Received: 29 October 2020 | Revised: 30 March 2021 | Accepted: 13 April 2021

DOI: 10.1111/1755-0998.13405

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Proteomic fingerprinting facilitates biodiversity assessments in understudied ecosystems: A case study on integrated taxonomy of deep sea copepods

Jasmin Renz[1] | Elena L. Markhaseva[2] | Silke Laakmann[3,4] | Sven Rossel[5,6] | Pedro Martinez Arbizu[5,6] | Janna Peters[1]

[1]German Centre for Marine Biodiversity Research (DZMB), Senckenberg Research Institute, Hamburg, Germany

[2]Laboratory of Marine Research, Zoological Institute of the Russian Academy of Sciences, St. Petersburg, Russia

[3]Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg (HIFMB), Oldenburg, Germany

[4]Alfred Wegener Institute Helmholtz Center for Polar and Marine Research, Bremerhaven, Germany

[5]German Centre for Marine Biodiversity Research (DZMB), Senckenberg Research Institute, Wilhelmshaven, Germany

[6]Marine Biodiversity Research, Institute for Biology and Environmental Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Correspondence
Jasmin Renz, German Centre for Marine Biodiversity Research (DZMB), Senckenberg Research Institute, Hamburg, Germany.
Email: jasmin.renz@senckenberg.de

Funding information
BMBF; DFG, Grant/Award Number: RE 2808/3-1/2; ZIN theme, Grant/Award Number: AAAA, -, A and 19-119020690072-9; Ministry for Science and Culture of Lower Saxony; “Niedersächsisches Vorab”, Grant/Award Number: ZN3285; SENCKENBERG GESELLSCHAFT FUR NATURFORSCHUNG

Abstract

Accurate and reliable biodiversity estimates of marine zooplankton are a prerequisite to understand how changes in diversity can affect whole ecosystems. Species identification in the deep sea is significantly impeded by high numbers of new species and decreasing numbers of taxonomic experts, hampering any assessment of biodiversity. We used in parallel morphological, genetic, and proteomic characteristics of specimens of calanoid copepods from the abyssal South Atlantic to test if proteomic fingerprinting can accelerate estimating biodiversity. We cross-validated the respective molecular discrimination methods with morphological identifications to establish COI and proteomic reference libraries, as they are a pre-requisite to assign taxonomic information to the identified molecular species clusters. Due to the high number of new species only 37% of the individuals could be assigned to species or genus level morphologically. COI sequencing was successful for 70% of the specimens analysed, while proteomic fingerprinting was successful for all specimens examined. Predicted species richness based on morphological and molecular methods was 42 morphospecies, 56 molecular operational taxonomic units (MOTUs) and 79 proteomic operational taxonomic units (POTUs), respectively. Species diversity was predicted based on proteomic profiles using hierarchical cluster analysis followed by application of the variance ratio criterion for identification of species clusters. It was comparable to species diversity calculated based on COI sequence distances. Less than 7% of specimens were misidentified by proteomic profiles when compared with COI derived MOTUs, indicating that unsupervised machine learning using solely proteomic data could be used for quickly assessing species diversity.

KEYWORDS
calanoid copepods, COI sequencing, deep sea biodiversity, MALDI-TOF MS, proteomic fingerprinting

1 | INTRODUCTION

In light of the worldwide observed and predicted changes in biodiversity, there is an urgent need to measure spatial and temporal variation in biodiversity from local to global scales, to understand what impacts these changes might have on communities and ecosystems, and how biodiversity can be conserved (Costello, 2001; Ash et al., 2009). The deep sea is largely underinvestigated even though it constitutes the largest environment on earth, and only little information on the diversity of different metazoan groups (Ramirez-Llodra et al., 2011), which varies on local, regional and global scales, is available (Rex & Etter, 2010; Vinogradova, 1997). The increasing interest in exploring resources provided by the deep sea, such as polymetallic nodules, and the contemplated mining activities, will probably have an immense, yet unforeseeable impact on the inhabiting fauna (Cuyvers et al., 2018). Therefore, a fast and reliable assessment of species diversity is crucial to set baselines in biodiversity, understand the relationship of species to the surrounding environment, detect the influence of anthropogenic disturbances on species compositions and allow a sustainable use of deep-sea resources. Valid species identification is the first step to understand population structures, abundance, and diversity of the communities in the deep sea. While some taxa, such as polychaetes, are thought to be more widespread (Watling et al., 2013), others are found to have many endemic species, with many of them occurring only in singletons and most of them being new to science (e.g., Rex & Etter, 2010 and references therein). This hinders any assessment of regional diversity and biogeographic patterns.

Calanoid copepods often dominate benthopelagic communities of the deep sea (Wishner, 1980), showing high diversity (Bradford-Grieve, 2004) that in some areas is hypothesized to be comparable to pelagic waters (Renz & Markhaseva, 2015). Approximately 2,800 species of calanoid copepods are currently described (Park & Ferrari, 2009), although a large number of so far undetected species is expected to exist in underinvestigated ecosystems such as the deep sea. This is reflected in the fact that >70% of genera detected in recent years in the abyssal benthopelagic realm of the South Atlantic and Southern Ocean were new to science, with most of them being endemic to the benthopelagic zone and specialized to living within the vicinity of the seabed (Renz & Markhaseva, 2015). The identification of individual calanoids within the benthic boundary layer is significantly impeded by the morphological conservation of the group (Renz & Markhaseva, 2015) and, moreover, as generally applicable for many taxa, by increasing lacking taxonomic expertise. Furthermore, almost no identification literature is available for juvenile stages of calanoid copepods.

Molecular techniques have added a new dimension to the traditional phenotypic approach and allow researchers to overcome taxonomic difficulties in the identification of species with different life history stages. They can improve and contribute to estimating species diversity and were successfully applied to understand the diversity of calanoid copepods (e.g., Blanco-Bercial et al., 2014; Bucklin, Hopcroft, et al., 2010; Bucklin, Ortman, et al., 2010; Laakmann et al., 2013), resolve problematic taxa (e.g., Aarbakke et al., 2014; Hill et al., 2001) or reveal cryptic and new species (Caudill & Bucklin, 2004; Chen & Hare, 2011; Goetze, 2003). In many of these studies, species identification was enabled by analysing the mitochondrial cytochrome c oxidase subunit I gene (COI) as the metazoan species-specific DNA sequence (Hebert et al., 2003; Hebert et al., 2003). However, COI barcoding of individual species is time- and cost-intensive.

Proteomic fingerprinting, a tool commonly used for species and strain identification in microbiology (as reviewed by Croxatto et al., 2012) is a relatively new approach in metazoan studies. Proteomic spectra are determined using matrix-assisted laser/desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS). Pilot studies on proteomic fingerprinting of metazoan taxa indicate the possibilities to distinguish different species (e.g., Feltens et al., 2010; Kaufmann et al., 2012; Mazzeo et al., 2008; Volta et al., 2012), including closely-related freshwater copepod species (Riccardi et al., 2012), calanoid copepods from the North Sea (Laakmann et al., 2013) and tropical Atlantic (Bode et al., 2017; Kaiser et al., 2018) and benthic harpacticoid species (Rossel and Martínez Arbizu, 2018, 2019). Most of these studies either aimed to proof the general concept of proteomic fingerprints for a certain taxon group or to apply the method to the field using a reference library with pre-identified species. However, in the benthopelagic layers of the deep sea, we expect a high number of new, undescribed species, making the establishment of a such a reference library rather time- and cost-intensive. The use of unsupervised classification methods with proteomic fingerprints is a promising technique that may allow a prediction of species diversity without prior species identification (Rossel & Martínez Arbizu, 2020). It implies that specimens can be separated on species level using the dissimilarity of their MALDI spectra as the only source of information. The aim of our study was therefore to (i) find an adequate unsupervised technique to estimate species richness of a community from proteomic profiles only, (ii) evaluate resolution, accuracy and efficiency of species separation based on proteomic fingerprinting using cross-validation with morphology and COI sequencing, and (iii) provide for the first time data on diversity and composition of benthopelagic abyssal copepods using an integrated morphological, genetic and proteomic approach.

2 | MATERIALS AND METHODS

2.1 | Sampling

Calanoid copepods were collected in the benthopelagic boundary layer (BBL) during RV Meteor cruise M79-1 (Supporting Information 1) (Project DIVA 3) in the South Atlantic Ocean from station 580 (14°58.91’ S, 29°56.48’ W) at a depth of 5139 m using an epibenthic sledge (EBS; Brenke, 2005). The sledge consists of a closable 500 µm epi- and supranet, each with an opening of 1 m width and 0.35 m height. Both nets end up in a cod end with a mesh size of 300 µm. The net openings are positioned 0.2–0.6 m (epinet) and 0.77–1.12 m above the seabed (supranet). The EBS was hauled over the seabed at 1 knot for 10 min. Nets were opened before starting the trawl and closed before starting to haul the net from the seabed. On board, the samples were immediately fixed in 96% pure undenatured ethanol. Ethanol was exchanged within 24 h of sampling and samples were constantly cooled for molecular analyses.

2.2 | Sorting, identification, and specimen preparation for molecular analysis

Calanoid copepods were sorted in the laboratory and all individuals were classified into adult females, adult males and copepodites (altogether 358 individuals, Table 1). Adult stages were identified to genus or species level if possible, using a stereomicroscope and a microscope, and assigned to a morphotype. In some cases, it was necessary to dissect oral limbs and swimming legs to allow for genus identification. Individuals of all morphotypes were transferred to the collection of the DZMB to allow for later descriptions of new species, leaving 259 specimens for molecular analyses.

Identifications were based on the taxonomic literature on calanoids, for example, Damkaer (1975), Park (1978, 1980, 1982, 1983a,1983b), Brodsky et al., (1983), Schulz (1989, 1998), Ohtsuka et al., (1993), Ohtsuka et al., (1994), Bradford-Grieve (1994, 1999), Markhaseva (1996), Bradford-Grieve et al., (1999) and Andronov (2002).

All of the 259 individuals were cut in half for further molecular analysis to allow for concurrent measurements of molecular genetic analysis and proteomic fingerprinting from the same specimens. Molecular genetic analyses were conducted using the metasome and the urosome, while proteomic mass spectra were established using only the cephalosome of the individuals (except for individuals > 4 mm, where only the anterior part of the cephalosome was taken for analysis). Pre-tests with epipelagic copepods showed no significant differences in proteomic composition using the whole body or only parts of the cephalosome (personal observation).

2.3 | Molecular genetic analysis

Genomic DNA of a total of 259 specimens (41 adult females and adult males and 218 copepodites) was extracted using the QIAamp DNA mini kit (Qiagen) following the manufacturer's protocol with an overnight lysis (Table 1). PCR amplifications were accomplished by illustra PuRe-Taq Ready-To-Go PCR Beads (GE Healthcare) using 4 µl of DNA templates in 25 µl reaction volumes. COI amplification and sequencing were performed using the primer pair LCO1490 and HCO2198 (Folmer et al., 1994) with a thermoprofile of 95°C (5 min) and 38 cycles with 95°C (45 s), 45°C (50 s) and 72°C (1 min), and a final elongation at 72°C (5 min). In case of PCR failures four other primer pair combinations (dgLCO1490 and dgHCO2198 (Meyer, (2003)), LCO1490-JJ and HCO2198-JJ (Astrin and Stüben, 2008), LCO1490 and HCO709 (Blank et al., 2008) and LCO1490 together with Cop-COI-2189 (Bucklin, Ortman, et al., 2010) were applied (Table 2). PCR products were purified with ExoSap-IT (0.25 µl exo, 1 µl SAP). Both, PCR products and purified PCR products were checked on an agarose gel (1%) with GelRed (0.1%). Strands were sequenced using the Big Dye Terminator (Applied Biosystems Inc., ABI). The sequences were run on an ABI 3130xl DNA sequencer (Macrogen, Amsterdam).

Sequences were assembled, edited and checked for reading frames using the software GENEIOUS prime 2019 created by Biomatters (available from http://www. geneious.com/). The data sets were translated into amino acid alignments and checked for stop codons to avoid pseudogenes. Using BLAST (Altschul et al., 1990), sequences were compared with those available in GenBank. All new sequences were deposited in GenBank (Table 1). Multiple alignments of COI were performed in MEGA version 6.06 (Tamura et al., 2013) using default settings and the muscle algorithm (Edgar, 2004).

2.4 | Protein mass fingerprinting analysis (MALDI-TOF MS)

Proteomic profiles were determined for 259 specimens. The copepod tissue was quickly dried at room temperature. Depending on sample size 5–10 µl matrix solution (α-cyano-4-hydroxycinnamic acid as saturated solution in 50% acetonitrile, 47.5% LC-MS grade water, and 2.5% trifluoroacetic acid) was added. After at least 10 min extraction 1.2 µl of each sample was added onto the target plate. Protein mass spectra were measured from 2 to 20 kD using a linear-mode MALDI-TOF System (Microflex LT/SH, Bruker Daltonics). Peak intensities were analysed during random measurement in the range between 2 and 10 kDa using a centroid peak detection algorithm, a signal to noise threshold of 2 and a minimum intensity threshold of 400 with a peak resolution higher than 400 for mass spectra evaluation. Proteins/oligonucleotide method was employed for fuzzy control with a maximal resolution 10 times above the threshold. For each sample 240 satisfactory shots were summed up.

Spectra were analysed using the R-packages MALDIQUANT (Gibb & Strimmer, 2012), and MALDIQUANTFOREIGN (Gibb, 2013). Peaks were detected using a signal to noise ratio (SNR) of 7 after square-root transformation, savitzky golay smoothing, baseline removal (SNIP-algorithm) and normalization (TIC) of spectra. Peaks were repeatedly binned until the intensity matrix reached a stable peak number (tolerance 0.002, strict approach) and missing values were interpolated from the corresponding spectrum. All signals below a SNR <1.75 were assumed to be below detection limit and set to zero in the final peak matrix.

2.5 | Prediction of species number and diversity

2.5.1 | COI barcodes

A COI fragment of 658 basepair (bp) (minimum sequence length 334 bp) was analysed by neighbour-joining analysis based on uncorrected pairwise genetic distances using the software MEGA

TABLE 1 Calanoid copepod specimens from the South Atlantic, Project DIVA 3, station 580, for morphological identification and molecular analysis; for the analysis using matrix-assisted laser desorption/ionisation time-of-flight (Maldi-TOF) spectra, the anterior part (cephalosome) of the individuals was used; for the molecular genetic analysis of COI the posterior part (metasome + urosome) of the individuals was used

Morphospecies (SP) (Nr.)	Genus/species	Number of ind. and developmental stages for morphology	Stage for molecular analysis	Total length (mm)	MOTU (M) (Nr.)	COI (Nr.)	GenBank accession number
1	*Parkius* sp.[A]	2F, 1 M	F	1.58		KJ 717	
2	*Tharybis* sp.[A]	14F, 2 M, 1CV	2F	1.19, 1.15	25	KJ 718[†], KJ747[†]	MW807597 MW807599
3	*Omorius* sp.[A]	4F	F	2.50	47	KJ 719[†]	MW807580
4	*Byrathis* sp.[A]	1F	F	2.25	10	KJ 720[†]	MW807425
5	*Zenkevitchiella* sp.[B]	2F, 1 M	F	2.40	11	KJ 721[†]	MW807603
6	*Paracomantenna* sp.[C]	1F	F	7.80	46	KJ 722[†]	MW807427
7	*Xancithrix ohmani* [A]	1F	F	4.10	4	KJ 723[†]	MW807600
8	*Paraeuchaeta* sp.[D]	5F	F	7.35	5	KJ 724[†]	MW807583
9	*Xanthocalanus* sp.[A]	3F	2F	7.90, 7.90	42	KJ 725[†], KJ735	MW807584
10	*Xanthocalanus* sp.[A]	1F	F	7.20		KJ 726	
11	*Xanthocalanus* sp.[A]	1 M	M	4.30	18	KJ 727[†]	MW807602
12	*Prolutamator* sp.[C]	1F	F	3.55	2	KJ 728[†]	MW807591
13	*Aetideopsis* sp.[C]	1 M	M	2.55	26	KJ 729[†]	MW807422
14	*Chiridius* sp.[C]	11 M	3 M	2.10, 2.03, 1.90		KJ 730, KJ 743, KJ 749	
15	cf. *Bradyidius* sp.[C]	1F	F	2.65		KJ 731	
16	Scolecitrichidae[A]	5F, 3 M, 2 CV	2F, M	5.05 (F), 5.15 (F) 4.58 (M)	3	KJ 732[†], KJ733[†], KJ 750[†]	MW807593 MW807594 MW807595
17	*Paramisophria* sp.[E]	1F	F	7.90	59	KJ 734[†]	MW807589
18	*Xanthocalanus* sp.[A]	3F	2F	3.30, 3.60		KJ 736, KJ 746	
19	*Scolecitrichopsis* sp.[A]	13F, 4 M	F	1.75	1	KJ 737[†]	MW807597
20	*Scolecitrichopsis* sp.[A]	1 M	M	1.90	6	KJ 738[†]	MW807596
21	*Alrhabdus* sp.[E]	3F	F	2.50	28	KJ 739[†]	MW807423
22	*Prolutamator hadalis*[C]	6F	2F	2.70, 2.80	2	KJ 740[†], KJ 742[†]	MW807590
23	*Foxtonia* cf. *barbatula*[F]	1F	F	1.05	251	KJ 741[†]	MW807578
24	*Xanthocalanus* sp.[A]	1F	F	9.00	39	KJ 744	
25	*Xanthocalanus* sp.[A]	2F	F	6.60	9	KJ 745[†]	MW807601
26	*Omorius* sp.[A]	1F	F	2.13	8	KJ 748[†]	MW807579
27	Arietellidae[E]	1F, 1CV	F	3.80	54	KJ 751[†]	MW807424
28	Diaixidae[A]	1F	F	2.70		KJ 752	
29	*Ryocalanus* sp.[G]	1F, 1CII					
30	Aetideidae[C]	1F					
31	*Byrathis* sp.[A]	1F					
32	*Byrathis* sp.[A]	5F					

(Continues)

TABLE 1 (Continued)

Morphospecies (SP) (Nr.)	Genus/species	Number of ind. and developmental stages for morphology	Stage for molecular analysis	Total length (mm)	MOTU (M) (Nr.)	COI (Nr.)	GenBank accession number
33	Bradfordian incertae sedis [A]	1F					
34	*Vensiasa* sp.[A]	2F					
35	*Sensiava* sp.[A]	3F					
36	Bradfordian group	7 M					
37	*Caudacalanus* sp.[F]	2F					
38	Diaixidae[A]	1F					
39	Arietellidae[E]	1F					
40	*Frankferrarius admirabilis*[H]	1 M, 1CII					
41	*Scolecithrix danae*[A]	1F					
42	*Yrocalanus antarcticus*[G]	1F					
	Byrathis sp.[A]	2CV					
	Paraeuchaeta sp.[D]	8 juv.	juv.		5	KJ753[†]–KJ756[†]; KJ765[†]–766; KJ768; KJ771[†]	MW807585–MW807587; MW807558-MW807562
	Calanoida juvenile	215 juv.	Juv.		7; 12–21; 27; 29–38; 40–41; 43–45; 48–53; 55–58; 60	KJ757–KJ764; KJ767; KJ769–KJ770; KJ772–KJ831; KJ552–KJ570 KJ571–KJ694	MW807426; MW807428–MW807557; MW807559–MW807561; MW807563–MW807577; MW807581–MW807582; MW807586; MW807588; MW807592; MW807595; MW807598
		∑ 358					

Abbreviations: [†]Pre-idebtified morphospecies, for which COI sequencing successful; A, Bradfordian families Scolecitrichidae, Tharybidae, Phaennidae, Diaididae, Parkiidae; B, Bathypontiidae; C, Aetideidae; CII, CV, copepodite stage II and V; D, Euchaetidae; E, Arietellidae; F, Arctokonstantinidae; F, female; G, Ryocalanidae; H, Augaptilidae; juv, juvenile; M, male.

version 6.06 (Tamura et al., 2013). For comparative purposes, neighbour-joining analysis based on the K80 model (Kimura 2-parameter (K2P) was performed with the following settings: equal base frequencies, one transition and one transversion rate; Kimura, 1980) and 10,000 bootstrap replicates using the software MEGA version 6.06 (Tamura et al., 2013). Both models resulted in comparable tree topologies. We therefore chose the tree topology based on uncorrected p-distances to represent our data set and determine molecular operational taxonomic units (MOTUs) in our data set.

The online-tool CD-HIT- Suite (Ying et al., 2010) was used to identify MOTUs using the pairwise alignment with a predefined similarity threshold of 0.97 (Weizhong & Godzik, 2006; Ying et al., 2010) (97% similarity), which corresponds to the universal DNA-barcoding threshold proposed by Hebert, Cywinska, et al., (2003). In order to assign species identifications to the COI barcodes, we analysed our data set together with representatives of obligate benthopelagic-deep sea calanoid copepods from Laakmann et al., (2019) and Renz et al., (2019) as well as representatives of those species matching our sequences on a similarity level >97% as revealed by the BLAST analysis. The harpacticoid copepod *Euterpina acutifrons* (Dana, 1847) (KT209043.1) was chosen as outgroup taxon.

2.5.2 | Agglomerative hierarchical clustering of protein mass spectra

Due to the expected highly unbalanced data set, as commonly found in deep-sea communities, agglomerative hierarchical clustering (HC) was chosen to identify species clusters in proteomic profiles. Agglomerative HC was performed using Euclidean distances based on Hellinger-transformed peak intensities and single linkage.

TABLE 2 Primers used for PCR and sequencing

Primer label	Sequence (5'–3')	References
LCO1490	GGTCAACAAATCATAAAGATATTGG	Folmer et al. (1994)
HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. (1994)
LCO1490-JJ	CHACWAAYCATAAAGATATYGG	Astrin and Stüben (2008)
HCO2198-JJ	AWACTTCVGGRTGVCCAAARAATCA	Astrin and Stüben (2008)
dgLCO1490	GGTCAACAAATCATAAAGAYATYGG	Meyer, (2003)
dgHCO2198	TAAACTTCAGGGTGACCAAARAAYCA	Meyer, (2003)
HCO709	AATNAGAATNTANACTTCNGGGT	Blank et al. (2008)
Cop-COI-2189	GGGTGACCAAAAAATCARAA	Bucklin, Ortman, et al. (2010)

Four commonly used internal cluster criteria were tested for their ability to separate on species level using the R-package NBClust (Charrad et al., 2014, Table 3): the variance ratio criterion by Caliński and Harabasz (1974), the Dunn index (Dunn, 1974), the silhouette analysis (Rousseeuw, 1987) and the gap analysis (Tibshirani et al., 2001). Each of these iterative applied algorithms uses a specific measure to find significant clusters in the data set. The first two follow a quite similar approach in providing the number of clusters that are best separated and most compact. The Dunn index uses a ratio of separation (a minimum of pairwise distances between clusters) and compactness (as maximum of pair-wise distances within the cluster), while the variance criterion is based on the ratio of the between cluster sum of squares and the within cluster sum of squares. The latter also includes a penalty factor for the number of clusters tested. The often-applied silhouette analysis uses the difference between (normalized) separation and compactness instead of a ratio. For each data point a silhouette width is calculated, and the average of these widths then provides the validation criterium for the tested solution. Another approach is followed in the gap analysis, which compares the compactness of the clusters within the data set, with that of clusters of a random data set to validate whether the solution is significantly different from a random structure. In order to test, which cluster validation method is most suited to identify species-level structures in the data, we applied all criteria on the subset of samples which species-level identity was verified (labelled) by COI barcode (N = 182). We evaluated the consistency between MALDI data and MOTUs using four external cluster validation criteria provided by the R-package ClusterR: Rand, Hubert & Arabie adjusted Rand, Fowlkes & Mallows measures and Jaccard (Table 3, for all indices as well as a detailed discussion of the indices, see Jaccard, 1908; Rand, 1971; Fowlkes & Mallows, 1983; Hubert & Arabie, 1985; Milligan & Cooper, 1986; Wagner & Wagner, 2007). Due to the good clustering results of the criterion by Caliński and Harabasz (1974), we then used this criterion to estimate the total number of species based on all MALDI samples (N = 259).

2.5.3 | Consensus clustering of protein mass spectra

To determine the stability of species clusters a consensus clustering was performed using HC with single linkage using the R-package ConsensusClusterPlus (Wilkerson & Hayes, 2010). A consensus matrix was calculated based on 100 repetitions of HC using Euclidean distance of Hellinger transformed peak intensities based on subsets of features (f), here compounds and samples (s), respectively (i.e., f = 0.8/s = 0.8, f = 1/s = 0.8, f = 0.8/s = 1, f = 0.5/s = 1). Outer clustering of the consensus matrix was again done using HC with single linkage and cluster stability was inspected visually. The number of clusters was inferred from each consensus analysis using the proportion of ambiguous clustering (PAC) as internal validation measure (Șenbabaoğlu et al., 2014). PAC is defined as the fraction of sample pairs with consensus values in the interval above 0 (i.e., sample pairs that are never in the same cluster) and below 1 (i.e., sample pairs that are always in the same cluster). In a truly stable clustering, a consensus matrix contains only 0 and 1, and the PAC would have a score of 0. Here, we used 0.1 as lower and 0.9 as upper boundary. From this we inferred the optimal number of clusters by the lowest PAC. As for the agglomerative clustering we applied consensus clustering to the labelled subset of specimens (N = 182) to first validate the method and then on the whole data set (N = 259), using all samples and only 50% of features (s = 1, f = 0.5), to predict species numbers.

2.5.4 | Calculation of diversity

We calculated diversity in the calanoid community using the Shannon Index [$H' = -\sum pi\ln(pi)$, where pi is the proportion of individuals found in the species] as well as the Pielou's Evenness [$J = \frac{H'}{H\,max}$].

3 | RESULTS

3.1 | Discrimination of morphospecies based on morphological analysis

The 358 individuals sorted for morphological analysis consisted of 127 adult and 231 juvenile individuals (Table 1). 133 individuals, including all 127 adult individuals as well as six juvenile specimens were assigned to overall 42 different morphospecies based on differences in morphological characteristics and length. The majority of juvenile stages could not be assigned to any morphospecies due

TABLE 3 Results of agglomerative hierarchical clustering using the criterion by Caliński and Harabasz (1974), Dunn index (Dunn, 1974), gap analysis by Tibshirani et al., (2001) and the silhouette analysis by Rousseeuw (1987)

			Hierachical clustering					Consensus clustering				
	(*N* = 132)	(*N* = 182)	(*N* = 259)	Labelled subset only (*N* = 182)				(*N* = 259)	Labelled subset only (*N* = 182)			
Criterion	Morphology	p-distance COI	Caliński and Harabasz (1974)	Caliński and Harabasz (1974)	Dunn (1974)	Tibshirani et al., (2001)	Rousseeuw (1987)	F = 0.5, s = 1, PAC	F = 0.5, s = 1, PAC	F = 1, s = 0.8, PAC	F = 0.8, s = 0.8, PAC	F = 0.8, s = 1, PAC
Taxonomical units (“species”)*			79*					84*				
Taxonomical units (“species”)	42	60	(60)	58	53	3	181	(62)	59	60	53	52
Rand			0.99	0.99	0.99	0.07	0.96	0.99	0.99	0.99	0.99	0.99
Hubert & Arabie's adj. Rand			0.93	0.93	0.91	0.00	0.00	0.92	0.93	0.93	0.91	0.89
Fowlkes & Mallows's index			0.93	0.93	0.91	0.20	0.04	0.93	0.93	0.93	0.91	0.90
Jaccard index			0.87	0.87	0.84	0.04	0.00	0.86	0.87	0.87	0.84	0.81
Diversity	3.08	3.58	3.83*	3.53	3.45	0.09	5.20	3.86*	3.55	3.56	3.45	3.42
Evenness	0.82	0.87	0.88*	0.87	0.87	0.09	1.00	0.87*	0.87	0.87	0.87	0.86

The revealed taxonomical units (morphospecies, molecular operational taxonomic units (MOTUs) as well as cluster number based on proteomic profiles), external cluster evaluation based on intercomparison of MOTU and clusters, as well as Shannon diversity and Evenness are provided. *cluster number and diversity was calculated using all specimens (*N* = 259), external validation measures were calculated on the subset of labelled (by COI) specimens (*N* = 182).

to undeveloped identification characters and lacking identification keys for non-adult stages.

Of the 42 morphospecies, 24 morphospecies belonged to the "Bradfordian families" Scolecitrichidae, Tharybidae, Phaennidae, Diaixidae and Parkiidae (in total 88 individuals, i.e., 66.6%), seven to Aetideidae (22 ind., 16.6%), four to Arietellidae (7 ind., 5%), two to Arctokonstantinidae (3 ind, 2.3%), two to Ryocalanidae (3 ind., 2.3%), one to Euchaetidae (5 ind., 3.8%), one to Augaptilidae (2 ind., 1.5%) and one to Bathypontiidae (2 ind., 1.5%).

Nine specimens could only be identified to family or higher taxon level, as these individuals most probably belonged to new, undescribed genera. Only 12 specimens out of 358 (3.3%) could be assigned to six already described species (*Xancithrix ohmani*, *Scolecithrix danae*, *Prolutamator hadalis*, *Foxtonia cf. barbatula*, *Yrocalanus antarcticus* and *Frankferrarius admirabilis*) without further time-intensive dissection and detailed study of morphological characters that would have made individuals unsuitable for molecular analysis, because of the handling procedure involved.

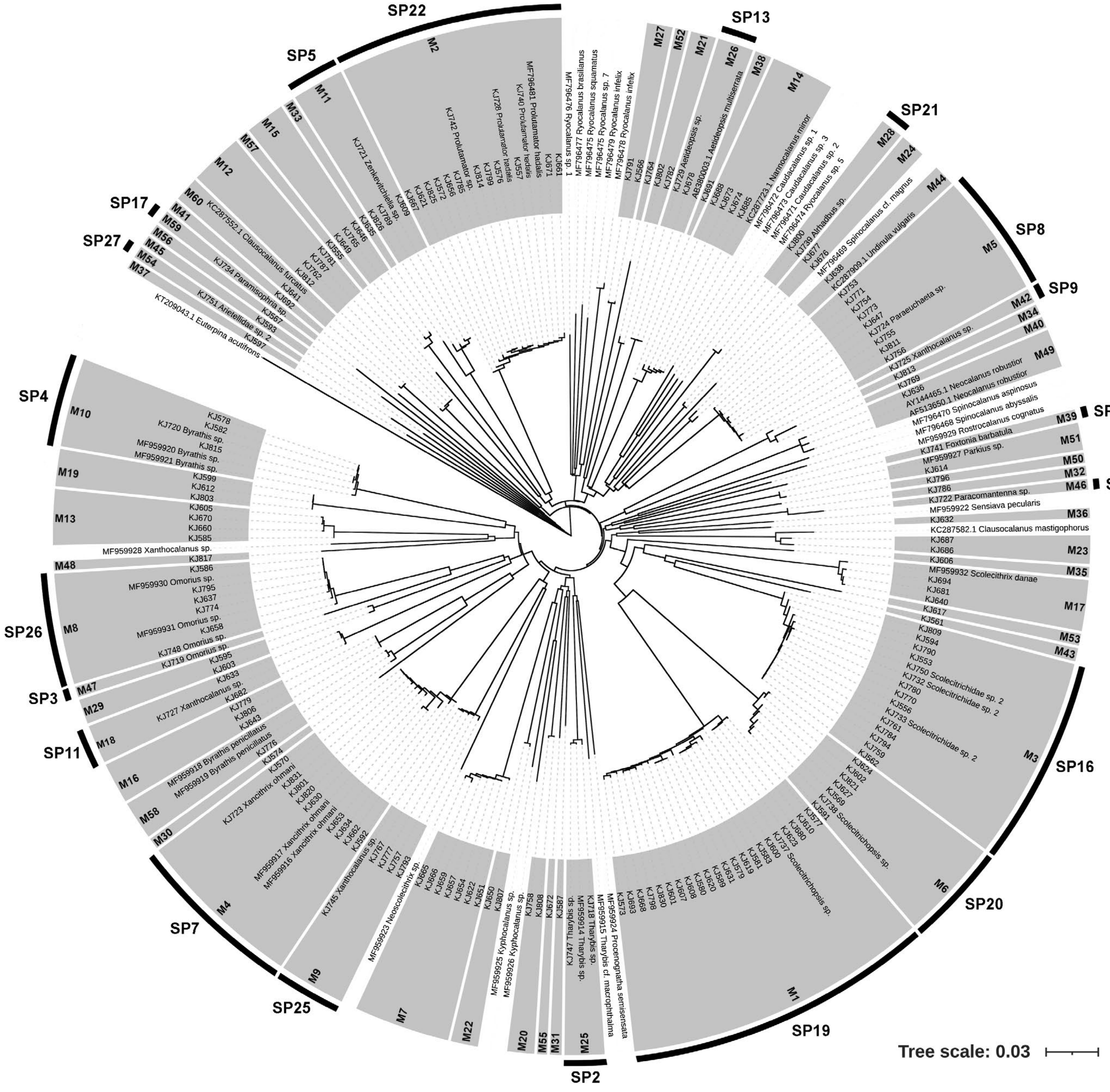

FIGURE 1 Neighbour-joining analysis of the 658 bp cytochrome c oxidase subunit I (COI) fragment based on uncorrected p-distances for benthopelagic calanoid copepods from the South Atlantic, including all sequences assigned to 60 identified MOTUs (M) by CD-hit based on sequence similarity of 97%; all sequences established during this publication are indicated by a KJ number. Sequences by Renz et al., (2019), Laakmann et al., (2019), and sequences with a similarity >97% as revealed by a BLAST search are indicated by GenBank Accession Numbers; SP, morphospecies-Nr., see also Table 1, grey areas: MOTUs (M) as revealed by the CD-hit analysis

A total of 16% (23 ind.) of those specimens assigned to a morphospecies were singletons, while all other morphospecies were represented by more than one specimen. Most specimens belonged to the morphospecies *Scolecitrichopsis* sp., followed by *Tharybis* sp. and *Paraeuchaeta* sp.

3.2 | Discrimination of molecular operational taxonomic units (MOTUs) based on mitochondrial COI

259 out of 358 specimens were available for molecular genetic analysis, while another 99 specimens were dissected in glycerin for morphological analysis and partly kept for later description of new species (Table 1). Morphotypes 28–41 could therefore not be included into molecular genetic analysis.

Amplification and sequencing were successful for 29 out of 44 pre-identified individuals (66%, including 24 adult specimens and five juvenile specimens), belonging to 19 morphospecies and 152 out of 215 unidentified juveniles (71%), leading to a COI sequence analysis for altogether 182 out of 259 individuals (70%) (Table 1).

Sequence length varied between 334 and 658 bp, with 77% of sequences longer than 650 bp. Identification of molecular operational taxonomic units (MOTUs) based on uncorrected p-distances at a similarity level >97% revealed 60 MOTUs (Figure 1), with 2 MOTUs based on sequences <500 bp. Of these 60 MOTUs, 29 (48%) were represented by singletons, 10 (17%) by two sequences and 21 (35%) by five or more sequences. Only two singleton MOTUs consisted of a sequence <500 bp, indicating only little influence of the inclusion of short sequence lengths on the estimation of diversity in this study. All morphospecies, for which sequencing was successful, were well reflected in the tree topology, forming different clades with sequence divergences >3%, except of one assigned morphospecies, as *Prolutamator* morphospecies 12 appeared in one clade with *Prolutamator hadalis* (morphospecies 22).

Seventeen MOTUs could be assigned to a priori assigned morphospecies that were identified to genus or species level, while two MOTUs could be assigned to a priori defined morphospecies that most likely represent undescribed genera within two different families of calanoids.

Searching the reference sequence of every MOTU against GenBank using the BLAST algorithm resulted in matching 7 MOTUs to already sequenced benthopelagic calanoid copepod species from the same area (*Prolutamator* sp., *Byrathis penicillatus* and *Byrathis* sp., *Omorius* sp., *Xancithrix ohmani*, *Tharybis* sp., *Parkius* sp.) (Laakmann et al., 2019; Renz et al., 2019), at a similarity level >97%. Seven MOTUs could be furthermore assigned to species known from the pelagial (*Neocalanus robustior*, *Undinula vulgaris*, *Clausocalanus furcatus* and *C. mastigophorus*, *Aetideopsis multiserrata*, *Foxtonia barbatula*, *Scolecithrix danae*), with three of them represented in the benthopelagic data set by singletons. Including morphological analysis as well as GenBank results using the Blast algorithm, altogether 26 out of 60 MOTUs could be assigned to a morphospecies, with some morphospecies most likely representing contamination from upper water layers during sampling, as for example *Clausocalanus* spp. and *Scolecithrix danae* can be considered to be typical pelagic inhabitants (see also Renz & Markhaseva, 2015).

3.3 | Discrimination of operational taxonomic units based on proteomic profiles (POTU)

Proteome profiles could be successfully measured for all of the 259 specimens that went into molecular analysis. In total 588 molecular compounds (with a signal to noise ratio >7 in at least one organism in the data set) were determined and used for all further analysis.

For 182 out of these 259 specimens COI barcodes were successfully determined and these specimens were used to validate the applied unsupervised approaches for species delimitation. Euclidean distances based on the proteome were generally lower within than between MOTUs (Figure 2) indicating that cluster analysis could reliably separate on the MOTU-level. However, the interpretation of cluster validation measures is not straight forward as different meaningful cluster structures and substructures may occur in a data set. Both, the variance ratio criterion (Caliński and Harabasz, 1974) as well as the Dunn index (Dunn, 1974) provided an optimal solution of 182 clusters, i.e., each specimen was forming an own cluster. Both methods also showed a distinct local maximum at a cluster number of 58 and 53, respectively (Figure 3a,b). The commonly used gap analysis (Tibshirani et al., 2001) as well as the silhouette analysis

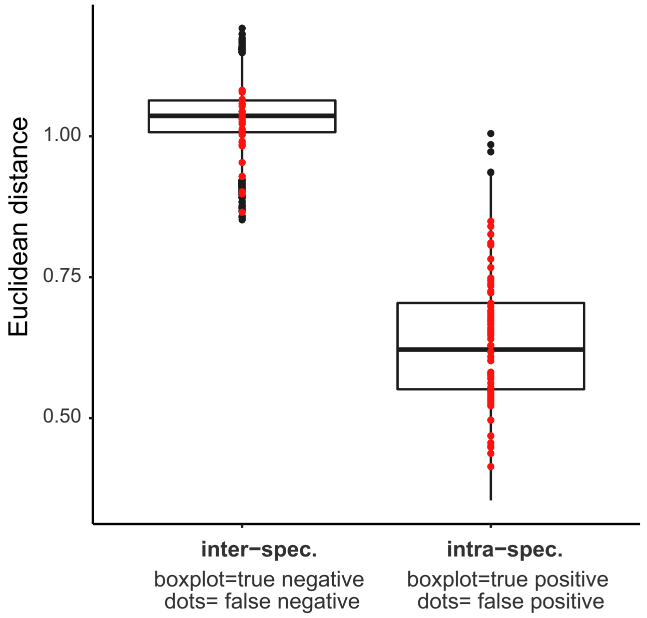

FIGURE 2 Euclidean distance based on Hellinger transformed peak intensities between (inter-spec.) and within (intra-spec.) POTUs: boxplots comprise true positives and true negatives, i.e., samples with consistent classification with MOTU assignment; red dots are Euclidean distances of false positive and negative samples, i.e., distances between the wrongly assigned samples with all other samples in the false cluster and the correct cluster, respectively

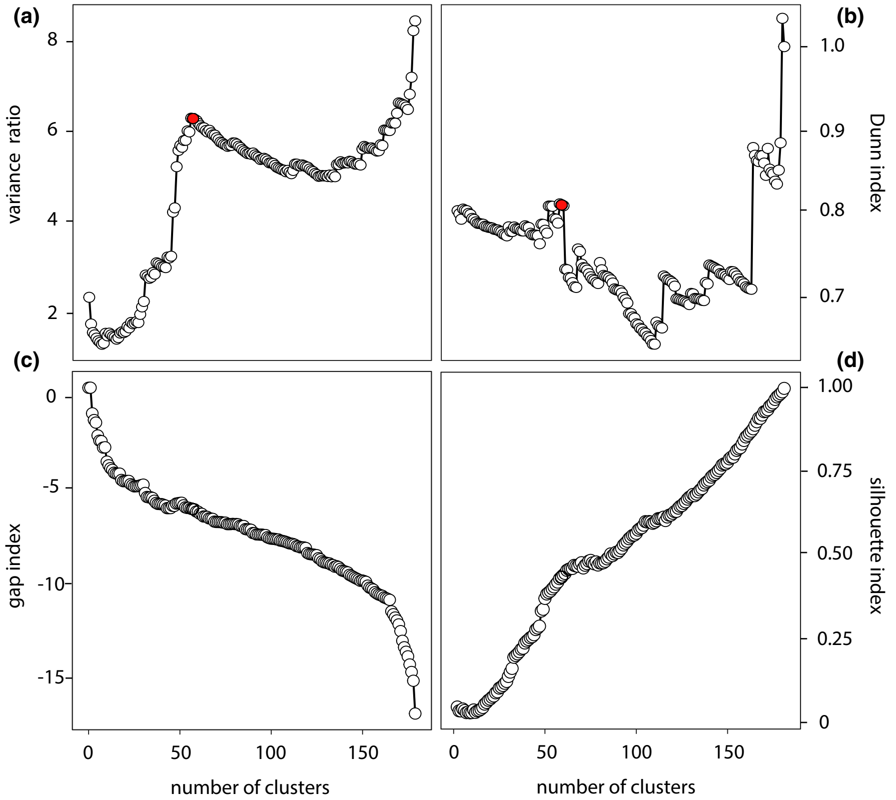

FIGURE 3 Identification of number of species clusters based on agglomerative hierarchical clustering using Euclidean distances derived from proteomic composition and internal cluster validation by the variance ratio (Caliński and Harabasz (1974); Figure 4a), the Dunn Index 1974 Dunn (1974); Figure 4b), the gap analysis (Tibshirani et al., 2001; Figure 4c) and the silhouette analysis (Rousseeuw, 1987; Figure 4d); red dots in 4a and 4b indicate the first maximum in the variance ratio and Dunn index, respectively

(Rousseeuw, 1987) did not reveal any of this substructure leading to an extreme under- and overestimation, respectively, when compared with MOTUs (Figure 3c,d). External cluster validation revealed highest success rates for delimiting clusters on MOTU-level (and thus probably species level) using the variance criterium (Caliński and Harabasz, 1974). We therefore defined these clusters as POTUs and their intercomparison with the identified MOTUs revealed that 13 out of 182 individuals (7%) were misidentified by proteomic-based clustering. However, MOTU identification in two of these clusters (M42 *Xanthocalanus* sp., M40 Indet., M34 Indet.) and (M46 Indet, M32 Indet) derived from short COI sequence lengths only, and MOTU delimitation within these clusters strongly dependeds on the model applied (personal observation). Thus, consistency between POTUs and species may be even higher. Euclidean distances of false-positive and false-negative assigned specimens were not distinctly different from the correct inter- and intraspecific distances, respectively (Figure 2).

Consensus clustering was used to estimate cluster stability. Between 52 and 60 clusters were inferred from consensus clustering followed by PAC depending on the percentage of samples and features (compounds) included (Table 3, Figure 4, Figure 5). The substructure of most clusters was low reflecting overall high stability of clustering results. However, some clusters with stronger substructure (e.g., M18 *Xanthocalanus* sp) and some linked cluster groups (e.g., M52 Indet., M15 Indet., M21 Indet) are more sensitive to incorrect delimitation. Clustering based on 100% of samples and 50% of compounds delimited 59 POTUs and misidentified 11 out of 182 specimens (6%) when compared to MOTUs. Six of the multi-MOTU clusters displayed no evident substructure.

3.4 | Biodiversity assessment

Morphology based diversity, calculated for the 132 individuals that were assigned to 42 morphotypes, resulted in a Shannon diversity of 3.1 and an evenness of 0.8. MOTU- and POTU-based diversity calculated for the 182 collectively determined samples was very similar with a Shannon diversity of 3.6 and 3.5, respectively and an evenness of 0.9. A biodiversity assessment based on all 259 specimens inferred the occurrence of 79 POTUs and a Shannon diversity of 3.8, while Evenness remained the same (0.9). Most POTUs occurred with only one (49% of all POTUs) or two (19% of all POTUs) individuals. Six POTUs occurred with higher specimen abundance: two POTUs of *Scolecitrichopsis* sp. (nine and 25 ind.), *Xancithrix ohmani* (10 ind.), *Paraeuchaeta* sp. (18 ind.), *Prolutamator* sp. (18 ind.) and Scolecitrichidae (19 ind.). Consistency for clustering was checked with the labelled subset of 182 specimens. High Rand indicators suggest that species composition was determined reasonably also with a larger data set using agglomerative hierarchical clustering.

4 | DISCUSSION

The Millennium Ecosystem Assessment provided strong evidence that the abundance of many species is declining, and that species distributions have been substantially altered due manifold anthropogenic activity (Millennium Ecosystem Assessment, 2005). A fast and reliable provision of comprehensive baselines for biodiversity is an urgent need for ecosystem management, yet still a strong challenge specifically in understudied marine ecosystems, such as the

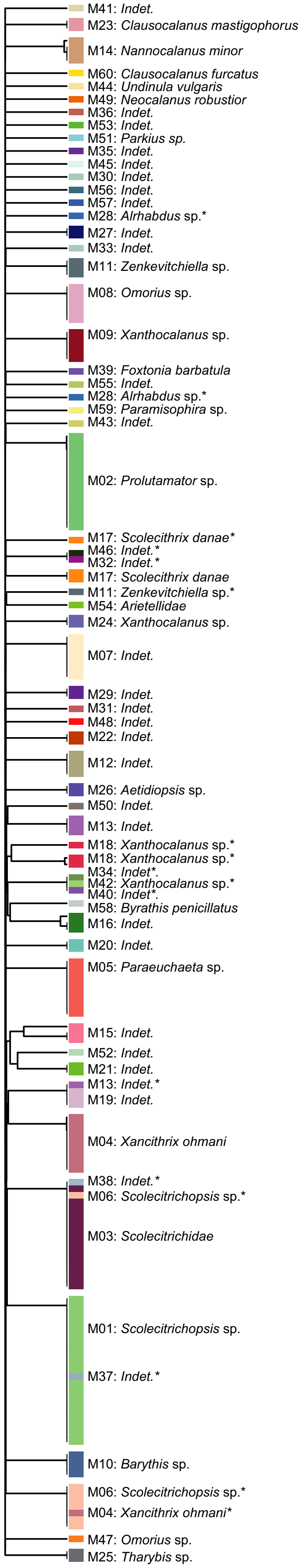

FIGURE 4 Hierarchical consensus clustering of proteomic profiles with 100% of samples and 50% of features (compounds) applying the proportion of ambiguous clustering as internal validation measure (Şenbabaoğlu et al., 2014) to identify cluster stability (here: 59 clusters). Colours of the clusters refer to the MOTUs as identified by COI; MOTU number and where possible the assigned name of the MOTU is provided; * indicates individuals where MOTU and POTU delimitation was not consistent

deep sea. Next to morphological identification, DNA based methods such as barcoding/metabarcoding, as well as the recently emerged rapid analyses using MALDI-TOF mass spectrometry to identify specimens using proteomic fingerprinting, were shown to accelerate the process of specimen identification in biodiversity assessments (Rossel et al., 2019). A crucial step in using these methods is to build reference libraries that connect morphological data to species-specific COI barcodes and proteome fingerprints. Here, for the first time, we report a study assessing species biodiversity of the highly specialized benthopelagic calanoid copepod community in the deep sea below 5,000 m in the South Atlantic, by a combined approach of morphological and molecular methods.

4.1 | Species identification by a combined morphological and molecular approach – a methodological evaluation

Four criteria of a method are substantial for biodiversity assessments: (i) the resolution, i.e., the taxonomic level of discrimination that can be reached, (ii) the accuracy, i.e., the percentage of correct classification, (iii) the net identification rate, i.e., the proportion of all available specimens of a sample that can be identified to species level, which is a result of resolution, accuracy and loss rate during the process and, finally (iv) the cost-benefit ratio, i.e., the effort in

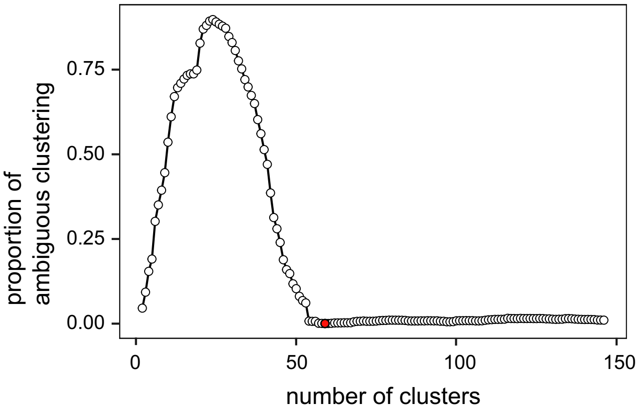

FIGURE 5 Identification of number of species clusters based on consensus clustering with 100% of samples and 50% of features (compounds) using Euclidean distances derived from proteomic composition and the proportion of ambiguous clustering (PAC) as internal validation measure (Şenbabaoğlu et al., 2014); the red dot indicates the first minimum of PAC

terms of resources in relation to the amount of information gained influencing thus the number of specimens or samples to infer biodiversity from.

4.1.1 | Taxonomic resolution

The barcode region of the mitochondrial COI gene is a well-established character for species-level identification of many marine metazoan taxa and has been shown to be successful in a vast range of studies on calanoid copepods (e.g., Blanco-Bercial et al., 2014; Bucklin, Hopcroft, et al., 2010; Bucklin, Ortman, et al., 2010; Laakmann et al., 2013; Machida et al., 2009). COI sequence analysis revealed 60 MOTUs when using the proposed sequence similarity of 97% (Hebert, Ratnasingham, et al., 2003). A less conservative estimate, allowing for a lower sequence similarity and a threshold of 95% delimiting MOTUs (see discussion in Blanco-Bercial et al., 2014), resulted in an estimate of 56 MOTUs including sequences >500 bp, and did thereby not significantly influence the estimate of diversity. Cryptic species or strong population structuring of widely distributed species may influence the threshold that has to be defined to estimate the number of MOTUs. However, little is known so far about the genetic diversity of benthopelagic calanoid copepod populations, their population dynamics and distribution in space and time. The number of MOTUs revealed by COI sequencing is therefore here considered to be a conservative first estimate for species diversity of calanoid copepods.

Inter- and intraspecific variation of proteomic fingerprints is far less understood. Several studies have proved that copepods show clear differences between species in proteomic mass spectra (Bode et al., 2017; Kaiser et al., 2018; Laakmann et al., 2013; Riccardi et al., 2012), and that these spectra can be used for species identification using supervised machine learning techniques with pre-established libraries (Rossel and Martínez Arbizu, 2018). Yet, no gold standard for unsupervised species delimitation has been developed. A previous study successfully applied partitioning around medoids (PAM) clustering in combination with the silhouette index to predict species number (Rossel and Martínez Arbizu, 2020). PAM and also k-means clustering were not applicable to our study due to the expected unbalanced data set with most probably many singletons and small sample size. Thus, we applied agglomerative hierarchical clustering (HC) in combination with different cluster validation methods. While the silhouette index seems to be very problematic for many singletons, the gap analysis only resolved larger structures in the data set. The two criteria based on the ratio of a separation- and compactness-measure of the clusters (Calinski-Harabasz and Dunn) were most accurate in delimiting POTUs on species-level. These strong differences emphasize that POTU identification is highly sensitive to the applied approach especially in “difficult” data sets with many singletons and/or unbalanced composition. However, regardless of the unsupervised algorithm applied,all approaches require a consistent, stable and taxon independent species gap in the similarity of mass spectra. At least for the species included in this study this prerequisite seems to be fulfilled, as classification of MOTUs and POTUs was generally consistent. More validation studies on species-level delimitation as well as the development of standard pipelines will be needed to establish proteomic fingerprinting as assessment tool for biodiversity in understudied species communities.

In conclusion, taxonomic resolution of all three methods was similar, with morphospecies, MOTUs and POTUs most likely representing distinct biological species.

4.1.2 | Accuracy

Another factor to be evaluated is the accuracy, i.e., percentage of specimens that can be correctly assigned to a species or genus name. This process is inevitably linked to morphological identification based on expert taxonomic knowledge, either during the study itself or by information coming from an integrative reference library. The genetic distance-based assignment of MOTUs using the basic local alignment search tool (BLAST) method (Altschul et al., 1990) provided the opportunity to assign MOTUs to already described and sequenced species, thereby adding to the species information obtained by the data set. This supports the importance of taxonomically comprehensive DNA barcode databases when morphological identification is not possible, for example, in juveniles. At present, proteomic profiles still require morphological or genetic intercalibration if more information than only diversity or species number is needed. Standard pipelines for supervised species identification in combination with central proteomic libraries still need to be established. The intercomparison of MOTUs and POTUs revealed that 7% of specimens were assigned differently. A slightly smaller error rate of 4% has been observed applying clustering on simulated data sets (Rossel and Martínez Arbizu, 2020). Misidentification of single specimens was neither detectable using a direct comparison of sample distances nor by consensus clustering, indicating that misidentification is probably caused by variance in the proteomic profiles. The stability of clusters was relatively high; however, some clusters seem to be more predestined for unstable delimitation than others. Overall, it is evident that the accuracy of proteomic fingerprinting for species discrimination of calanoid copepods is lower than of COI sequencing and also morphological identification.

4.1.3 | Net identification rate

The net identification rate was lowest for morphological identification, as it allowed morphospecies delimitation for only 37% of the specimens found, due to the tedious work and expert knowledge required for species identification, as well as the almost exclusive limitation to adult individuals. Molecular methods contributed significantly to the estimation of diversity by providing the possibility of including juvenile stages into the analysis. COI sequencing was successful in 70% of the specimens that were included in the molecular analyses. The loss rate of 30% most likely originated from the need for taxa-specific optimization; i.e., the sometimes low affinity of the universal COI barcoding primers by Folmer et al., (1994) could, in our case, not always be countered by group-specific primers for calanoid copepods. Proteomic mass spectra were successfully determined for all specimens included, and cross validation with identified MOTUs revealed an identification rate of 93% by proteomic based clustering. Proteomic fingerprinting therefore proved to be the most robust and comprehensive approach of all three methods for biodiversity assessments. It should be kept in mind that the net identification rate is not a stable number, but increases with the progress of the respective method, for example, the number of species descriptions, development of identification keys, increasing number of barcodes in GenBank and growing reference data bases, taxonomic knowledge and time spent for identification.

4.1.4 | Cost-benefit ratio

Required resources in terms of time- and cost-effort per sample as well as handling expertise are, next to identification rate, resolution and accuracy, additional criteria which need to be considered when assessing the value of a method for routine assessments of biodiversity. While optical identification of morphospecies is cheap in terms of consumables, this method requires, in absence of keys or species descriptions, the most personal costs, as the process is long lasting and can only be performed by highly experienced taxonomists. Barcoding also requires experienced staff (at least well-trained technician level) and the cost of consumables are generally high, more than 5 Euro per specimen (Rossel et al., 2019). MALDI-TOF, on the other hand, is a fast and low-cost (0.1 Euro consumables) method (Rossel et al., 2019), which can be accomplished easily after very short training, resulting in relatively low personal costs.

4.2 | Species richness and diversity

The deep sea is by far less explored than coastal areas, although it occupies 60% of the planet (Costello et al., 2010), and its stability and large area may accommodate a high species richness (Grassle, 1989). Calanoid copepods form the most numerous taxon in pelagic waters, often making up 80% of the zooplankton biomass in the water column (Mauchline, 1998). However, knowledge about their diversity in the deep is scarce, due to the lack of taxonomic knowledge and the occurrence of many species with strikingly similar morphology. By this, they provide an ideal model taxon for a biodiversity study in the deep sea. The combined morphological and molecular approach applied in this study revealed species diversity of calanoids in the abyssal region of the South Atlantic to be as high as in many pelagic systems. The biodiversity assessment based on proteomic fingerprints, validated with morphological identification and COI sequence analysis, estimated a taxonomic richness of 79 species and, as a result of low within-species abundance, a diversity of 3.5. This diversity is comparable to the highest diversity estimates of calanoid copepods in pelagic studies (e.g., Fernández de Puelles et al., 2019; Hwang et al., 2009). Information on copepod diversity and species richness in the pelagic habitat of the tropical South Atlantic is sparse and almost lacking for benthopelagic layers. For the latter, Renz and Markhaseva (2015) identified a minimum number of approximately 25 genera at stations below 5,000 m. Most were endemic to the abyssal habitat close to the seabed. A similar number of genera was shown for pelagic calanoids from the water column by Woodd-Walker et al. (2002) in the same region. For the whole South Atlantic, 97 pelagic genera were reported by Boltovskoy (1999), including near-shore genera. Of these 97 genera only six were reported for abyssal depths. The most detailed information is given in the database by Razouls et al. (2005-2020, accessed 061219), reporting approximately 300 pelagic calanoid copepod species for the Central South Atlantic. This database includes 15 obligate benthopelagic species described in recent years from abyssal depth. In a study off the Brazilian Coast, Dias et al., (2018) detected 111 species of calanoids from 44 genera down to 1,200 m depth. The present study revealed one of the highest species numbers detected so far for the benthopelagic layers of the abyssal deep sea in the South Atlantic.

In conclusion, cross-validation of proteomic fingerprinting with morphology and COI sequencing proved a generally consistent species discrimination of calanoid copepods for all three methods. Based on this, proteomic fingerprinting added significantly to the biodiversity assessment, as it was the only method allowing for a successful analysis of all individuals examined. With this method, an extremely high species diversity of calanoids, as well as a high degree of singletons, could be detected. Morphological information revealed that most of these species are new to science. Therefore, we consider proteomic fingerprinting to be an accurate, fast, inexpensive, and therefore highly promising assessment tool, which can provide comprehensive baselines of species diversity, not only in epipelagic monitoring studies, but in deep-sea studies with high number of unknown species as well. The method is still in its infancy in marine science. Reference libraries allowing taxonomic information to be assigned to species-specific proteomic features need to be established and filled before the method can be applied as a "stand-alone" tool. Also, we will have to enhance our understanding on the uncertainties and pitfalls of the method. However, although taxonomic expertise remains the keystone for any biodiversity assessment, and COI barcodes provide reliable information for species discrimination and assignment, an integration of proteomic fingerprinting will clearly enhance and accelerate the identification processes in biodiversity studies.

ACKNOWLEDGEMENTS

Sampling was done during RV Meteor cruise M79-1 financed by DFG and BMBF. The authors thank the Working Group on Morphological and Molecular Taxonomy of the International Council for the Exploration of the Sea (ICES) for facilitating this research and are grateful to Karen Jeskulke, who assisted during laboratory work for genetic analysis. This work was supported by the DFG initiative 1991 "Taxon-omics" (grant number RE 2808/3-1/2). Elena Markhaseva acknowledges support by ZIN theme AAAA-A19-119020690072-9. HIFMB is a collaboration between the Alfred-Wegener-Institute, Helmholtz-Center for Polar and Marine Research, and the Carl-von-Ossietzky University Oldenburg, initially funded by the Ministry for Science and Culture of Lower Saxony and the Volkswagen Foundation through the "Niedersächsisches Vorab" grant program (grant no. ZN3285). This is publication number 12 of Senckenberg am Meer Proteome Laboratory and number 80 of Senckenberg am Meer Metabarcoding and Molecular Laboratory. Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

Jasmin Renz, Janna Peters, Silke Laakmann, Pedro Martinez Arbizu and Sven Rossel designed the general approach of this study. Jasmin Renz and Pedro Martinez Arbizu participated in sample collection. The analyses were performed by Jasmin Renz, Janna Peters and Elena L. Markhaseva. Jasmin Renz and Janna Peters analysed the data sets and wrote the manuscript. All authors contributed to writing the manuscript.

DATA AVAILABILITY STATEMENT

DNA sequences (including morphological identification and sampling location) have been submitted to GenBank and accession numbers (BankIt 2443312) (MW807422 - MW807603) are provided in Table 1. Proteomic profiles have been submitted to Dryad, https://doi.org/10.5061/dryad.8cz8w9gpx.

ORCID

Jasmin Renz https://orcid.org/0000-0002-2658-445X
Sven Rossel https://orcid.org/0000-0002-1187-346X

REFERENCES

Aarbakke, O. N. S., Bucklin, A., Halsband, C., & Norrbin, F. (2014). Comparative phylogeography and demographic history of five sibling species of *Pseudocalanus* (Copepoda: Calanoida) in the North Atlantic Ocean. *Journal of Experimental Marine Biology and Ecology*, *461*, 479–488.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.

Andronov, V. N. (2002). The calanoid copepods (Crustacea) of the genera *Diaixis* Sars, 1902, *Parundinella* Fleminger, 1957, *Undinella* Sars 1900 and *Tharybis* Sars, (1902). *Arthropoda Selecta*, *11*(1), 1–80.

Ash, N., Jürgens, N., Leadley, P., Alkemade, R., Araújo, M. B., & Asner, G. P. (2009). bioDISCOVERY: Assessing, monitoring and predicting biodiversity. DIVERSITAS Report No. 7.

Astrin, J. J., & Stüben, P. E. (2008). Phylogeny in cryptic weevils: molecules, morphology and new genera of western Palaearctic Cryptorhynchinae (Coleoptera: Curculionidae). *Invertebrate Systematics*, *22*, 503–522.

Blanco-Bercial, L., Cornils, A., Copley, N., & Bucklin, A. (2014). DNA barcoding of marine copepods: assessment of analytical approaches to species identification. *PLoS Currents*, *6*, 62.

Blank, M., Laine, A. O., Jürss, K., & Bastrop, R. (2008). Molecular identification key based on PCR/RFLP for three polychaete sibling species of the genus *Marenzelleria*, and the species' current distribution in the Baltic Sea. *Helgoland Marine Research*, *62*(2), 129.

Bode, M., Laakmann, S., Kaiser, P., Hagen, W., Auel, H., & Cornils, A. (2017). Unravelling diversity of deep-sea copepods using integrated morphological and molecular techniques. *Journal of Plankton Research*, *39*, 600–617. https://doi.org/10.1093/plankt/fbx031.

Boltovskoy, D. (1999). *South Atlantic Zooplankton*. Backhuys Publishers.

Bradford-Grieve, J. M. (1994). The marine fauna of New Zealand: pelagic calanoid Copepoda: Megacalanidae, Calanidae, Paracalanidae, Mecynoceridae, Eucalanidae, Spinocalanidae, Clausocalanidae. *New Zealand Oceanographic Institute Memoir*, *102*, 5–156.

Bradford-Grieve, J. M. (1999). The marine fauna of New Zealand: pelagic calanoid copepoda: Bathypontiidae, Arietellidae, Augaptilidae, Heterorhabdidae, Lucicutiidae, Metridinidae, Phyllopodidae, Centropagidae, Pseudodiaptomidae, Temoridae, Candaciidae, Pontellidae, Sulcanidae, Acartiidae, Tortanidae. *NIWA Biodiversity Memoir*, *111*, 1–268.

Bradford-Grieve, J. M. (2004). Deep-sea benthopelagic calanoid copepods and their colonization of the near-bottom environment. *Zoological Studies*, *43*(2), 276–291.

Bradford-Grieve, J. M., Markhaseva, E. L., Rocha, C. E. F., & Abiahy, B. (1999). Copepoda. In B. Boltovskoy (Ed.), *South Atlantic Zooplankton* (pp. 869–1098). Backhuys publication.

Brenke, N. (2005). An epibenthic sledge for operations on marine soft bottom and bedrock. *Marine Technology Society Journal*, *39*, 10–19.

Brodsky, K. A., Vyshkvartzeva, N. V., Kos, M. S., & Markhaseva, E. L. (1983). Oar-footed crustaceans (Copepoda: Calanoida) of the seas of USSR and adjacent waters. *Opredeliteli Po Faune SSSR*, *Izdavaemye Zoologicheskim Institutom Akademii Nauk SSSR*, *Leningrdad*, *Nauka*, *1*, 1–356.[In Russian].

Bucklin, A., Hopcroft, R. R., Kosobokova, K. N., Nigro, L. M., Ortman, B. D., Jennings, R. M., & Sweetman, C. J. (2010). DNA barcoding of Arctic Ocean holozooplankton for species identification and recognition. *Deep Sea Research Part II: Topical Studies in Oceanography*, *57*(1–2), 40–48.

Bucklin, A., Ortman, B. D., Jennings, R. M., Nigro, L. M., Sweetman, C. J., Copley, N. J., & Wiebe, P. H. (2010). A "Rosetta Stone" for metazoan zooplankton: DNA barcode analysis of species diversity of the Sargasso Sea (Northwest Atlantic Ocean). *Deep Sea Research Part II: Topical Studies in Oceanography*, *57*(24–26), 2234–2247.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Caudill, C. C., & Bucklin, A. (2004). Molecular phylogeography and evolutionary history of the estuarine copepod, *Acartia tonsa*, on the Northwest Atlantic coast. *Hydrobiologia*, *511*(1–3), 91–102.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36. http://www.jstatsoft.org/v61/i06/

Chen, G., & Hare, M. P. (2011). Cryptic diversity and comparative phylogeography of the estuarine copepod *Acartia tonsa* on the US Atlantic coast. *Molecular Ecology*, *20*(11), 2425–2441.

Costello, M. J. (2001). To know, research, manage, and conserve marine biodiversity. *Océanis*, *24*(4), 25–49.

Costello, M. J., Coll, M., Danovaro, R., Charrad, M., Halpin, P., Ojaveer, H., & Miloslavich, P. (2010). A census of marine biodiversity knowledge, resources and future challenges. *PLoS ONE*, e12110.

Croxatto, A., Prod'hom, G., & Greub, G. (2012). Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews*, *36*(2), 380–407.

Cuyvers, L., Berry, W., Gjerde, K., Thiele, T., & Wilhem, C. (2018). *Deep seabed mining: A rising environmental challenge* (p. 74). IUCN and Gallifrey Foundation.

Damkaer, D. M. (1975). *Calanoid copepods of the genera* Spinocalanus *and* Mimocalanus *from the central Arctic Ocean, with a review of the Spinocalanidae*, Vol. *55* (p. 55). US Department of Commerce.

Dias, C. D. O., Araujo, A. V. D., & Bonecker, S. L. C. (2018). Vertical distribution and structure of copepod (Arthropoda: Copepoda) assemblages in two different seasons down to 1,200 m in the tropical Southwestern Atlantic. *Zoologia (Curitiba)*, *35*, 1–11. https://doi.org/10.3897/zoologia.35.e13886.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797.

Feltens, R., Görner, R., Kalkhof, S., Gröger-Arndt, H., & von Bergen, M. (2010). Discrimination of different species from the genus Drosophila by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry. *BMC Evolutionary Biology*, *10*(1), 95.

Fernández de Puelles, M., Gazá, M., Cabanellas-Reboredo, M., Santandreu, M., Irigoien, X., González-Gordillo, J. I., & Hernández-León, S. (2019). Zooplankton abundance and diversity in the tropical and subtropical ocean. *Diversity*, *11*(11), 203.

Folmer, O. M., Black, W., Hoen, R., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*, 294–299.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

Gibb, S. (2013). MALDIquantForeign: Import/Export routines for MALDIquant. R Pack. Ver., 9. http://CRAN.R-project.org/package=MALDIquantForeign.

Gibb, S., & Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, *28*(17), 2270–2271.

Goetze, E. (2003). Cryptic speciation on the high seas; global phylogenetics of the copepod family Eucalanidae. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1531), 2321–2331.

Grassle, J. F. (1989). Species diversity in deep-sea communities. *Trends in Ecology & Evolution*, *4*(1), 12–15.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 313–321.

Hebert, P. D. N., Ratnasingham, S., & DeWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 96–99.

Hill, R., Allen, L., & Bucklin, A. (2001). Multiplexed species-specific PCR protocol to discriminate four N. Atlantic *Calanus* species, with an mtCOI gene tree for ten *Calanus* species. *Marine Biology*, *139*(2), 279–287.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Hwang, J. S., Souissi, S., Dahms, H. U., Tseng, L. C., Schmitt, F. G., & Chen, Q. C. (2009). Rank-abundance allocations as a tool to analyze planktonic copepod assemblages off the Danshuei river estuary (Northern Taiwan). *Zoological Studies*, *48*(1), 49–62.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin De La Societe Vaudoise Des Sciences Naturelles*, *44*, 223–270.

Kaiser, P., Bode, M., Cornils, A., Hagen, W., Martínez Arbizu, P., Auel, H., & Laakmann, S. (2018). High-resolution community analysis of deep-sea copepods using MALDI-TOF protein fingerprinting. *Deep-Sea Research Part I*, *138*, 122–130.

Kaufmann, C., Schaffner, F., Ziegler, D., Pflueger, V., & Mathis, A. (2012). Identification of field-caught *Culicoides* biting midges using matrix-assisted laser desorption/ionization time of flight mass spectrometry. *Parasitology*, *139*(2), 248–258.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120.

Laakmann, S., Gerdts, G., Erler, R., Knebelsberger, T., Martínez Arbizu, P., & Raupach, M. (2013). Comparison of molecular species identification for North Sea calanoid copepods (Crustacea) using proteome fingerprints and DNA sequences. *Molecular Ecology Resources*, *13*, 862–876.

Laakmann, S., Markhaseva, E. L., & Renz, J. (2019). Do molecular phylogenies unravel the relationships among the evolutionary young "Bradfordian" families (Copepoda; Calanoida). *Molecular Phylogenetics and Evolution*, *130*, 330–345.

Machida, R., Hashiguchi, Y., Nishida, M., & Nishida, S. (2009). Zooplankton diversity analysis through single-gene sequencing of a community sample. *BMC Genomics*, *10*, 438.

Markhaseva, E. L. (1996). Calanoid copepods of the family Aetideidae of the World ocean. *Trudy Zoologicheskogo Instituta RAN, St. Petersburg*, *268*, 331p.

Mauchline, J. (1998). The Biology of Calanoid Copepods. *Advances in Marine Biology*, *33*, 710.

Mazzeo, M. F., Giulio, B. D., Guerriero, G., Ciarcia, G., Malorni, A., Russo, G. L., & Siciliano, R. A. (2008). Fish authentication by MALDI-TOF mass spectrometry. *Journal of Agricultural and Food Chemistry*, *56*(23), 11071–11076.

Meyer, C. P. (2003). Molecular systematics of cowries (Gastropoda: Cypraeidae) and diversification patterns in the tropics. *Biological Journal of the Linnean Society*, *79*, 401–459. https://doi.org/10.1046/j.1095-8312.2003.00197.x.

Millennium Ecosystem Assessment (2005). *Ecosystems and Human Well-being: Biodiversity Synthesis*. World Resources Institute.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, *21*(4), 441–458.

Ohtsuka, S., Boxshall, G. A., & Roe, H. S. J. (1994). Phylogenetic relationships between arietellid genera (Copepoda: Calanoida), with the establishment of three new genera. *Bulletin of the Natural History Museum London (Zool.)*, *60*(2), 105–172.

Ohtsuka, S., Roe, H. S. J., & Boxshall, G. A. (1993). A new family of calanoid copepods, the Hyperbionycidae, collected from the deep-sea hyperbenthic community in the northeastern Atlantic. *Sarsia*, *78*(1), 69–82.

Park, E. T., & Ferrari, F. D. (2009). Species diversity and distributions of Pelagic Copepods from the Southern Ocean. In I. Krupnik, M. A. Lang, & S. E. Miller (Eds.), *A Selection from Smithsonian at the Poles Contributions to International Polar year* (pp. 143–180). Smithsonian Institution Scholarly Press.

Park, T. S. (1978). Calanoid copepods (Aetideidae and Euchaetidae) from antarctic and subantarctic waters. *Biology of the Antarctic Seas VII*, *27*, 91–290.

Park, T. S. (1980). Calanoid copepods of the genus Scolecithricella from Antarctic and Subantarctic Seas. In: Biology of the Antarctic seas. IX. Antarctic Research Series Washington, *31*(2): 25–79.

Park, T. S. (1982). Calanoid copepods of the genus Scaphocalanus from Antarctic and Subantarctic waters. In: Biology of the Antarctic Seas, XI. Antarctic Research Series Washington, *34*(2): 75–127.

Park, T. S. (1983a). Calanoid copepods of some scolecithricid genera from antarctic and subantarctic waters. *Biology of the Antarctic Seas XIII*, *38*, 165–213.

Park, T. S. (1983b). Calanoid copepods of the family Phaennidae from Antarctic and Subantarctic waters. *Biology of the Antarctic Seas. XIV. Antarctic Research Series*, *39*(5), 317-368317-368.

Ramirez-Llodra, E., Tyler, P. A., Baker, M. C., Bergstad, O. A., Clark, M. R., Escobar, E., Levin, L. A., Menot, L., Rowden, A. A., Smith, C. R., & Van Dover, C. L. (2011). Man and the last great wilderness: human impact on the deep sea. *PLoS One*, *6*(8), e22588.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Razouls, C., de Bovée, F., Kouwenberg, J., & Desreumaux, N. (2005–2020). *Diversity and Geographic Distribution of Marine Planktonic*

Copepods. Sorbonne University. Available at http://copepodes.obs-banyuls.fr/en.

Renz, J., & Markhaseva, E. L. (2015). First insights into genus level diversity and biogeography of deep sea benthopelagic calanoid copepods in the South Atlantic and Southern Ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, *105*, 96–110.

Renz, J., Markhaseva, E. L., & Laakmann, S. (2019). The phylogeny of Ryocalanoidea (Copepoda, Calanoida) based on morphological and a multi-gene analysis with a description of new ryocalanoidean species. *Zoological Journal of the Linnean Society*, *185*(4), 925–957. https://doi.org/10.1093/zoolinnean/zly069.

Rex, M. A., & Etter, R. J. (2010). *Deep-sea Biodiversity: Pattern and Scale*. Harvard University Press.

Riccardi, N., Lucini, L., Benagli, C., Welker, M., Wicht, B., & Tonolla, M. (2012). Potential of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) for the identification of freshwater zooplankton: a pilot study with three *Eudiaptomus* (Copepoda: Diaptomidae) species. *Journal of Plankton Research*, *34*(6), 484–492.

Rossel, S., Khodami, S., & Martinez Arbizu, P. (2019). Comparison of rapid biodiversity assessment of meiobenthos using MALDI-TOF MS and Metabarcoding. *Frontiers in Marine Science*, *6*, 659.

Rossel, S., & Martínez Arbizu, P. (2018). Automatic specimen identification of Harpacticoids (Crustacea: Copepoda) using Random Forest and MALDI-TOF mass spectra, including a post hoc test for false positive discovery. *Methods in Ecology and Evolution*, *9*, 1421–1434.

Rossel, S., & Martínez Arbizu, P. (2019). Revealing higher than expected diversity of Harpacticoida (Crustacea: Copepoda) in the North Sea using MALDI-TOF MS and molecular barcoding. *Scientific Reports*, *9*(1), 9182.

Rossel, S., & Martínez Arbizu, P. (2020). Unsupervised biodiversity estimation using proteomic fingerprints from MALDI-TOF MS data. *Limnology and Oceanography Methods*, *18*(5), 183–195.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Schulz, K. (1989). Notes on rare spinocalanid copepods from the eastern North Atlantic, with descriptions of new species of the genera Spinocalanus and Teneriforma (Copepoda: Calanoida). *Mitteilungen Aus Dem Hamburgischen Zoologischen Museum Und Institut*, *86*, 185–208.

Schulz, K. (1998). A new species of *Xantharus* Andronov, 1981 (Copepoda: Calanoida) from the mesopelagic zone of the Antarctic Ocean. *Helgoländer Meeresuntersuchungen*, *52*(1), 41.

Şenbabaoğlu, Y., Michailidis, G., & Li, J. Z. (2014). Critical limitations of consensus clustering in class discovery. *Scientific Reports*, *4*(1), 1–13.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Vinogradova, N. G. (1997). Zoogeography of the abyssal and hadal zones. *Advances in Marine Biology*, Vol. *32* (pp. 325–387). Academic Press.

Volta, P., Riccardi, N., Lauceri, R., & Tonolla, M. (2012). Discrimination of freshwater fish species by matrix-assisted laser Desorption/Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS): a pilot study. *Journal of Limnology*, *71*(1), e17.

Wagner, S., & Wagner, D. (2007). *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik.

Watling, L., Guinotte, J., Clark, M. R., & Smith, C. R. (2013). A proposed biogeography of the deep ocean floor. *Progress in Oceanography*, *111*, 91–112.

Weizhong, L., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*, 1658–1659.

Wilkerson, D. M., & Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence *assessments and item tracking*. *Bioinformatics*, *26*(12), 1572–1573.

Wishner, K. (1980). Aspects of the community ecology of deep-sea benthopelagic plankton, with special attention to gymnopleid copepods. *Marine Biology*, *60*, 179–187.

Woodd-Walker, R. S., Ward, P., & Clarke, A. (2002). Large-scale patterns in diversity and community structure of surface water copepods from the Atlantic Ocean. *Marine Ecology Progress Series*, *236*, 189–203.

Ying, H., Beifang, N., Ying, G., Limin, F., & Weizhong, L. (2010). CD-HIT: Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, *26*, 680–682. https://doi.org/10.1093/bioinformatics/btq003.

SUPPORTING INFORMATION

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

How to cite this article: Renz J, Markhaseva EL, Laakmann S, Rossel S, Martinez Arbizu P, Peters J. Proteomic fingerprinting facilitates biodiversity assessments in understudied ecosystems: A case study on integrated taxonomy of deep sea copepods. *Mol Ecol Resour*. 2021;21:1936–1951. https://doi.org/10.1111/1755-0998.13405