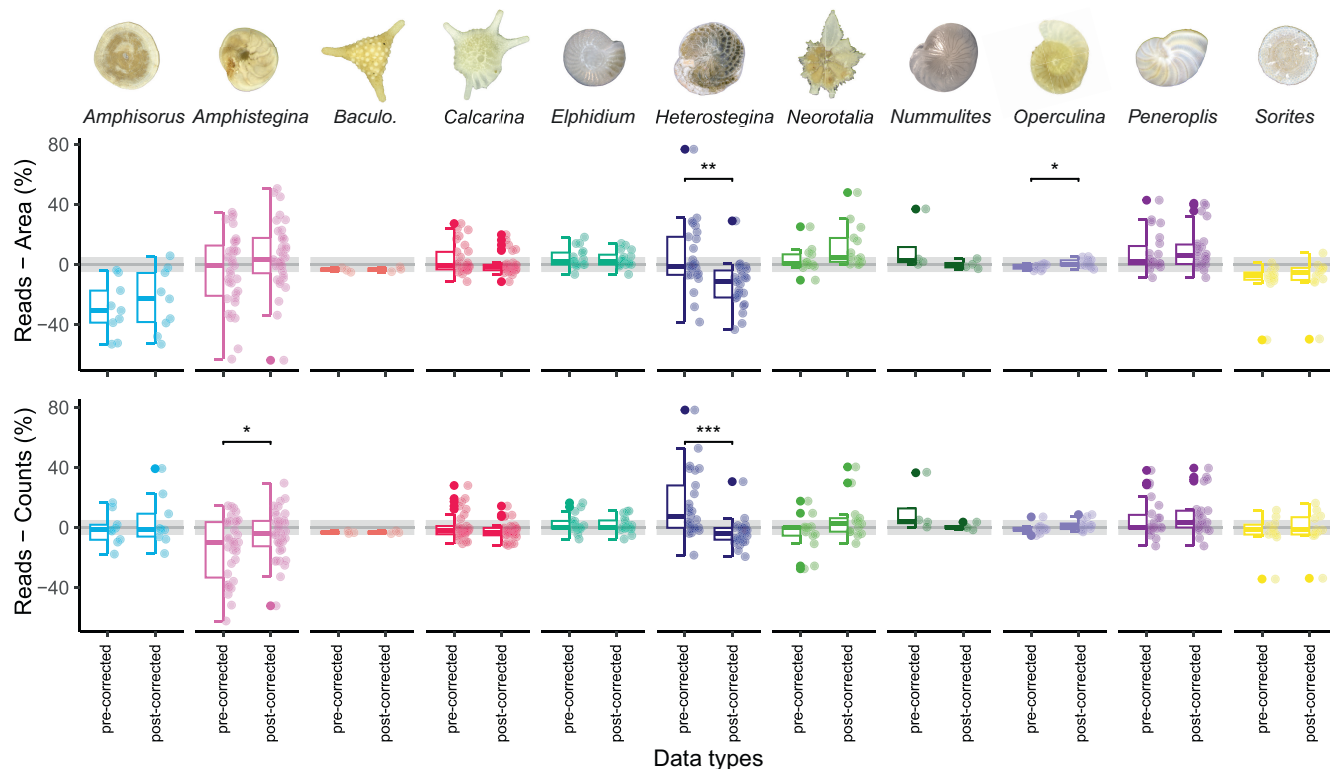
TABLE 2 From quality-controlled dataset to working dataset.

### 3.3 | Congruence of the metabarcoding outputs between mock and bulk samples

The pre-corrected datasets of the mock and the bulk samples were very similar with no statistical difference based on the ANOSIM results (sample types:  $p$ -value > .05 and  $R$  < .05) (Figure 7a-c) and the replicates were not significantly different based on the pairwise  $t$ -test results (Table S8). The analyses have shown that there was a statistically significant difference between the sampling sites with an  $R$  value above .5 (ANOSIM, sampling sites:  $p$ -value = .001 and  $R$  > .5) (Figure 7a-c). In a few cases, we observed certain differences in genus relative abundance based on the relative number of reads, for example, *Sorites* was very abundant at shallow depths by Pajenekang in the bulk sample but not as abundant in the mock sample (Figure S1). Nevertheless, the genus richness was the same on average between the bulk and the mock samples (Figure 7d).

## 4 | DISCUSSION

We tested the accuracy of quantitatively assessing large benthic foraminifera (LBF) community composition from metabarcoding data to extract proportional surface area, a good proxy for biovolume, and the relative abundance in coral reef sediment from Indonesia. We expected that the relative number of reads can estimate more precisely the proportional surface area of LBF taxa rather than the traditional relative abundance, because mitochondria are generally more abundant in larger specimens and therefore correlated to size. Contrary to our hypothesis, the corrections on the number of reads significantly improved the estimations of relative abundance, by reducing the difference between the measured and estimated relative abundance to  $\pm 5\%$  on average, but only superficially reduced the difference between the measured and estimated proportional surface area, with a remaining difference of  $\pm 10\%$  in abundant genera. In most cases, the estimations for rare taxa and taxa in low abundance did not improve with any of the corrections. Similar outcomes were

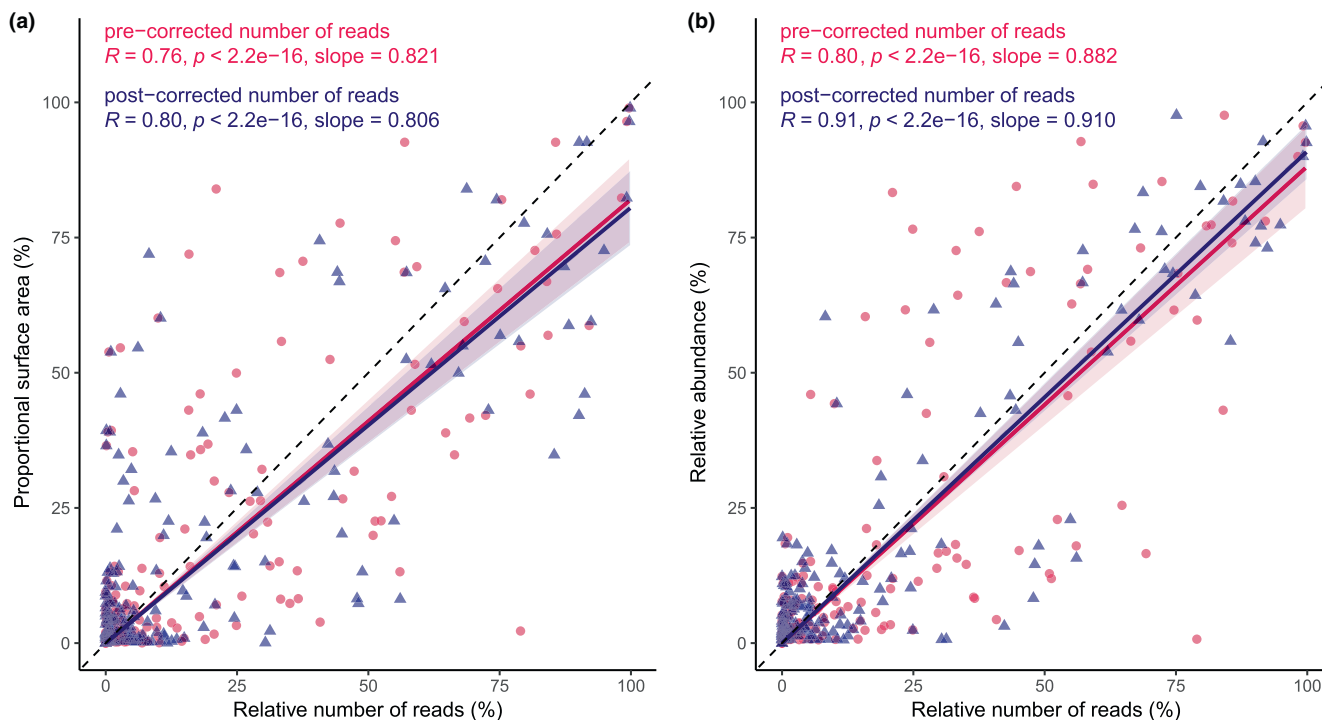


**FIGURE 5** Difference in proportions between the data types (relative pre-corrected and post-corrected reads) and the proportional surface area and the relative abundance at the genus level (top: Relative number of reads minus proportional surface area; bottom: Relative number of reads minus relative abundances). The grey-shaded region shows the  $\pm 5\%$  difference zone. The difference shows how far from the morphological assessment the proportions per genus per sample are. Negative values indicate an underestimation of the proportion by the number of reads and positive values indicate an overestimation. *T*-tests were performed to assess significance between the data types (pre- and post-corrected number of reads); significance is indicated by one ( $p$ -value  $< .05$ ), two ( $p$ -value  $< .01$ ) or three asterisks ( $p$ -value  $< .001$ ).

described in other recent studies, where dominant species drove the correlation between population size and number of reads (Martin et al., 2022; Skelton et al., 2022). Accurately retrieving quantitative information on species assemblage composition from a metabarcoding dataset depends also on the complexity of the community structure, in addition to inherent biological and technical biases (Bell et al., 2019; Piñol et al., 2019). In other words, the more species the less accurate the approach might be. This phenomenon is in accordance with our data, because, in reef sediments, LBF assemblages are often dominated by one to two species, whereas species richness can reach more than 40 species across Indo-Pacific reefs (Förderer et al., 2018). In light of these results, one could argue that raw number of reads, without corrections, is sufficient to roughly estimate proportional biovolume and calcite volume, as the corrections did not improve the results significantly, providing clear general trends. The corrections were however essential to obtain accurate estimates of relative abundance and to make molecular data comparable to past research for which solely traditional counting methods were used, with a conversion of counts to relative abundances.

Further on, similarly to morphological community assessment, metabarcoding methods can also be prone to false positives and negatives. For instance, false positives can occur due to

tag-switching, cross-contamination and chimaeras in spite of all the efforts during the experiments (Bell et al., 2019; Esling et al., 2015; Ficetola et al., 2016). Likewise, false negatives can occur by choosing inadequate primer sets (Krehenwinkel et al., 2017; Piñol et al., 2019), due to heterogeneity of samples before division or PCR stochasticity. Those effects lead to discrepancies between community composition of control samples (e.g., the mock samples) and environmental samples (e.g., the bulk samples). For that reason, we used recently developed degenerate primers to target the mitochondrial Leray region in all groups of Foraminifera (Girard, Macher, et al., 2022; Macher et al., 2022), and primers with a level of degeneracy that have been shown to reduce amplification bias to some degree (Elbrecht & Leese, 2017; Krehenwinkel et al., 2017; Marquina et al., 2019). Therefore, the small differences, sometimes larger than others, observed between the mock and bulk samples in our study, partly arose due to heterogeneous subsample division and remaining biases inherent to the PCR process. The greater presence of extracellular DNA present in bulk sediment samples, compared to the mock samples, added a degree of variability and participated in the differences observed between the bulk and the mock communities. Extracellular DNA comes partly from decomposing cells or excretion from living cells (Nielsen et al., 2007), and its degradation



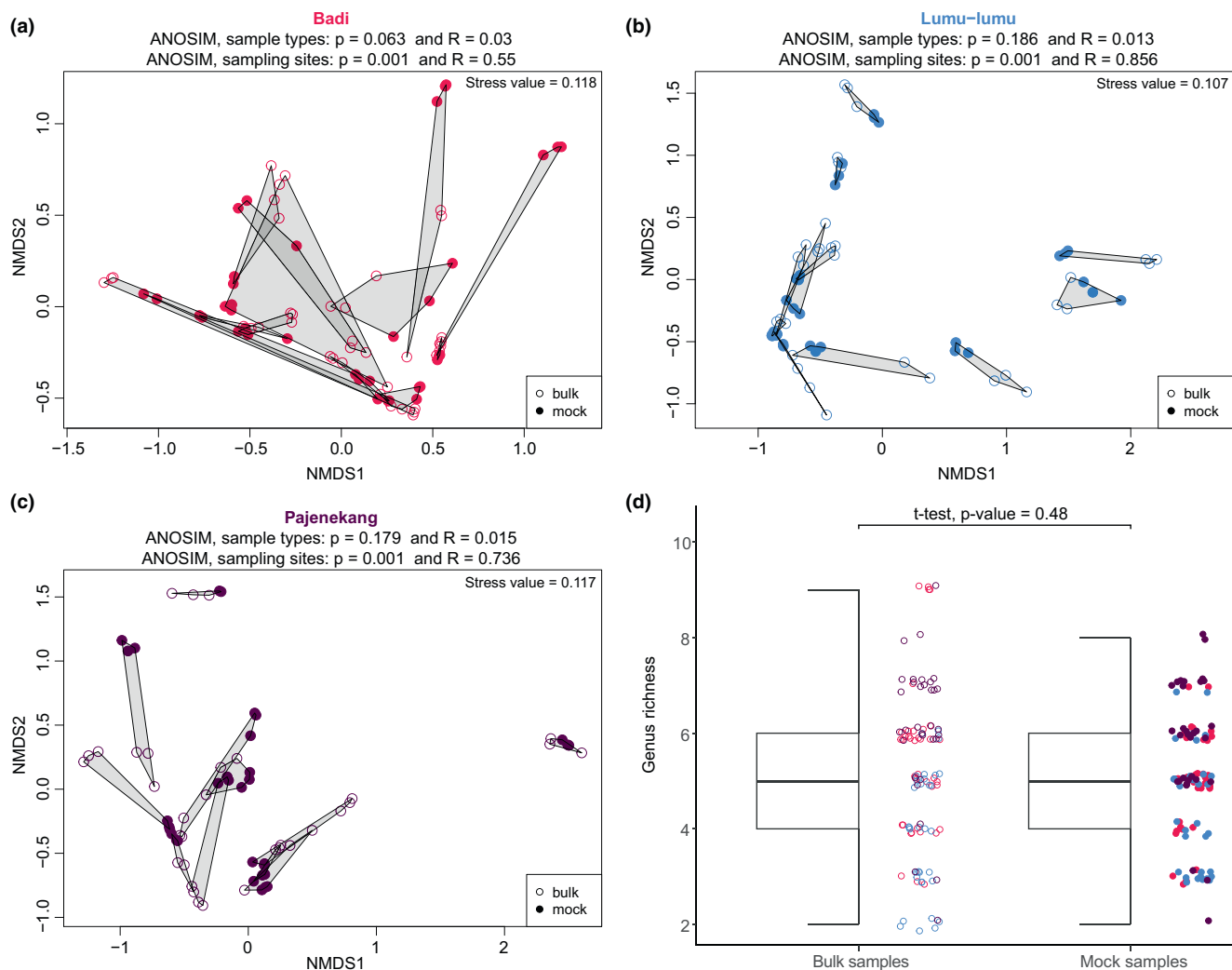
**FIGURE 6** Comparison between the four data types obtained from the mock samples. The proportional surface area (a) and the relative abundance (b) are compared to the relative pre-corrected (red) and post-corrected (dark blue) number of reads. ANOSIM, sampling sites:  $p$ -value = .001,  $R = .653$  and data types:  $p$ -value = .025,  $R = .019$  (see also pairwise  $t$ -test in Table S7).

rate is highly variable and depends on the environmental conditions (Corinaldesi et al., 2008; Torti et al., 2015). The extracellular DNA pool in marine sediments has been characterized by highly diverse sequences of low abundance (Corinaldesi et al., 2008), which were mostly filtered out during data processing and quality control steps. The remaining reads associated with extracellular DNA probably had little effect on the dataset when used for monitoring and LBF community composition. However, it led to non-significant, small differences in taxa richness and read abundances, as well as proportionality. The presence and amplification of extracellular DNA will have a greater impact and should be carefully considered when accounting for rare taxa and smaller foraminifera.

We examined community composition at the genus level due to unresolved morphological and molecular taxonomic understanding in certain genera, especially for *Calcarina*, *Peneroplis* and *Sorites* (Renema, 2010). Other recent studies used a quantitative approach on different groups (e.g., plankton, diatoms, arthropods, fish larvae, marine mammal) at the species (Ershova et al., 2023; Thomas et al., 2016; Vasselon et al., 2018), family (Ratcliffe et al., 2021), order (Krehenwinkel et al., 2017) or higher taxon level (Ershova et al., 2023; Martin et al., 2022) despite several orders of magnitude variation in gene copy numbers. Those studies resolved some technical and biological metabarcoding biases and stated that it generally provided a more accurate assessment of taxa community composition for proportional biomass, biovolume, surface area and relative abundance, compared to uncorrected metabarcoding outputs. Hence, corrections of the data allowed for a more comprehensive interpretation of metabarcoding data with the use of quantitative

information to answer ecological questions, similarly to interpretations based on morphological (directly observed) datasets.

In foraminifera, with an adequate reference database, species-level assessment of the community is possible with quantitative metabarcoding, admitting the marker used resolves all species. The nuclear SSU region (18S rRNA gene) is unique to foraminifera species studied thus far with molecular techniques and has an extensive reference database (Barrenechea Angeles et al., 2024; Guillou et al., 2013; Pawlowski & Holzmann, 2014), with 4442 available reference sequences in the PR2 database (<https://app.pr2-database.org/pr2-database/>). However, this marker includes hypervariable regions producing intra-specimen genetic variability with a high number of replicates in the nuclear genome (Girard, Langerak, et al., 2022; Weber & Pawlowski, 2014), which makes the use of quantitative information from metabarcoding outputs difficult (but see Weber and Pawlowski (2013)). On the contrary, the mitochondrial genome was recently amplified (Macher et al., 2023), and has a very limited reference database (Girard, Langerak, et al., 2022; Macher et al., 2022), with 209 sequences and 75 species (<https://www.ncbi.nlm.nih.gov/>). The mitochondrial marker used in this study (COI of the Leray region, after Macher, Wideman, et al. (2021)) is too conserved to resolve species in all foraminiferal groups, but a longer or alternative region in the mitochondrial genome could offer such resolution. It however offers the possibility to assess communities quantitatively similar to relative abundance, as shown in our study. With the growing and affordable accessibility to sequencing technologies in the last years, a multiple marker approach, for example, by combining the nuclear SSU and the mitochondrial COI



**FIGURE 7** Comparison between the community composition obtained from the relative pre-corrected number of reads from bulk samples (empty circles) and the mock samples (full circles) displayed in an NMDS at Badi (a), Lumu-lumu (b) and Pajenekang (c) islands. The ANOSIM was performed on two levels: Grouping sampling sites (significantly different at all islands) and grouping sample types (not significantly different at all islands). (d) The boxplots show the genus richness between the two sample types. The dot colours are associated with the island (red = Badi, blue = Lumu-lumu, purple = Pajenekang).

markers would offer the fine-scale diversity of the assemblage and the quantitative information. Assessing taxon-specific community to species level is essential when monitoring single species known to be bioindicators for pollutants or other marginal environmental conditions (Dean, 2008; Frontalini et al., 2018; Girard, Estradivari, et al., 2022; Jaanus et al., 2009).

Additionally, proportion estimations to proportional biovolume, calcite volume and calcite weight in the sediment can be extrapolated from the number of reads, since our results also demonstrated a very strong correlation of those three metrics to surface area. Estimates of absolute calcite volume and weight could be obtained using quantitative metabarcoding by measuring number of gene copies in environmental samples on the ddPCR as well as weighing bulk and mock samples for controls and with a sampling design with standardized volume of collected sediment and surface area sampled. This highlights the application of the quantitative metabarcoding approach to monitoring, among others, the budget of carbonate standing crops

on (sub)tropical carbonate shelves. For this purpose, the mitochondrial marker used in our study is suitable, as the taxonomic resolution in metabarcoding datasets does not need to reach species level. In fact, some of the key variables to quantify carbonate production in reefs are foraminiferal test shapes, census counts (for test density) and turnover rates (life history) (Narayan et al., 2022). Test shapes are very distinct between genera (Renema, 2018), the same resolution at which our method was developed. On average LBF have comparable carbonate production rates to corals, coralline algae and macrobenthos (Hallock, 1981; Narayan et al., 2022). This raises the question of how meaningful the information about foraminiferal carbonate standing crops, in addition to community composition, can be as an indication of reef health, since both separately (i.e., foraminifera and carbonate budgets) have been suggested as indicators for coral growth and reef health (Girard, Estradivari, et al., 2022; Lange et al., 2020; Narayan et al., 2022; Prazeres et al., 2020, and references therein).

By using single-cell quantitative PCR, Micro-CT scanning and metabarcoding methods, we demonstrated that quantitative information using the mitochondrial marker can be retrieved to accurately estimate living LBF community composition, in terms of relative abundance and proportional surface area (a proxy for biovolume), from metabarcoding data. This quantitative approach allows for comparison of past morphological datasets with future molecular ones. Our results also highlight the application of our quantitative approach to support reef monitoring operations by capturing fine-scale processes, such as depth gradients, seasonal and anthropogenic impacts on communities, that require high throughput sampling treatment. With our method, a census-based approach that requires a deeper taxonomic knowledge is not essential to assess LBF communities. However, in regions where the foraminifera community is not well known, combining molecular and morphological techniques for quantifying community composition in foraminifera is still recommended to improve the sequence reference database and eventually add missing taxa.

#### AUTHOR CONTRIBUTIONS

EBG, RM and WR designed the research. EBG and RM performed the explorative work. EBG, WR and AMAP conducted fieldwork and collected the samples. EBG, EAD and CR performed the laboratory work (from morphological assessment and DNA extraction to metabarcoding data). EBG conducted the data analysis and wrote the first draft of the manuscript. All authors participated in the improvement of the manuscript until its present form.

#### ACKNOWLEDGEMENTS

We thank Max Janse for the sediment samples from Burger's Zoo Aquarium. We thank Tisja Meijers for her help as technician at the Naturalis Biodiversity Center molecular laboratory. We also thank Jan-Niklas Macher and Frithjof Ehlers for their support with data processing and analysis. We thank the Indonesian authorities for approving the collection of reef sediment samples (permit holder: Elsa Girard, permit number: 2C11FB0145-W). This study was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (4D-REEF, No 813360). RM was supported by the Cluster of Excellence 'The Ocean Floor—Earth's Uncharted Interface' (EXC-2077, Project 390741603) funded by the German Research Foundation (DFG).

#### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

#### OPEN RESEARCH BADGES



This article has earned Open Data, Open Materials and Preregistered Research Design badges. Data, materials and the preregistered design and analysis plan are available at [<https://github.com/EBGirard/QuantMetaForam>].

#### DATA AVAILABILITY STATEMENT

All R scripts and datasets are available at <https://github.com/EBGirard/QuantMetaForam>. Raw sequence reads and metadata were deposited in the Sequence Read Archive – SRA (BioProject PRJNA1035292).

#### BENEFIT-SHARING STATEMENTS

A research collaboration was developed with scientists from Indonesia providing genetic samples, with collaborators active in fieldwork included as co-authors. The results of research have been shared with the provider communities. Benefits from this research accrue from the sharing of our data and results on public databases as described above.

#### ORCID

Elsa B. Girard  <https://orcid.org/0000-0001-5785-542X>

#### REFERENCES

- Barrenechea Angeles, I., Nguyen, N.-L., Greco, M., Tan, K. S., & Pawlowski, J. (2024). Assigning the unassigned: A signature-based classification of rDNA metabarcodes reveals new deep-sea diversity. *PLoS One*, *19*, e0298440.
- Bell, K. L., Burgess, K. S., Botsch, J. C., Dobbs, E. K., Read, T. D., & Brosi, B. J. (2019). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology*, *28*, 431–455.
- Brinkmann, I., Schweizer, M., Singer, D., Quinchar, S., Barras, C., Bernhard, J. M., & Filipsson, H. L. (2023). Through the eDNA looking glass: Responses of fjord benthic foraminiferal communities to contrasting environmental conditions. *Journal of Eukaryotic Microbiology*, *70*, e12975.
- Buchner, D., Macher, T.-H., & Leese, F. (2022). APSCALE: Advanced pipeline for simple yet comprehensive analyses of DNA metabarcoding data. *Bioinformatics*, *38*, 4817–4819.
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*, 2639–2643.
- Cavaliere, M., Barrenechea Angeles, I., Montresor, M., Bucci, C., Brociani, L., Balassi, E., Margiotta, F., Francescangeli, F., Bouchet, V. M. P., Pawlowski, J., & Frontalini, F. (2021). Assessing the ecological quality status of the highly polluted Bagnoli area (Tyrrhenian Sea, Italy) using foraminiferal eDNA metabarcoding. *Science of the Total Environment*, *790*, 147871.
- Cock, P. J. A., Cooke, D. E. L., Thorpe, P., & Pritchard, L. (2023). THAPBI PICT—a fast, cautious, and accurate metabarcoding analysis pipeline. *PeerJ*, *11*, e15648.
- Cordier, T., Alonso-Sáez, L., Apothéoz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*, 2937–2958.
- Corinaldesi, C., Beolchini, F., & Dell'Anno, A. (2008). Damage and degradation rates of extracellular DNA in marine sediments: Implications for the preservation of gene sequences. *Molecular Ecology*, *17*, 3939–3951.
- Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferreira, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., & Gabriel, S. (2018). Characterization and remediation of sample index swaps

- by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, 19, 332.
- Dawson, J. L., Smithers, S. G., & Hua, Q. (2014). The importance of large benthic foraminifera to reef Island sediment budget and dynamics at Raine Island, northern great barrier reef. *Geomorphology*, 222, 68–81.
- Dean, H. K. (2008). The use of polychaetes (Annelida) as indicator species of marine pollution: A review. *Revista de Biología Tropical*, 56, 11–38.
- Di Muri, C., Lawson Handley, L., Bean, C. W., Li, J., Peirson, G., Sellers, G. S., Walsh, K., Watson, H. V., Winfield, I. J., & Hänfling, B. (2020). Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. *Metabarcoding and Metagenomics*, 4, e56959.
- Doo, S. S., Hamylton, S., Finfer, J., & Byrne, M. (2017). Spatial and temporal variation in reef-scale carbonate storage of large benthic foraminifera: A case study on one tree reef. *Coral Reefs*, 36, 293–303.
- Downs, C. A., Woodley, C. M., Richmond, R. H., Lanning, L. L., & Owen, R. (2005). Shifting the paradigm of coral-reef "health" assessment. *Marine Pollution Bulletin*, 51, 486–494.
- Elbrecht, V., & Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers of Environmental Science & Engineering in China*, 5, 11. <https://doi.org/10.3389/fenvs.2017.00011>
- Eqbal, A.-E., Fabio, F., Eszter, B., Sandra, B., Shaker, A.-H., Fadila, S., Ahmad, B. A., Jan, P., & Fabrizio, F. (2022). Benthic foraminifera as proxies for the environmental quality assessment of the Kuwait Bay (Kuwait, Arabian gulf): Morphological and metabarcoding approaches. *Science of the Total Environment*, 833, 155093.
- Ershova, E. A., Wangensteen, O. S., & Falkenhaus, T. (2023). Mock samples resolve biases in diversity estimates and quantitative interpretation of zooplankton metabarcoding data. *Marine Biodiversity*, 53, 66.
- Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, 43, 2513–2524.
- Farouni, R., Djambazian, H., Ferri, L. E., Ragoussis, J., & Najafabadi, H. S. (2020). Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing. *Nature Communications*, 11, 2704.
- Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, 16, 604–607.
- Förderer, M., Rödder, D., & Langer, M. R. (2018). Patterns of species richness and the center of diversity in modern Indo-Pacific larger foraminifera. *Scientific Reports*, 8, 8189.
- Friedman, J. R., & Nunnari, J. (2014). Mitochondrial form and function. *Nature*, 505, 335–343.
- Frontalini, F., Cordier, T., Balassi, E., Arminot du Chatelet, E., Cermakova, K., Apothéoz-Perret-Gentil, L., Martins, M. V. A., Bucci, C., Scantamburlo, E., Treglia, M., Bonamin, V., & Pawlowski, J. (2020). Benthic foraminiferal metabarcoding and morphology-based assessment around three offshore gas platforms: Congruence and complementarity. *Environment International*, 144, 106049.
- Frontalini, F., Greco, M., Di Bella, L., Lejzerowicz, F., Reo, E., Caruso, A., Cosentino, C., Maccotta, A., Scopelliti, G., Nardelli, M. P., Losada, M. T., Arminot du Châtelet, E., Coccioni, R., & Pawlowski, J. (2018). Assessing the effect of mercury pollution on cultured benthic foraminifera community using morphological and eDNA metabarcoding approaches. *Marine Pollution Bulletin*, 129, 512–524.
- Gielings, R., Fais, M., Fontaneto, D., Creer, S., Costa, F. O., Renema, W., & Macher, J.-N. (2021). DNA Metabarcoding methods for the study of marine benthic Meiofauna: A review. *Frontiers in Marine Science*, 8, 730063. <https://doi.org/10.3389/fmars.2021.730063>
- Girard, E. B., Estradivari, Ferse, S., Ambo-Rappe, R., Jompa, J., & Renema, W. (2022). Dynamics of large benthic foraminiferal assemblages: A tool to foreshadow reef degradation? *Science of the Total Environment*, 811, 151396.
- Girard, E. B., Langerak, A., Jompa, J., Wangensteen, O. S., Macher, J.-N., & Renema, W. (2022). Mitochondrial cytochrome oxidase subunit 1: A promising molecular marker for species identification in foraminifera. *Frontiers in Marine Science*, 9, 809659. <https://doi.org/10.3389/fmars.2022.809659>
- Girard, E. B., Macher, J.-N., Jompa, J., & Renema, W. (2022). COI metabarcoding of large benthic foraminifera: Method validation for application in ecological studies. *Ecology and Evolution*, 12, e9549.
- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., Del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., ... Christen, R. (2013). The Protist ribosomal reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 41, D597–D604.
- Hallock, P. (1981). Production of carbonate sediments by selected large benthic foraminifera on two Pacific coral reefs. *Journal of Sedimentary Research*, 51, 467–474.
- Hallock, P., Lidz, B. H., Cockey-Burkhard, E. M., & Donnelly, K. B. (2003). Foraminifera as bioindicators in coral reef assessment and monitoring: The ForAM index. *Environmental Monitoring and Assessment*, 81, 221–238.
- Hallock, P., & Reymond, C. E. (2022). Contributions of Trimorphic life cycles to dispersal and evolutionary trends in large benthic foraminifera. *Journal of Earth Science*, 33, 1425–1433.
- Hassan, S., Sabreena, Pocza, P., Ganai, B. A., Almalki, W. H., Gafur, A., & Sayyed, R. Z. (2022). Environmental DNA metabarcoding: A novel contrivance for documenting terrestrial biodiversity. *Biology*, 11, 1297. <https://doi.org/10.3390/biology11091297>
- Hughes, T. P., Barnes, M. L., Bellwood, D. R., Cinner, J. E., Cumming, G. S., Jackson, J. B. C., Kleypas, J., van de Leemput, I. A., Lough, J. M., Morrison, T. H., Palumbi, S. R., van Nes, E. H., & Scheffer, M. (2017). Coral reefs in the Anthropocene. *Nature*, 546, 82–90.
- Humphreys, A. F., Purkis, S. J., Wan, C., Aldrich, M., Nichols, S., & Garza, J. (2022). A new foraminiferal bioindicator for long-term heat stress on coral reefs. *Journal of Earth Science*, 33, 1451–1459. <https://doi.org/10.1007/s12583-021-1543-7>
- Jaanus, A., Toming, K., Hällfors, S., Kaljurand, K., & Lips, I. (2009). Potential phytoplankton indicator species for monitoring Baltic coastal waters in the summer period. In J. H. Andersen & D. J. Conley (Eds.), *Eutrophication in coastal ecosystems: Towards better understanding and management strategies selected papers from the second international symposium on research and Management of Eutrophication in coastal ecosystems, 20–23 June 2006, Nyborg, Denmark* (pp. 157–168). Springer.
- Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. R package version 0.7.2. <https://rpkgs.datanovia.com/rstatix/>
- Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, 7, 17668.
- Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, 28, 420–430.
- Lange, I. D., Perry, C. T., & Alvarez-Filip, L. (2020). Carbonate budgets as indicators of functional reef "health": A critical review of data underpinning census-based methods and current knowledge gaps. *Ecological Indicators*, 110, 105857.
- Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (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, 34.

- Macher, J.-N., Bloska, D. M., Holzmann, M., Girard, E. B., Pawlowski, J., & Renema, W. (2022). Mitochondrial cytochrome c oxidase subunit I (COI) metabarcoding of foraminifera communities using taxon-specific primers. *PeerJ*, *10*, e13952.
- Macher, J.-N., Coots, N. L., Poh, Y.-P., Girard, E. B., Langerak, A., Muñoz-Gómez, S. A., Sinha, S. D., Jirsová, D., Vos, R., Wissels, R., Gile, G. H., Renema, W., & Wideman, J. G. (2023). Single-cell genomics reveals the divergent mitochondrial genomes of Retaria (Foraminifera and Radiolaria). *MBio*, *14*, e0030223.
- Macher, J.-N., Prazeres, M., Taudien, S., Jompa, J., Sadekov, A., & Renema, W. (2021). Integrating morphology and metagenomics to understand taxonomic variability of Amphisorus (Foraminifera, Miliolida) from Western Australia and Indonesia. *PLoS One*, *16*, e0244616.
- Macher, J.-N., Wideman, J. G., Girard, E. B., Langerak, A., Duijm, E., Jompa, J., Sadekov, A., Vos, R., Wissels, R., & Renema, W. (2021). First report of mitochondrial COI in foraminifera and implications for DNA barcoding. *Scientific Reports*, *11*, 1–9.
- Madin, J. S., & Madin, E. M. P. (2015). The full extent of the global coral reef crisis. *Conservation Biology*, *29*, 1724–1726.
- Marquina, D., Andersson, A. F., & Ronquist, F. (2019). New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Molecular Ecology Resources*, *19*, 90–104.
- Martin, J. L., Santi, I., Pitta, P., John, U., & Gypens, N. (2022). Towards quantitative metabarcoding of eukaryotic plankton: An approach to improve 18S rRNA gene copy number bias. *Metabarcoding and Metagenomics*, *6*, e85794.
- Milivojević, T., Rahman, S. N., Raposo, D., Siccha, M., Kucera, M., & Morard, R. (2021). High variability in SSU rDNA gene copy number among planktonic foraminifera revealed by single-cell qPCR. *ISME Communications*, *1*, 1–8.
- Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*, 1422–1428.
- Miya, M. (2022). Environmental DNA Metabarcoding: A novel method for biodiversity monitoring of marine fish communities. *Annual Review of Marine Science*, *14*, 161–185.
- Moinard, S., Piau, D., Laporte, F., Rioux, D., Taberlet, P., Gonindard-Melodelima, C., & Coissac, E. (2023). Towards quantitative DNA metabarcoding: A method to overcome PCR amplification bias. <https://doi.org/10.1101/2023.10.03.560640>
- Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schaffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *15*, 1757–1764.
- Narayan, G. R., Reymond, C. E., Stuhr, M., Doo, S., Schmidt, C., Mann, T., & Westphal, H. (2022). Response of large benthic foraminifera to climate and local changes: Implications for future carbonate production. *Sedimentology*, *69*, 121–161.
- Nielsen, K. M., Johnsen, P. J., Bensasson, D., & Daffonchio, D. (2007). Release and persistence of extracellular DNA in the environment. *Environmental Biosafety Research*, *6*, 37–53.
- Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Cáceres, M., Durand, S., ... Weedon, J. (2022). *Vegan: Community Ecology Package*. R package version 2.6-4.
- Pawlowski, J., & Holzmann, M. (2014). A plea for DNA barcoding of foraminifera. *Journal of Foraminiferal Research*, *44*, 62–67.
- Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, *55*, 12–25.
- Pawluczyk, M., Weiss, J., Links, M. G., Egaña Aranguren, M., Wilkinson, M. D., & Egea-Cortines, M. (2015). Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, *407*, 1841–1848.
- Pillet, L., Fontaine, D., & Pawlowski, J. (2012). Intra-genomic ribosomal RNA polymorphism and morphological variation in *Elphidium macellum* suggests inter-specific hybridization in foraminifera. *PLoS One*, *7*, e32373.
- Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*, 407–419.
- Prazeres, M., Martínez-Colón, M., & Hallock, P. (2020). Foraminifera as bioindicators of water quality: The FoRAM index revisited. *Environmental Pollution*, *257*, 113612.
- Prazeres, M., & Renema, W. (2019). Evolutionary significance of the microbial assemblages of large benthic foraminifera. *Biological Reviews of the Cambridge Philosophical Society*, *94*, 828–848.
- Ratcliffe, F. C., Uren Webster, T. M., Rodríguez-Barreto, D., O'Rourke, R., Garcia de Leaniz, C., & Consuegra, S. (2021). Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples. *Ecological Applications*, *31*, e02284.
- Renema, W. (2010). Is increased calcarinid (foraminifera) abundance indicating a larger role for macro-algae in Indonesian Plio-Pleistocene coral reefs? *Coral Reefs*, *29*, 165–173.
- Renema, W. (2018). Terrestrial influence as a key driver of spatial variability in large benthic foraminiferal assemblage composition in the central Indo-Pacific. *Earth Science Reviews*, *177*, 514–544.
- Shelton, A. O., Gold, Z. J., Jensen, A. J., Agnese, D. E., Andruszkiewicz Allan, E., Van Cise, A., Gallego, R., Ramón-Laca, A., Garber-Yonts, M., Parsons, K., & Kelly, R. P. (2023). Toward quantitative metabarcoding. *Ecology*, *104*, e3906.
- Silverman, J. D., Bloom, R. J., Jiang, S., Durand, H. K., Dallow, E., Mukherjee, S., & David, L. A. (2021). Measuring and mitigating PCR bias in microbiota datasets. *PLoS Computational Biology*, *17*, e1009113.
- Skelton, J., Cauvin, A., & Hunter, M. E. (2022). Environmental DNA metabarcoding read numbers and their variability predict species abundance, but weakly in non-dominant species. *Environmental DNA*, *5*, 1092–1104. <https://doi.org/10.1002/edn3.355>
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050.
- Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*, 714–726.
- Torti, A., Lever, M. A., & Jørgensen, B. B. (2015). Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Marine Genomics*, *24*(Pt 3), 185–196.
- Uthicke, S., & Nobes, K. (2008). Benthic foraminifera as ecological indicators for water quality on the great barrier reef. *Estuarine, Coastal and Shelf Science*, *78*, 763–773.
- Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., Tapolczai, K., & Domaizon, I. (2018). Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*, *9*, 1060–1069.
- Weber, A. A.-T., & Pawlowski, J. (2013). Can abundance of protists be inferred from sequence data: A case study of foraminifera. *PLoS One*, *8*, e56739.
- Weber, A. A.-T., & Pawlowski, J. (2014). Wide occurrence of SSU rDNA intragenomic polymorphism in foraminifera and its implications for molecular species identification. *Protist*, *165*, 645–661.
- Wu, M., Kalyanasundaram, A., & Zhu, J. (2013). Structural and biomechanical basis of mitochondrial movement in eukaryotic cells. *International Journal of Nanomedicine*, *8*, 4033–4042.



Zhao, F., Filker, S., Xu, K., Li, J., Zhou, T., & Huang, P. (2019). Effects of intragenomic polymorphism in the SSU rRNA gene on estimating marine microeukaryotic diversity: A test for ciliates using single-cell high-throughput DNA sequencing. *Limnology and Oceanography: Methods*, *17*, 533–543.

#### 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:** Girard, E. B., Didaskalou, E. A., Pratama, A. M. A., Rattner, C., Morard, R., & Renema, W. (2024). Quantitative assessment of reef foraminifera community from metabarcoding data. *Molecular Ecology Resources*, *00*, e14000. <https://doi.org/10.1111/1755-0998.14000>