Received: 4 October 2022 | Revised: 13 February 2023 | Accepted: 20 February 2023

DOI: 10.1111/1755-0998.13779

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Perspectives of species identification by MALDI-TOF MS in monitoring—Stability of proteomic fingerprints in marine epipelagic copepods

Janna Peters[1] | Silke Laakmann[2,3,4] | Sven Rossel[4] | Pedro Martínez Arbizu[4] | Jasmin Renz[1]

[1]German Center for Marine Biodiversity Research (DZMB), Senckenberg am Meer, Hamburg, Germany

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

[3]Alfred Wegener Institute, Helmholtz-Centre for Polar and Marine Research (AWI), Bremerhaven, Germany

[4]German Center for Marine Biodiversity Research (DZMB), Senckenberg am Meer, Wilhelmshaven, Germany

Correspondence
Janna Peters, German Center for Marine Biodiversity Research (DZMB), Senckenberg am Meer, Hamburg, Germany.
Email: janna.peters@senckenberg.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: RE2808/3-1/2; Volkswagen Foundation, Grant/Award Number: ZN3285

Handling Editor: Michael M. Hansen

Abstract

We analysed the robustness of species identification based on proteomic composition to data processing and intraspecific variability, specificity and sensitivity of species-markers as well as discriminatory power of proteomic fingerprinting and its sensitivity to phylogenetic distance. Our analysis is based on MALDI-TOF MS (matrix-assisted laser desorption ionization time of flight mass spectrometry) data from 32 marine copepod species coming from 13 regions (North and Central Atlantic and adjacent seas). A random forest (RF) model correctly classified all specimens to the species level with only small sensitivity to data processing, demonstrating the strong robustness of the method. Compounds with high specificity showed low sensitivity, that is identification was based on complex pattern-differences rather than on presence of single markers. Proteomic distance was not consistently related to phylogenetic distance. A species-gap in proteome composition appeared at 0.7 Euclidean distance when using only specimens from the same sample. When other regions or seasons were included, intraspecific variability increased, resulting in overlaps of intra and inter-specific distance. Highest intraspecific distances (>0.7) were observed between specimens from brackish and marine habitats (i.e., salinity probably affects proteomic patterns). When testing library sensitivity of the RF model to regionality, strong misidentification was only detected between two congener pairs. Still, the choice of reference library may have an impact on identification of closely related species and should be tested before routine application. We envisage high relevance of this time- and cost-efficient method for future zooplankton monitoring as it provides not only in-depth taxonomic resolution for counted specimens but also add-on information, such as on developmental stage or environmental conditions.

KEYWORDS
MALDI-TOF MS, marine zooplankton, marker sensitivity, marker specificity, monitoring, proteomic fingerprinting

1 | INTRODUCTION

Marine zooplankton species are useful indicators of environmental variation and climate change as they rapidly respond to changes in biological and physical conditions. Awareness of the importance of time-series-based zooplankton monitoring is increasingly growing. Time- and cost-efficient species identification methods are urgently needed not only for time-series but also for many fields of marine science, such as assessment of community turnover or biodiversity in the context of ecosystem-based management. Manual counts of these small organisms require well-trained personnel, as the taxonomic resolution is often very limited due to high morphological similarity or absence of diagnostic features in young developmental stages. Although taxonomic expertise remains a keystone for community monitoring, integration with molecular approaches can enhance and accelerate identification processes. However, DNA barcoding of single organisms is not yet suitable for routine species identification in time-series as it requires numerous steps in the working procedure, which come with high costs. Genetic multispecies approaches such as organismal metabarcoding of bulk samples are increasingly finding their way into zooplankton monitoring (Bucklin et al., 2016, 2019, 2021) as they combine comprehensive information on species occurrences with good methodological efficiency (Laakmann et al., 2020). In addition, approaches focusing on environmental DNA metabarcoding are increasingly being applied (Djurhuus et al., 2020, Suter et al., 2021). Although these multispecies approaches provide valuable extensive species information, they remain semiquantitative so far.

Proteomic fingerprinting as a fast, efficient and low-cost method for species identification (Renz et al., 2021; Rossel et al., 2019) has considerable potential to evolve to a valuable add-on to the current classical and molecular toolbox in zooplankton identification. In short, sample tissue is extracted in a matrix solution, which is then applied onto a target plate. The extracted compounds, mainly consisting of small cytosolic proteins and peptides (Ryzhov & Fenselau, 2001), are measured by matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) producing species-specific mass spectra, allowing the differentiation and, in combination with a reference database, the identification of specimens. Within the last 10 years several proof-of-concept studies on metazoans have revealed its general applicability for invertebrates (Murugaiyan & Roesler, 2017), specifically for insects (Campbell, 2005; Chavy et al., 2019; El Hamzaoui et al., 2018; Hasnaoui et al., 2022; Lawrence et al., 2019) and arachnids (Diarra et al., 2017; Gittens et al., 2020; Karger et al., 2019; Ngoy et al., 2021), which are relevant as vectors and pests. Far fewer studies have addressed taxa less relevant to human health. In the aquatic realm, research is mainly focused on groups which hold pivotal positions in marine and limnic food webs, namely copepods (Bode et al., 2017; Kaiser et al., 2018; Laakmann et al., 2013; Renz et al., 2021; Riccardi et al., 2012; Rossel & Martínez Arbizu, 2019; Yeom et al., 2021), cladocerans (Hynek et al., 2018) and fish (Maasz et al., 2017; Rossel et al., 2021; Volta et al., 2012).

To advance the method on its way to becoming a standard tool for zooplankton identification, a thorough evaluation of the sensitivity and specificity of proteomic fingerprinting and the intraspecific variance of fingerprints is essential. For bacteria it has been shown that culture conditions may influence peak numbers, spectrum quality and identification success based on MALDI-TOF MS (Balážová et al., 2014; Goldstein et al., 2013). Knowledge on spectra variations and factors influencing them is limited in metazoans; for example, spectra of ticks varied with season and habitat (Karger et al., 2019), and population-specific patterns were identified for bed bugs (Benkacimi et al., 2020) and mosquitoes (Müller et al., 2013). However, the underlying causes of differences between regions, either of genetic or environmental origin, remain unclear. To the best of our knowledge, there is as yet no information on seasonal and regional variability of proteomic patterns in marine invertebrates, their resilience against physiological or environmental impacts or the stability of markers between genetically more distant populations. As a first approach to answering these questions, we have analysed mass spectra of abundant epipelagic copepods from various zooplankton monitoring stations in the Atlantic and adjacent seas. In doing so, we have covered a wide range of environments from the Arctic to temperate zones, brackish and euryhaline waters, as well as neritic and oceanic regimes and as many seasons as possible.

The aim of our study was, based on data from various marine copepod populations and species, (i) to validate the general robustness of the species differentiation and identification approach to data processing and data variance, (ii) to determine the specificity and sensitivity of single proteomic markers for the species, (iii) to estimate the discriminatory power of proteomic fingerprinting and its sensitivity to phylogenetic distance, (iv) to estimate inter- and intraspecific variability of spectra searching for stable species gaps and the impact of variation on identification success, and finally (v) to present perspectives of proteomic fingerprinting for marine zooplankton studies.

2 | MATERIALS AND METHODS

2.1 | Sample collection

Specimens were derived from ethanol samples (>96%), which were collected during diverse monitoring programmes or field campaigns and from a copepod culture (Figure 1; Table S1). Age and storing conditions varied between samples; samples were stored at 4°C from 2018 onwards. Adult female copepods were identified morphologically to the species level and stored in ethanol until further processing at 4°C. In total 752 specimens from 32 species and 13 different regions were used for proteomic fingerprinting analyses. Selected samples from the North Sea have been part of the pilot study on using MALDI-TOF MS for species discrimination of North Sea copepods (Laakmann et al., 2013).

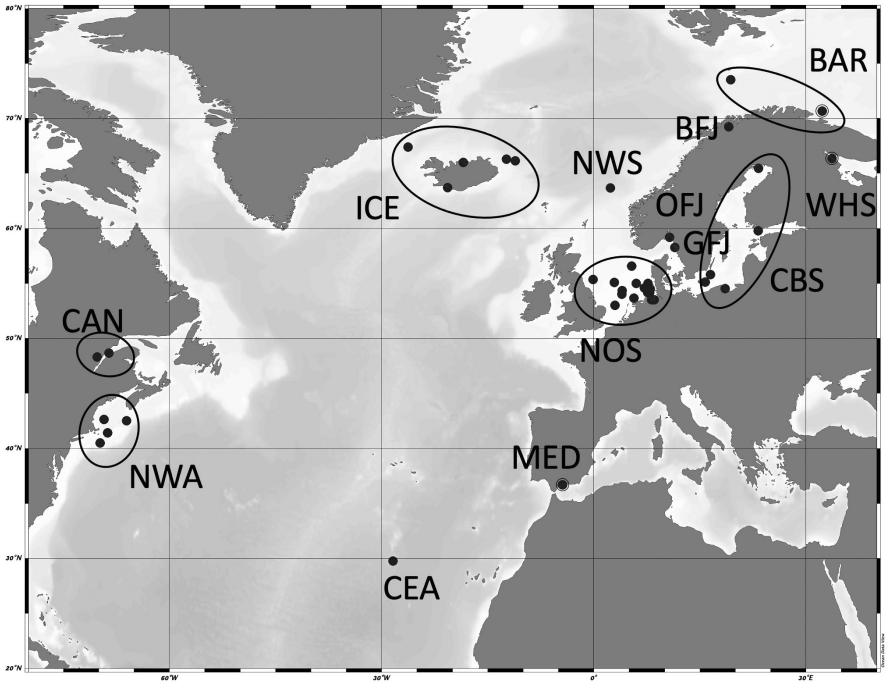

FIGURE 1 Overview of included regions. BAR, Barents Sea; BFJ, Balsfjord; CAN, Canada; CBS, Central Baltic Sea; CEA, Central-East Atlantic; GFJ, Gullmarsfjord; ICE, Icelandic waters; MED, Mediterranean; NOS, North Sea; NWA, North-West Atlantic; NWS, Norwegian Sea; OFJ, Oslofjord; WHS, White Sea (also see Table S1).

2.2 | Proteomic measurement

Proteomic profiles were determined for all 752 specimens. For small copepods (<2 mm) the whole specimen was used, while for larger copepods (such as *Calanus* spp., *Metridia* sp., *Paraeuchaeta* spp.) a piece of the cephalosome was briefly dried at room temperature and kept in an Eppendorf tube. Depending on sample size, 5–10 μL of 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 (Williams et al., 2003). After at least 10 min of extraction, 1.2 μL of each sample was added onto the target plate, with two or three replicates. Protein mass spectra were measured from 2 to 20 kDa 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 20 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 >400 for mass spectra evaluation. A protein/oligonucleotide method was used for fuzzy control with a maximal resolution 10 times above the threshold. For each sample 240 satisfactory shots were summed.

2.3 | Data processing

Spectra in the range 2–20 kDa were processed with MALDIQUANT (Gibb & Strimmer, 2012) and MALDIQUANTFOREIGN (Gibb, 2017) using square root transformation, Savitzky–Golay smoothing with a half window size of 10, baseline removal by the statistics-sensitive nonlinear iterative peak-clipping algorithm (SNIP, Ryan et al., 1988) and normalization setting the total ion current to one. Normalized spectra of technical replicates were averaged. Optimal peak detection parameters were derived by varying the signal to noise ratio (SNR) thresholds for peak identification and the half window size (HWS) of peak picking, both in the range 3–15 with species classification success of the random forest (RF) model (for method see below) as the target variable. We also determined the proportion of closely spaced peaks after binning (i.e., mass difference between peaks <6 Da), that have a risk of being misassigned to a specific *m/z* value. The highest classification success was reached with an SNR of 4 and an HWS of 3, but the proportion of closely spaced peaks was still high and strongly decreased at SNR and HWS values of 8. These peak detection parameters (SNR = 8, HWS = 8) were applied to the final data set. Picked peaks were repeatedly binned to compensate for small variation in the *m/z* values between measurements until the intensity matrix reached a stable peak number (with a tolerance of 2000 ppm, strict approach). All signals below the SNR were set to zero in the final peak matrix. For all further analysis peak intensities were Hellinger-transformed (Legendre & Gallagher, 2001) using the R package VEGAN (Oksanen et al., 2019) as this proved to be beneficial for proteomic data (Rossel & Martínez Arbizu, 2018a).

2.4 | Species classification

Species classification was performed by an RF model using the R package RANDOMFOREST (Liaw & Wiener, 2002) using 2000 trees and the square root of peak number as randomly sampled variables at each split. To avoid overrepresentation of the most abundant species, the number of subsampled specimens per species in each of the 2000 decision trees was limited to the abundance of the least abundant species ($N = 4$), respectively. Under-sampling the majority class can reduce the risk of overfitting, but it also comes with potential loss of important information. Species with fewer than four specimens were excluded as we observed a strong increase in the out of bag error with

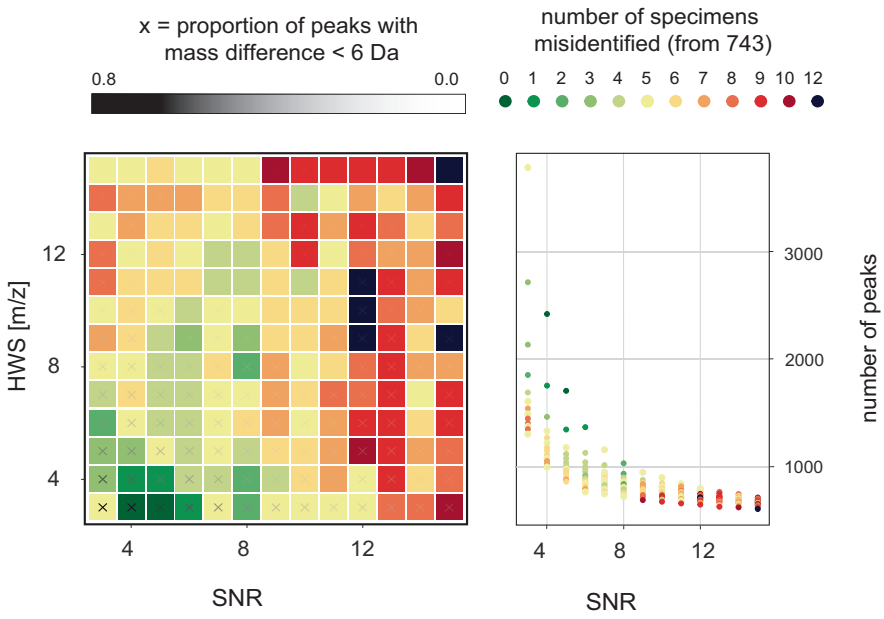

FIGURE 2 Left panel: impact of peak detection parameters (SNR = signal to noise ratio threshold for peak picking, influencing resolution on the intensity axis and HWS [*m/z*] = half window size of peak picking algorithm, influencing resolution on the *m/z* axis) on species classification success of the random forest model shown as colour code from 0 to 12 misidentified specimens out of 743; the proportion of peaks with a mass difference <6 Da are given as crosses. Right panel: impact of SNR on number of peaks; colour code gives number of misidentified specimens.

smaller sample size (unpublished data). For all other species the classification RF model was applied to all specimens from these 27 species: *Acartia* (*Acanthacartia*) *bifilosa* (*N* = 5), *A.* (*Acanthacartia*) *tonsa* (*N* = *34*), *A.* (*Acartiura*) *clausi* (*N* = 48), *A.* (*Acartiura*) *longiremis* (*N* = 81), *A.* (*Acartia*) *danae* (*N* = 18), *A.* (*Acartia*) *negligens* (*N* = 6), *Anomalocera patersonii* (*N* = 9), *Calanus finmarchicus* (*N* = 77), *Ca. helgolandicus* (*N* = 29), *Ca. glacialis* (*N* = 10), *Ca. hyperboreus* (*N* = 29), *Centropages bradyi* (*N* = 4), *Ce. typicus* (*N* = 47), *Ce. hamatus* (*N* = 53), *Ce. chierchiae* (*N* = 6), *Ditrichocorycaeus anglicus* (*N* = 10), *Eurytemora affinis affinis* (*N* = 12), *Limnocalanus macrurus macrurus* (*N* = 12), *Metridia longa* (*N* = 13), *M. lucens* (*N* = 24), *Microcalanus* sp. (*N* = 12), *Nannocalanus minor* (*N* = 24), *Paraeuchaeta norvegica* (*N* = 10), *Pseudocalanus elongatus* (*N* = 16), *P. moultoni* (*N* = 15), *Temora longicornis* (*N* = 91) and *T. sytlifera* (*N* = 16). Note that these are the taxon names accepted as valid by the World Register of Marine Species, but for simplification we will use genus and species names only from here on.

2.5 | Species-specific markers

The most important peaks for species identification including all peaks with a maximum of class-specific mean decrease in accuracy of more than 0.015 (leading to 180 important peaks) were derived from the final species RF model. Species-mean Euclidean distances were calculated using the R package VEGAN (Oksanen et al., 2019) based on individual distances using intensities of the whole peak spectrum and used for hierarchical clustering with ward.D2 linkage.

2.6 | Phylogenetic patterns

All genera represented by at least two congener species in the data set (i.e., *Acartia*, *Calanus*, *Centropages*, *Metridia*, *Pseudocalanus* and *Temora*) were included in a principal coordinates analysis (PCoA) based on Euclidean distances derived from Hellinger-transformed peak intensities. The analysis was performed using the R package APE (Paradis & Schliep, 2019).

2.7 | Stability of proteomic signals and choice of reference library

The stability of proteomic signals between samples and regions was analysed comparing the intraspecific variances of Euclidean distances between all specimens within a sample, within a region from different samples and seasons as well as between all specimens coming from different regions. In a next step we tested how reliable species identification can be done using inter-regional libraries: species RF models excluding all specimens from a certain region were used to assign these specimens to an RF model species class, respectively. A post hoc test for false positive discovery of proteomic profile-based RF models (Rossel & Martínez Arbizu, 2018a, https://github.com/pmartinezarbizu/RFtools) was applied, using a significance of <.01 to reject a classification as potentially false positive. Heatmaps presenting the distance between single specimens from different regions were created for congener pairs *A. clausi* (three regions) and *A. longiremis* (six regions), *Ce. typicus* (two regions) and *Ce. hamatus* (five regions), *T. stylifera* (one region) and *T. longicornis* (seven regions) as well as *Ca. hyperboreus* (three regions) and *Ca. finmarchicus* (six regions).

3 | RESULTS

3.1 | Species classification and species-specific markers

The main two parameters of resolution for peak detection, namely HWS on the *m/z* axis and SNR on the intensity axis, influenced the success of species classification. The out-of-bag error rate was below 0.8% in the range of SNR 4–8 and HWS 4–8 (Figure 2). Peak number

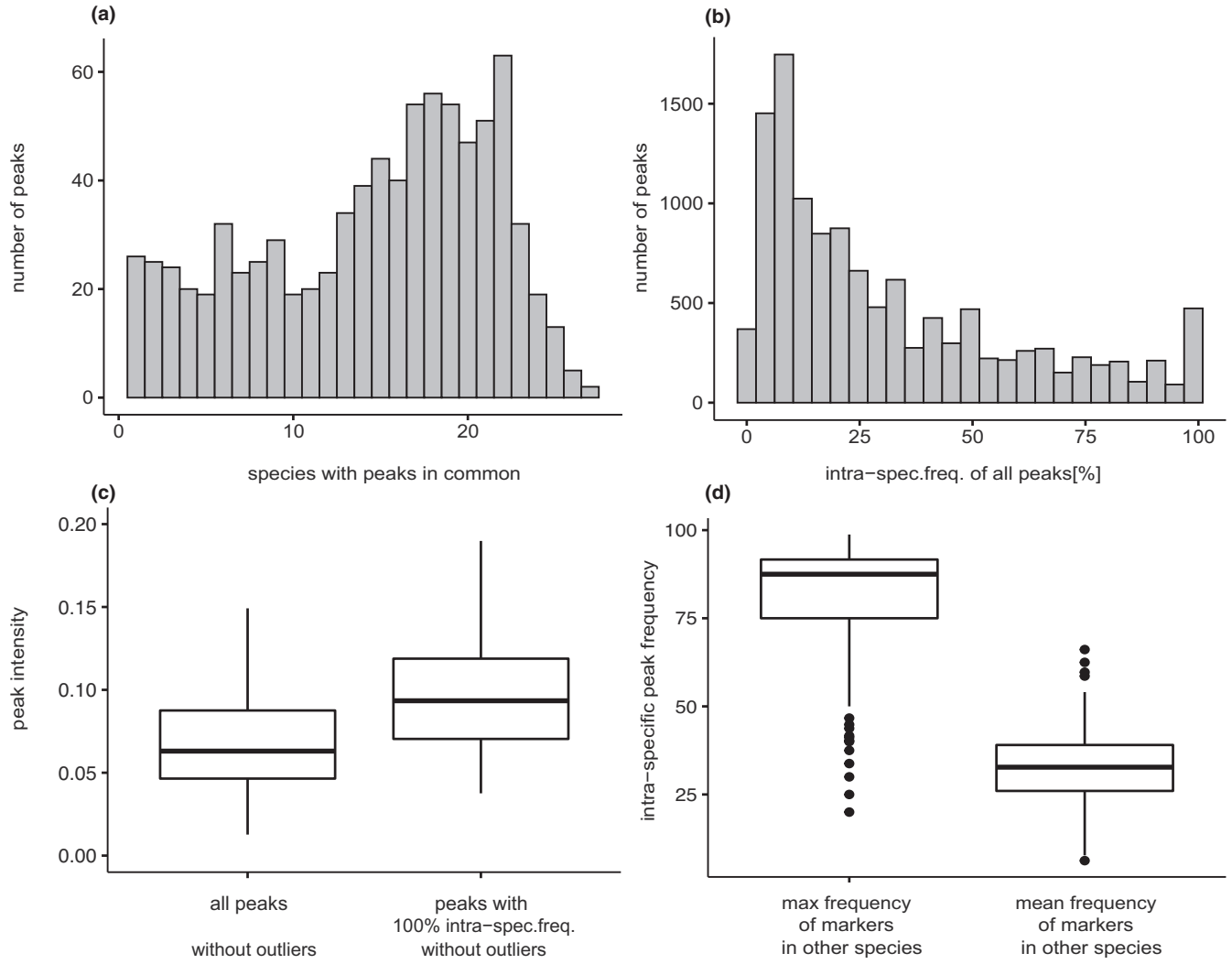

FIGURE 3 (a) Number of peaks, grouped by number of species with these peaks in common; (b) number of peaks, grouped by intraspecific frequency; (c) peak intensity as a boxplot (without outliers) for all peaks and for peaks with 100% intraspecific frequency; (d) maximum and mean intraspecific peak frequency of the 282 potential single markers in other species (100% intraspecific frequency in only one species).

in the whole data set decreased with increasing SNR from over 3000 to quite stable 600 from an SNR of 10 onwards. All specimens were correctly identified by the RF model at an SNR of 4 and an HWS of 3, but 70% of peaks were closely spaced with less than 6 Da distance. This proportion decreased to 10% for an HWS and SNR combination of 8. As misidentification was also low (<0.3%) this parameter setting was then used for all further analysis.

Overall, 838 different peaks from all species and specimens were included in the feature matrix for the analysis. Medium peak number per spectrum ranged between 50 and 227 with an average of around 140 peaks per individual. To search for ubiquitous compounds and to estimate the specificity and sensitivity of specific peaks, compound occurrence and abundance was analysed. Only two peaks were present in all species, albeit not in all specimens: *m/z* 3088 (in 44% of all specimens) and 3401 (in 70% of all specimens). Common peaks between species (disregarding varying intraspecific peak frequency) showed a left skewed distribution of occurrence, with 70% of all peaks occurring in more than 10 species (Figure 3a). In total, on 457 occasions we observed peaks with a 100% intraspecific frequency, that is they occurred in all specimens of a species (Figure 3b). Of these peaks, 282 were found with 100% in only one species and only one peak (*m/z* 3401) was found with 100% in a maximum of eight species. No peak with 100% frequency was observed in *T. longicornis*, *Ce. hamatus* and *A. longiremis*. These are the species in the data set with most included regions and/or regions with strong environmental variation in salinity and temperature. Potential marker peaks (i.e., those with 100% intraspecific frequency) were generally of higher intensity than average peaks (Figure 3c). Peaks with highest specificity (i.e., the 282 peaks with 100% intraspecific frequency in only one species) were compared for occurrence in other species, a measure of the sensitivity of potential markers. Mean intraspecific frequency of these peaks varied around 30% and maximum frequency around 80% (Figure 3d) in other species. Hence, although peaks with high specificity occurred, no single species-specific marker with strong

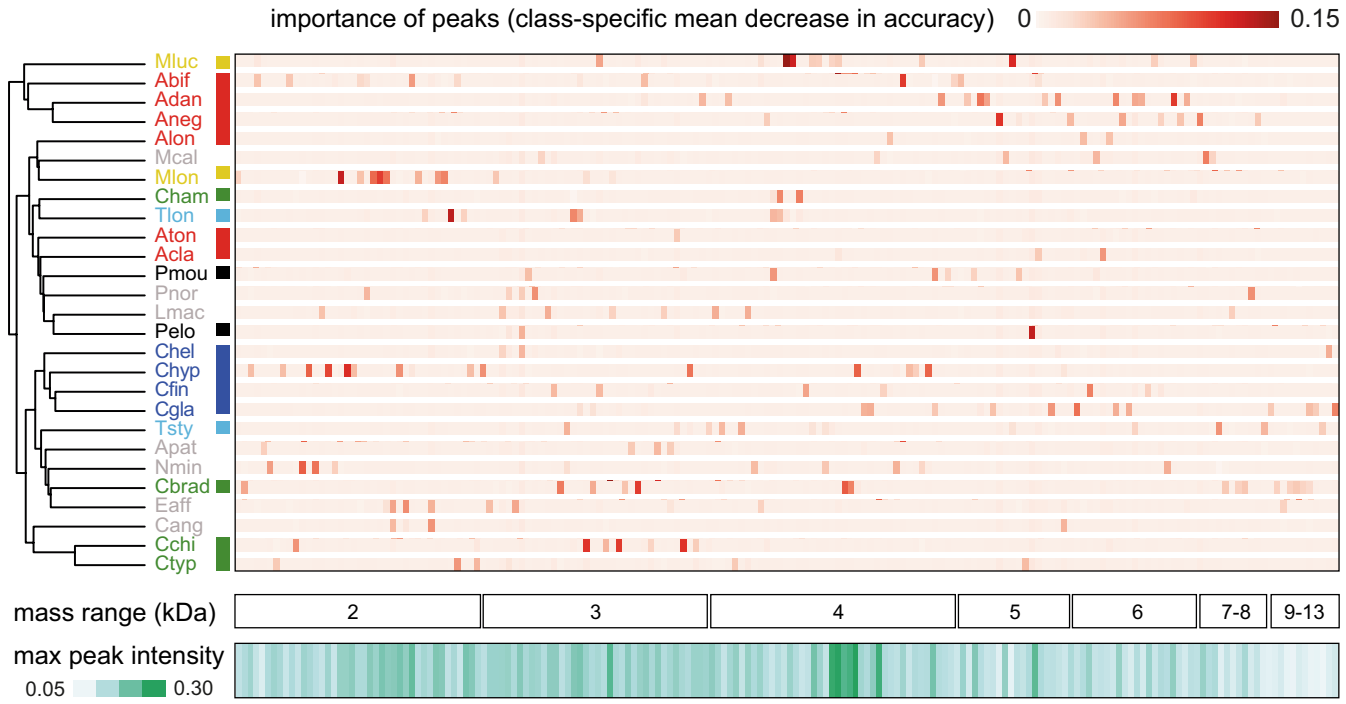

FIGURE 4 Heatmap of the 180 most important peaks for the species classification random forest model (peaks with maximum class-specific mean decrease in accuracy of >0.015 are presented); clustering of species is based on hierarchical clustering (Ward linkage) of the species-mean Euclidean distance based on the whole peak spectrum. The annotation gives the maximum peak intensity of the given *m/z* peak over the whole data set; heatmap scaling: 0–0.15 class-specific mean decrease in accuracy, maximum peak intensity is given in green, and boxes below the heatmap indicate the mass range (kDa) in which important peaks have been observed. Species included in this analysis: *Acartia bifilosa* (*Abif*), *A. clausi* (*Acla*), *A. danae* (*Adan*), *A. negligens* (*Aneg*), *A. longiremis* (*Alon*), *A. tonsa* (*Aton*), *Calanus finmarchicus* (*Cfin*), *Ca. helgolandicus* (*Chel*), *Ca. glacialis* (*Cgla*), *Ca. hyperboreus* (*Chyp*), *Centropages bradyi* (*Cbra*), *Ce. typicus* (*Ctyp*), *Ce. hamatus* (*Cham*), *Ce. chierchiae* (*Cchi*), *Metridia longa* (*Mlon*), *M. lucens* (*Mluc*), *Pseudocalanus elongatus* (*Pelo*), *P. moultoni* (*Pmou*), *Temora longicornis* (*Tlon*), *T. stylifera* (*Tsty*), *Paraeuchaeta norvegica* (*Pnor*), *Microcalanus* sp. (*Mcal*), *Anomalocera patersonii* (*Apat*), *Nannocalanus minor* (*Nmin*), *Eurytemora affinis* (*Eaff*), *Limnocalanus macrurus* (*Lmac*), *Corycaeus anglicus* (*Cang*).

sensitivity could be identified in the proteomic spectra of the copepods, when integrating over samples, seasons and regions.

Nevertheless, species identification was reliable using RF with only five misidentifications out of 743 (<1%). The 180 most important markers given by the class-specific mean decrease in accuracy (i.e., those peaks with high discriminatory power in the nodes of the decision trees) were extracted from the RF model (Figure 4). These discriminant peaks were quite evenly distributed over the whole *m/z* range of 2–13 kDa, also including peaks of different intensities. Generally, species-specific peaks were of lower importance in species that included specimens from many regions compared to species analysed in only one or two regions.

3.2 | Phylogenetic patterns

No strong multispecies clusters (based on species means) could be identified for a particular genus (Figure 4), although there is a tendency of some congener pairs to form common clusters, for example in *Calanus* and some *Acartia* and *Centropages* species. Prominent similarity was found between the congener pairs *A. danae* and *A. negligens*, as well as *Ce. chierchiae* and *Ce. typicus*. The PCoA on proteomic spectra of congener species revealed a general similarity between congeners of *Acartia*, *Calanus*, *Centropages*, *Metridia*, *Pseudocalanus* and *Temora* (Figure 5), but with overlap between congener groups.

3.3 | Stability of proteomic signals and choice of reference library

To evaluate the stability of proteomic signals of a species between regions and seasons we compared mean intra- and interspecific Euclidean distances (Figure 6). Intraspecific variability was lowest between animals from the same sample (i.e., from the same sampling event). It increased when animals from different sampling events within each region were included, reflecting the range of spectra within a region. This variance increased further when animals of the species from all regions were compared. The species gap between the lower 10% quantile of interspecific distances and the 90% quantile of intraspecific distances was quite prominent and large when only specimens from single samples were included. The threshold was around a Euclidean distance of 0.7. This species gap narrowed strongly with increasing intraspecific variance in multisample/

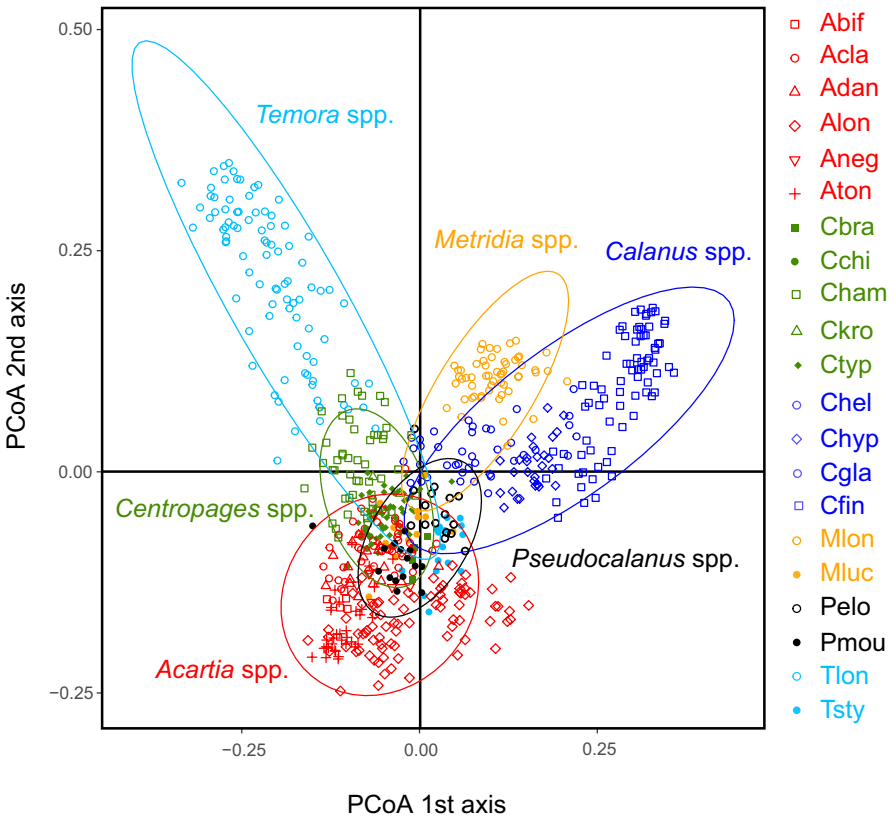

FIGURE 5 Principal coordinates analysis (PCoA) on proteomic spectra of congener species; species included: *Acartia bifilosa* (*Abif*), *A. clausi* (*Acla*), *A. danae* (*Adan*), *A. negligens* (*Aneg*), *A. longiremis* (*Alon*), *A. tonsa* (*Aton*), *Calanus finmarchicus* (*Cfin*), *Ca. helgolandicus* (*Chel*), *Ca. glacialis* (*Cgla*), *Ca. hyperboreus* (*Chyp*), *Centropages bradyi* (*Cbra*), *Ce. kroyeri* (*Ckro*), *Ce. typicus* (*Ctyp*), *Ce. hamatus* (*Cham*), *Ce. chierchiae* (*Cchi*), *Metridia longa* (*Mlon*), *M. lucens* (*Mluc*), *Pseudocalanus elongatus* (*Pelo*), *P. moultoni* (*Pmou*), *Temora longicornis* (*Tlon*), *T. stylifera* (*Tsty*).

season specimens and nearly closed when specimens from all regions were included.

To test whether the wide plasticity of proteomic profiles will be relevant for species identification and the choice of reference library, we determined the species classification success of RF models, excluding specimens from specific regions respectively (Table 1). The impact of locality of the reference library was very low for most species, except for Baltic Sea *A. longiremis* and *T. longicornis* (with 18% and 25% false positives, respectively), for North Sea *Ce. hamatus* (with nearly 50% false positives) and NW Atlantic *N. minor* (with 75% false positives). All these misidentifications were, however, reliably detected using the RF post hoc test. In contrast, post hoc test corrected error rates remained high for *Ce. typicus* and *A. danae*, with up to 60% of unrevealed false positives. While all *Ce. typicus* specimens from the North Sea were identified as *Ce. chierchiae*, specimens of *A. danae* from the Central East Atlantic were assigned to *A. negligens*. Application of the post hoc test resulted in overall higher rejection rates of identification, with up to 100% rejection in *N. minor*. Mean Euclidean distances within regions varied from 0.3 to 0.6, and the maximum observed distance between specimens from different regions was between 0.5 and 0.8 (Table 1). The latter distances were in the same range as the observed interspecific distances (Figure 6).

To evaluate variances between regions we compared the distances of the congener pairs *A. clausi* and *A. longiremis*, *T. stylifera* and *T. longicornis*, *Ca. hyperboreus* and *Ca. finmarchicus* as well as *Ce. typicus* and *Ce. hamatus* (Figure 7). Strongest homogeneity was observed for the *Calanus* species; only some specimens showed distances on the interspecific level (i.e., distances which can also been observed between specimens of different species). However, also for *Calanus* some substructures on a regional level occurred; that is, specimens from different regions were less similar to each other. These were much more distinct for *A. longiremis*, *Ce. hamatus* and *T. longicornis*. Here we observed differences in the proteomic spectrum between specimens from different regions that were nearly on the interspecies level. For *A. longiremis* more distinct subclusters could be identified, specifically for specimens from the Baltic Sea (see also Figure S1). Similarly, specimens for the Baltic Sea formed strong subclusters in *T. longicornis* and *Ce. hamatus*. North Sea specimens of *Ce. hamatus* showed two subgroups, one forming a distinct cluster and one clustering together with animals from the White Sea.

4 | DISCUSSION

4.1 | Species identification and species-specific markers

This study focused on characterizing the variability and stability of proteomic fingerprints using widely distributed, epipelagic copepod species from various coastal zooplankton communities around the North Atlantic as a model case.

In line with previous studies on marine copepods (Riccardi et al., 2012, Laakmann et al., 2013, Bode et al., 2017, Kaiser et al., 2018, Rossel & Martínez Arbizu, 2018a, Rossel et al., 2019, Renz et al., 2021, Yeom et al., 2021), our results clearly support the high discriminatory power of proteomic fingerprinting on the

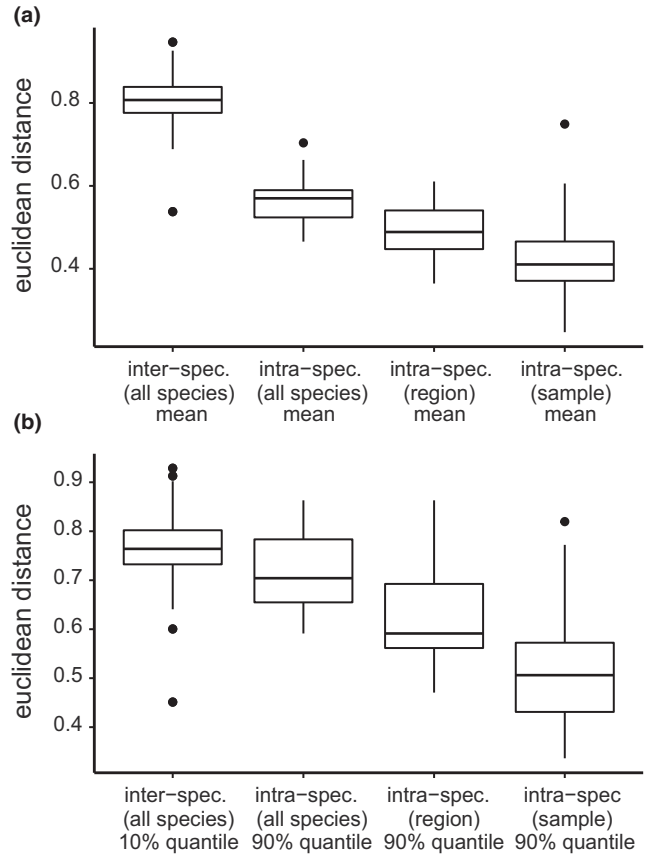

FIGURE 6 Boxplots based on species-specific means (upper panel) and 10% or 90% quantiles (lower panel) of Euclidean distances, providing interspecific distances and intraspecific distances based on specimens from different regions, from only the same region and the same sample, respectively.

species level in this taxonomic group. Nearly all specimens were correctly assigned to the different 27 species by an RF model for classification with only low sensitivity to data processing, indicating high robustness of the method. Many of the included species (e.g., *Acartia* spp. and *Calanus* spp.) can either only be separated with time-consuming morphological analyses, such as preparation of the fifth swimming leg, or by genetic analysis, such as the cryptic species of *Pseudocalanus*. Their reliable identification by proteomic fingerprinting again highlights the high value that MALDI-TOF MS could contribute to routine monitoring of marine communities, especially since most of the species investigated here are dominant at many North Atlantic monitoring sites.

In-depth analysis of proteomic spectra revealed that no discrete species-markers exist, and that identification is based on complex pattern differences rather than on the presence or absence of single compounds. Although several peaks show high specificity for a species, their sensitivity remains too low to serve as a single marker. The uniqueness of compounds in a species becomes increasingly blurred as more species and more specimens from different regions and seasons are included in the analysis. This has strong implications for the applicability of the method in multispecies research. Since the mere presence of individual species-specific peaks cannot serve as an indicator for the presence of a species in a bulk sample, multiplexing of proteomic fingerprints similar to metabarcoding appears to remain unlikely. Therefore, future application of this method in zooplankton monitoring will probably focus more on quantitative approaches as part of a bug-by-bug strategy and thereby improve species resolution and identification of challenging taxa. Given that the time- and cost-effectiveness of MALDI-TOF is much better than, for example, DNA barcoding (Renz et al., 2021; Rossel et al., 2019), it is nevertheless a powerful tool to accelerate biodiversity assessment and facilitate early and timely detection of changes in communities and ecosystems.

It is remarkable that apparently no conservative homologous compound was detected in all specimens and only two peaks were expressed in all species, although not in all specimens, but only in 44% and 70% of the animals, respectively. This variance in peak abundance may be due to genotypic or natural physiology-related variability, or probably at least partly to methodological or sample quality-related variability. One essential step in data processing is a slight shift of *m/z* values to align potential homologous peaks during binning and to account for observed small mass deviations caused by, for example, the relatively short trajectory of the Biotyper. As peak alignment is not independent of peak-neighbours, universal compounds may therefore be hidden in the fuzziness of the method. However, as we detected more than 280 peaks with 100% intraspecific frequency, and as more than 70% of peaks were detected in more than 10 of the 27 species, we argue that most homologous peaks were correctly identified in most cases and assume that this methodological effect is unlikely to be the main reason for an absence of universal peaks. No peak with 100% intraspecific frequency was found in those species with highest diversity in terms of included regions and environmental surroundings. This suggests that the observed variance in patterns is probably driven more by different genotypes or by phenotypic expression driven by physiological state.

Despite the absence of discrete peaks, species were reliably identified by RF based on several discriminant peaks. These were found in all mass ranges and intensities. In general, peaks with higher *m/z* values were of lower intensity. Consistent with the findings on intraspecific peak frequencies, the importance of species-specific peaks in the model decreased with increasing sample size. The disconnection between peak intensity and peak importance for species identification has previously been observed in other taxa, for example insects (Dieme et al., 2014; Müller et al., 2013) and crustaceans (Paulus et al., 2022).

4.2 | Phylogenetic patterns

We observed some phylogenetic structure within the data, such as highest similarities between certain congener pairs and an overall resemblance of congeneric species. Consistently, only very small deviations in proteomic pattern have been reported in cryptic species complexes (Dieme et al., 2014; Müller et al., 2013), specifically those with only recent speciation (Maasz et al., 2020; Paulus et al., 2022). However, when including all six calanoid genera with congeners in

TABLE 1 Classification success of specimens from different regions by random forest (RF) models (based on *Acartia bifilosa*, *A. clausi*, *A. danae*, *A. negligens*, *A. longiremis*, *A. tonsa*, *Calanus finmarchicus*, *Ca. helgolandicus*, *Ca. glacialis*, *Ca. hyperboreus*, *Centropages bradyi*, *Ce. kroyeri*, *Ce. typicus*, *Ce. hamatus*, *Ce. chierchiae*, *Metridia longa*, *M. lucens*, *Pseudocalanus elongatus*, *P. moultoni*, *Temora longicornis*, *T. sytlifera*, *Paraeuchaeta norvegica*, *Microcalanus* sp., *Anomalocera patersonii*, *Nannocalanus minor*, *Eurytemora affinis*) excluding specimens from the respective region; only species from multiple regions are presented.

	Region	*N*	RF error	RF error corr. (sign <.01)	Rejec. Rate	Mean dist. Intra-reg.	Mean dist inter-reg. (max)
Acartia clausi	MED	12	0.00	0.00	0.08	0.49	0.65
	NOS	28	0.00	0.00	0.11	0.45	0.59
	OFJ	8	0.00	0.00	0.00	0.35	0.65
Acartia danae	**NWA**	**6**	**0.33**	**0.33**	**0.67**	**0.45**	**0.71**
	CEA	**12**	**0.83**	**0.58**	**0.42**	**0.48**	**0.71**
Acartia longiremis	BFJ	12	0.00	0.00	0.08	0.35	0.73
	CAN	16	0.00	0.00	0.50	0.34	0.77
	CBS	34	0.18	0.00	0.85	0.55	0.77
	ICE	12	0.00	0.00	1.00	0.35	0.77
	OFJ	1	0.00	0.00	0.00	-	0.67
	WHS	6	0.00	0.00	0.00	0.34	0.74
Calanus finmarchicus	BAR	21	0.00	0.00	0.05	0.49	0.65
	BFJ	4	0.00	0.00	0.25	0.43	0.65
	CAN	4	0.00	0.00	0.25	0.39	0.58
	ICE	35	0.00	0.00	0.09	0.55	0.61
	NWA	12	0.00	0.00	0.00	0.39	0.63
	OFJ	1	0.00	0.00	0.00	–	0.60
Calanus glacialis	CAN	4	0.00	0.00	0.50	0.41	0.72
	ICE	1	0.00	0.00	0.00	–	0.72
	OFJ	2	1.00	0.00	1.00	0.44	0.72
	WHS	3	0.00	0.00	0.00	0.30	0.72
Calanus hyperboreus	CAN	12	0.08	0.00	0.75	0.47	0.63
	ICE	16	0.00	0.00	0.00	0.43	0.52
	NWA	1	0.00	0.00	1.00	–	0.63
Centropages hamatus	CBS	19	0.06	0.05	0.37	0.55	0.82
	NOS	28	0.46	0.00	0.75	0.61	0.80
	NWA	1	1.00	0.00	1.00	–	0.82
	OFJ	1	0.00	0.00	0.00	–	0.79
	WHS	4	0.25	0.00	0.50	0.46	0.77
Centropages typicus	**NOS**	**23**	**0.70**	**0.57**	**0.17**	**0.41**	**0.54**
	NWA	24	0.17	0.04	0.83	0.47	0.54
Metridia longa	BAR	12	0.00	0.00	0.25	0.52	0.62
	BFJ	1	0.00	0.00	0.00	–	0.62
	CAN	11	0.00	0.00	0.00	0.44	0.61
	ICE	6	0.00	0.00	0.00	0.32	0.55
	NWS	6	0.00	0.00	0.17	0.50	0.64
	OFJ	1	0.00	0.00	0.00	–	0.64
	WHS	19	0.00	0.00	0.00	0.43	0.62
Nannocalanus minor	NWA	12	0.75	0.08	0.92	0.41	0.63
	CEA	12	0.08	0.00	1.00	0.40	0.63

(Continues)

TABLE 1 (Continued)

	Region	N	RF error	RF error corr. (sign <.01)	Rejec. Rate	Mean dist. Intra-reg.	Mean dist inter-reg. (max)
Paraeuchaeta norvegica	BAR	4	0.25	0.00	0.75	0.51	0.66
	NWS	6	0.67	0.00	0.50	0.51	0.66
Temora longicornis	CAN	1	0.00	0.00	0.00	0.52	0.68
	CBS	41	0.10	0.00	0.15	0.44	0.78
	ICE	12	0.25	0.00	0.58	0.49	0.78
	NOS	12	0.00	0.00	0.08	0.52	0.73
	NWA	3	0.00	0.00	0.33	0.52	0.75
	OFJ	6	0.00	0.00	0.00	0.41	0.69
	WHS	12	0.00	0.00	0.00	0.46	0.69

Abbreviations: Fo abbreviations see Figure 1, *N*, number of specimens; RF error, rate of false negative specimens; RF error corr., rate of false negative specimens after a post hoc test for false positive discovery (Rossel & Martínez Arbizu, 2018a); rejec. rate, rate of specimen identified as false positive by a post hoc test; mean dist. intrareg., mean intraregional Euclidean distance; mean dist. inter-reg. (max), mean inter-regional Euclidean distance—only the highest distance within the species is given.

the analysis, similarity was not consistently related to phylogenetic distance and in some cases was higher between noncongeners than between congeners. Phylogenetic relationships have successfully been identified using proteomic composition (Maltseva et al., 2020; Telleria et al., 2010) and it was suggested that proteomic fingerprints may describe phylogenetic relationships (Zurita et al., 2019). However, our data indicate that proteomic fingerprints are not suitable to address phylogenetic questions in calanoid copepods. This makes sense as proteomic fingerprints are a potpourri of over 100 mainly cytosolic molecules with genes of quite different mutation rates behind them, also influenced by various physiological processes.

For most genera the higher similarity between congeners did not influence species identification success and is therefore probably not of practical relevance. However, the strongest misidentification while testing library robustness against regionality (i.e., the library did not include specimens from the respective region, but only from other regions), derived from the highly similar congener pairs from the same subgenus *A. danae* and *A. negligens*, as well as *Ce. typicus* and *Ce. chierchiae*. This misidentification was not resolvable by the post hoc test, which has been shown to detect false positives quite reliably (Rossel & Martínez Arbizu, 2018a). Since this only occurred when a nonregion-specific library was used, this problem may only be relevant to monitoring studies in which a rare species in the habitat or a neobiota of a very similar congener pair is not included in the library used. We have demonstrated here that the composition of the reference library can have a significant impact on the identification of closely related species and therefore needs to be thoroughly tested.

4.3 | Stability of proteomic signals

To specify a potential stable species gap in proteomic composition, we compared the intra- and interspecific variability of proteomic fingerprints using Euclidean distance as a measure. A distinct gap at a Euclidean distance of ~0.7 occurred when intraspecific variability was minimized by excluding variation between samples, seasons and regions. A similar threshold between inter- and intraspecific distances (i.e., 0.8) was observed, for example, for calanoid deep-sea copepods (Renz et al., 2021). Intraspecific variability increased when specimens from different regions or sampling seasons were included, leading to a stronger overlap of the maximal intraspecific and minimal interspecific distance. Sample history, for example sample storage conditions (temperature, pH, organic material in sample, etc.) and sample age may impact proteomic spectra (Rossel & Martínez Arbizu, 2018b). However, we suspect that the narrowing of the species gap is also driven by changes in proteomic spectra based on population-specific patterns and environment-induced variations in cell composition. In line with this interpretation of our data, proteomic patterns of mosquitos discriminated between colonies (Müller et al., 2013) and those of ticks varied with season and habitat (Karger et al., 2019). Strongest intraspecific distances, on the level of species distance, were observed between specimens from the brackish Baltic Sea and those from other regions, here *A. longiremis*, *Ce. hamatus* and *T. longicornis*. In these cases, species identification by means of a universal classification approach seems to come to its limits. A larger number of Baltic Sea specimens could not be determined unambiguously based on a database only containing North Atlantic animals. Apparently, salinity has a quite strong additive effect on proteomic patterns compared to other factors. This seems conclusive as copepods have been found to change protein expression not only under thermal stress (Rahlff et al., 2017) and over the seasonal cycle (Semmouri et al., 2020) but also under osmotic stressful conditions (DeBiasse et al., 2018). Copepods are able to osmoregulate and change the osmolarity of the haemolymph (Lee et al., 2012; Roddie et al., 1984). Specifically, marine copepods in brackish environments need to permanently control osmotic and cellular volume (Dutz & Christensen, 2018). Although most of the changes described so far in functional proteins are in the size fraction larger than measured

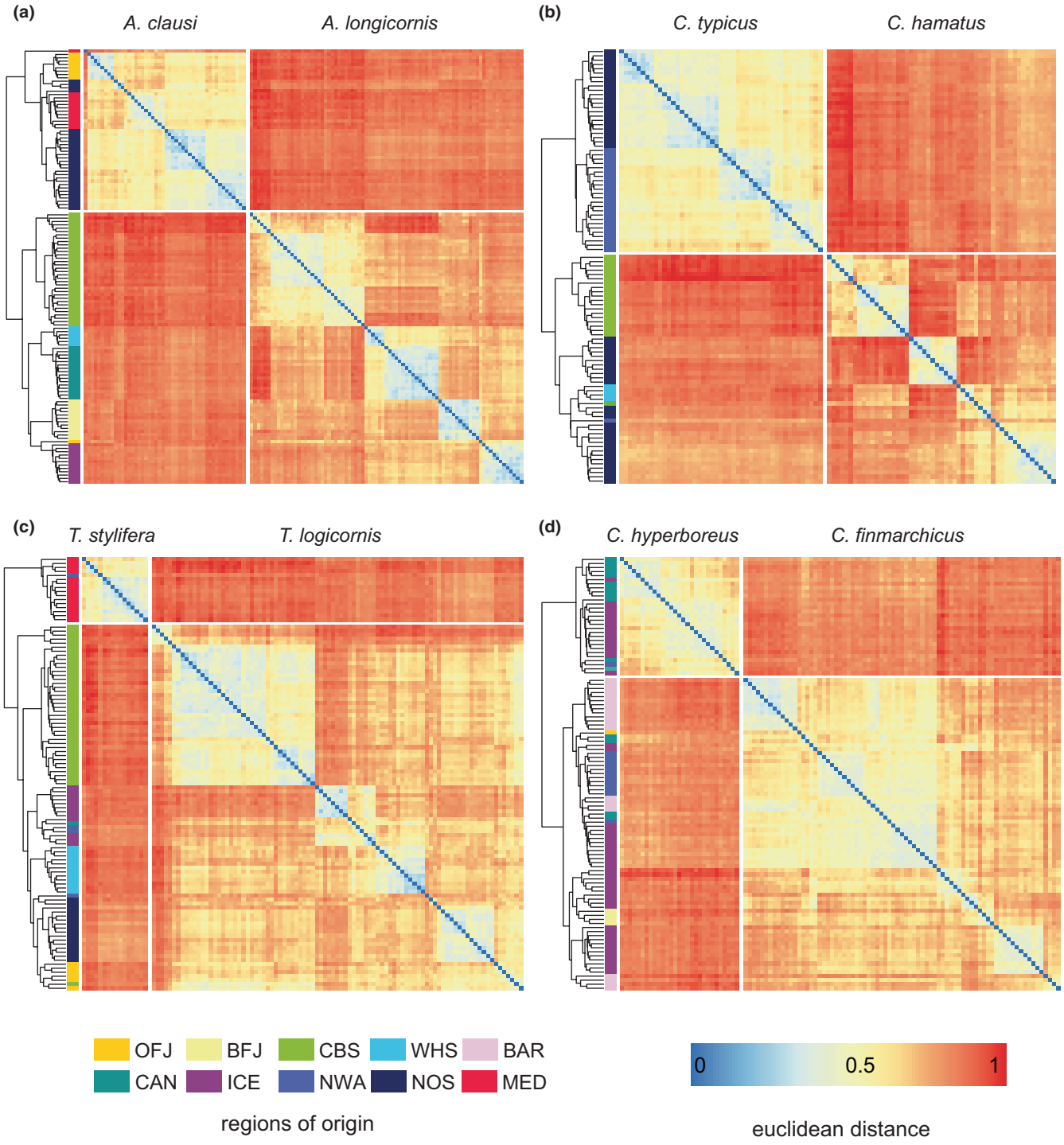

FIGURE 7 Heatmaps of Euclidean distance based on the proteomic spectrum of specimens from different regions (annotation, abbreviations see Figure 1), hierarchical clustering with average linkage; congener pairs included: *Acartia clausi* and *A. longiremis*, *Centropages typicus* and *Ce. hamatus*, *Temora stylifera* and *T. longicornis*, and *Calanus hyperboreus* and *Ca. finmarchicus*.

by MALDI-TOF MS, it seems realistic that changes in cell physiology will become visible to some extent in the proteomic fingerprint from 2 to 20 kDa.

In addition to a physiological response, the observed pattern could also be due to a population-specific genetic aspect. The Baltic Sea was suggested to act as a diversification hotspot to many native inhabitants (Geburzi et al., 2022) and a reduced gene flow between North and Baltic Sea populations was observed (Sjöqvist et al., 2015). To our knowledge, no information is available on population genetics and connectivity of Baltic *A. longiremis*, *T. longicornis* and *Ce. hamatus* with the North Sea, the North Atlantic and Arctic populations. Further field and experimental studies will be necessary and useful to disentangle the multiple effects of environment, ontogeny and underlying genetic differences on the expression of

proteomic spectra and to assess their impact on the ability to discriminate at the species level.

4.4 | Perspectives of proteomic fingerprinting for marine zooplankton monitoring studies

Proteomic fingerprinting has been widely used for pathogen screening in medicine (Croxatto et al., 2012) and has been extended from there to human vector identification (Murugaiyan & Roesler, 2017), food safety monitoring (Wang et al., 2013) and even palaeoproteomics (Sinet-Mathiot et al., 2019). MALDI-TOF-MS analyses have been successfully used in initial field surveys, demonstrating their general value for invertebrate monitoring (Müller et al., 2020). Nevertheless, the method is still far from being established and validated for biodiversity assessments or time-series approaches, and to our knowledge it has not yet found its way into standard protocols for metazoan monitoring programmes.

All recent findings on metazoans promise great potential, and proteomic fingerprinting has several advantages that argue for its own role in species identification. Sample processing is quite easy, fast and cost-efficient (Rossel et al., 2019) and measurement success rates are extremely high (Renz et al., 2021), if sample quality is sufficient (Rossel & Martínez Arbizu, 2018b, Rossel et al., 2021). These properties make the method a potential gap-filler in marine monitoring approaches despite the rapidly evolving use of single- and multi-species genetic methods. Standard morphological identification and counting procedures in zooplankton monitoring would not need to be changed as the approach is not intended to replace established routines but to provide additional rapid in-depth taxonomic resolution of specimens where needed. Formalin sampling is increasingly being replaced or supplemented by alcohol sampling in marine zooplankton research, opening the floor for many molecular approaches to add to classical counting, including proteomics. Formaldehyde fixation is not optimal for MALDI-TOF MS due to its covalent bonding with cellular proteins and need for additional processing. MALDI-TOF MS spectra were not obtained from formaldehyde-fixed bacteria samples, but no difference was observed between fresh and 70% ethanol-preserved samples (Williams et al., 2003). As a result, proteomic fingerprinting is better suited for future sample collections but less for older samples, unless new calibrations and methods are established.

Bearing an additional physiological signature, the fingerprint may provide information beyond the pure species name, such as on developmental stage (Karger et al., 2012; Laakmann et al., 2013; Rossel et al., 2023), gender (Lafri et al., 2016), environmental conditions (Karger et al., 2019) or feeding status (Hlavackova et al., 2019; Niare et al., 2017; Tandina et al., 2018). Successful detection of microplastics by MALDI-TOF MS (Adhikari et al., 2022) may even open future options for simultaneous detection of contamination. However, the method is still in its infancy, proteomic barcodes need to be established for most marine taxa, species delimitation models are under development and collateral information of proteomic signals still needs to be deciphered. Our understanding of marker variability is growing, and for marine copepods we show here that due to regional variability construction of a local database covering seasonal variability is strongly recommended. The impacts of library composition need to be thoroughly tested. The establishment of curated, open-access databases (ideally with web-based access), accompanied by the development of standardized, time-efficient data processing steps and adapted classification algorithms that allow central database queries (along the lines of GenBank or BOLD), will be a fundamental step in elevating the method from an experimental state to a standard applied procedure in marine science.

AUTHOR CONTRIBUTIONS

J.P., J.R. and S.L. designed the study and identified all taxa morphologically, J.P. and S.L. performed all molecular analysis, S.R. and M.-A. contributed to data analysis and data interpretation, and J.P. wrote the manuscript with significant contributions by all authors.

ACKNOWLEDGEMENTS

We are very grateful for the generous provision of samples by Ann Bucklin, Leo Blanco-Bercial, Astthor Gislason, Maiju Lehtiniemi, Tone Falkenhaug, Piotr Margonski, Lidia Yebra, Jens-Peter Herrmann, Jens Flöter, Gesche Winkler, Saskia Brix and Elena Markhasheva. We thank the two reviewers and the editor for valuable comments on the manuscript. This study was supported by the DFG initiative 1991 "Taxono-omics" (grant no. RE2808/3-1/2). HIFMB is a collaboration between the Alfred-Wegener-Institute, Helmholtz-Center for Polar and Marine Research, and the Carl-vonOssietzky University Oldenburg, initially funded by the Ministry for Science and Culture of Lower Saxony and the Volkswagen Foundation through the "Niedersächsisches Vorab" grant programme (grant no. ZN3285). 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. This is publication no. 19 of Senckenberg am Meer Proteome Laboratory. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors do not have conflicts of interests associated with the content of this study.

DATA AVAILABILITY STATEMENT

All data and applied scripts are available under: Peters et al., 2023; Proteomic spectra of epipelagic copepods; Dryad: https://doi.org/10.5061/dryad.jwstqjqf2

BENEFIT-SHARING STATEMENT

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

ORCID

Janna Peters https://orcid.org/0000-0002-0941-8257
Sven Rossel https://orcid.org/0000-0002-1187-346X
Jasmin Renz https://orcid.org/0000-0002-2658-445X

REFERENCES

Adhikari, S., Kelkar, V., Kumar, R., & Halden, R. U. (2022). Methods and challenges in the detection of microplastics and nanoplastics: A mini-review. *Polymer International*, *71*(5), 543–551. https://doi.org/10.1002/pi.6348

Balážová, T., Makovcová, J., Šedo, O., Slaný, M., Faldyna, M., & Zdráhal, Z. (2014). The influence of culture conditions on the identification of *mycobacterium* species by MALDI-TOF MS profiling. *FEMS Microbiology Letters*, *353*(1), 77–84. https://doi.org/10.1111/1574-6968.12408

Benkacimi, L., Gazelle, G., El Hamzaoui, B., Bérenger, J.-M., Parola, P., & Laroche, M. (2020). MALDI-TOF MS identification of *Cimex lectularius* and *Cimex hemipterus* bedbugs. *Infection, Genetics and Evolution*, *85*, 104536. https://doi.org/10.1016/j.meegid.2020.104536

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*(4), 600–617. https://doi.org/10.1093/plankt/fbx031

Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A., & Lehtiniemi, M. (2016). Metabarcoding of marine zooplankton: Prospects, progress and pitfalls. *Journal of Plankton Research*, *38*(3), 393–400. https://doi.org/10.1093/plankt/fbw023

Bucklin, A., Peijnenburg, K. T. C. A., Kosobokova, K. N., O'Brien, T. D., Blanco-Bercial, L., Cornils, A., Falkenhaug, T., Hopcroft, R. R., Hosia, A., Laakmann, S., Li, C., Martell, L., Questel, J. M., Wall-Palmer, D., Wang, M., Wiebe, P. H., & Weydmann-Zwolicka, A. (2021). Toward a global reference database of COI barcodes for marine zooplankton. *Marine Biology*, *168*(6), 78. https://doi.org/10.1007/s00227-021-03887-y

Bucklin, A., Yeh, H. D., Questel, J. M., Richardson, D. E., Reese, B., Copley, N. J., & Wiebe, P. H. (2019). Time-series metabarcoding analysis of zooplankton diversity of the NW Atlantic continental shelf. *ICES Journal of Marine Science*, *76*(4), 1162–1176. https://doi.org/10.1093/icesjms/fsz021

Campbell, P. M. (2005). Species differentiation of insects and other multicellular organisms using matrix-assisted laser desorption/ionization time of flight mass spectrometry protein profiling. *Systematic Entomology*, *30*, 186–190. https://doi.org/10.1111/j.1365-3113.2004.00279.x

Chavy, A., Nabet, C., Normand, A. C., Kocher, A., Ginouves, M., Prévot, G., Vasconcelos dos Santos, T., Demar, M., Piarroux, R., & de Thoisy, B. (2019). Identification of French Guiana sand flies using MALDI-TOF mass spectrometry with a new mass spectra library. *PLoS Neglected Tropical Diseases*, *13*(2), e0007031. https://doi.org/10.1371/journal.pntd.0007031

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. https://doi.org/10.1111/j.1574-6976.2011.00298.x

DeBiasse, M. B., Kawji, Y., & Kelly, M. W. (2018). Phenotypic and transcriptomic responses to salinity stress across genetically and geographically divergent *Tigriopus californicus* populations. *Molecular Ecology*, *27*(7), 1621–1632. https://doi.org/10.1111/mec.14547

Diarra, A. Z., Almeras, L., Laroche, M., Berenger, J.-M., Koné, A. K., Bocoum, Z., Dabo, A., Doumbo, O., Raoult, D., & Parola, P. (2017). Molecular and MALDI-TOF identification of ticks and tick-associated bacteria in Mali. *PLoS Neglected Tropical Diseases*, *11*(7), e0005762. https://doi.org/10.1371/journal.pntd.0005762

Dieme, C., Yssouf, A., Vega-Rúa, A., Berenger, J.-M., Failloux, A.-B., Raoult, D., Parola, P., & Almeras, L. (2014). Accurate identification of Culicidae at aquatic developmental stages by MALDI-TOF MS profiling. *Parasites & Vectors*, *7*(1), 544. https://doi.org/10.1186/s13071-014-0544-0

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., Montes, E., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B., & Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, *11*(1), 254. https://doi.org/10.1038/s41467-019-14105-1

Dutz, J., & Christensen, A. M. (2018). Broad plasticity in the salinity tolerance of a marine copepod species, Acartia longiremis, in the Baltic Sea. *Journal of Plankton Research*, *40*(3), 342–355. https://doi.org/10.1093/plankt/fby013

El Hamzaoui, B., Laroche, M., Almeras, L., Bérenger, J.-M., Raoult, D., & Parola, P. (2018). Detection of Bartonella spp. In fleas by MALDI-TOF MS. *PLoS Neglected Tropical Diseases*, *12*(2), e0006189. https://doi.org/10.1371/journal.pntd.0006189

Geburzi, J. C., Heuer, N., Homberger, L., Kabus, J., Moesges, Z., Ovenbeck, K., Brandis, D., & Ewers, C. (2022). An environmental gradient dominates ecological and genetic differentiation of marine invertebrates between the north and Baltic Sea. *Ecology and Evolution*, *12*(5), e8868. https://doi.org/10.1002/ece3.8868

Gibb, S. (2017). MALDIquantForeign: Import/Export routines for MALDIquant. A Package for R. https://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. https://doi.org/10.1093/bioinformatics/bts447

Gittens, R. A., Almanza, A., Bennett, K. L., Mejía, L. C., Sanchez-Galan, J. E., Merchan, F., Kern, J., Miller, M. J., Esser, H. J., Hwang, R., Dong, M., De León, L. F., Álvarez, E., & Loaiza, J. R. (2020). Proteomic fingerprinting of Neotropical hard tick species (Acari: Ixodidae) using a self-curated mass spectra reference library. *PLoS Neglected Tropical Diseases*, *14*(10), e0008849. https://doi.org/10.1371/journal.pntd.0008849

Goldstein, J. E., Zhang, L., Borror, C. M., Rago, J. V., & Sandrin, T. R. (2013). Culture conditions and sample preparation methods affect spectrum quality and reproducibility during profiling of *Staphylococcus aureus* with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Letters in Applied Microbiology*, *57*(2), 144–150. https://doi.org/10.1111/lam.12092

Hasnaoui, B., Diarra, A. Z., Berenger, J.-M., Medkour, H., Benakhla, A., Mediannikov, O., & Parola, P. (2022). Use of the proteomic tool MALDI-TOF MS in termite identification. *Scientific Reports*, *12*(1), 718. https://doi.org/10.1038/s41598-021-04574-0

Hlavackova, K., Dvorak, V., Chaskopoulou, A., Volf, P., & Halada, P. (2019). A novel MALDI-TOF MS-based method for blood meal identification in insect vectors: A proof of concept study on phlebotomine sand flies. *PLoS Neglected Tropical Diseases*, *13*(9), e0007669. https://doi.org/10.1371/journal.pntd.0007669

Hynek, R., Kuckova, S., Cejnar, P., Junková, P., Přikryl, I., & Říhová Ambrožová, J. (2018). Identification of freshwater zooplankton species using protein profiling and principal component analysis: Freshwater zooplankton protein profiling. *Limnology and Oceanography: Methods*, *16*(3), 199–204. https://doi.org/10.1002/lom3.10238

Kaiser, P., Bode, M., Cornils, A., Hagen, W., Arbizu, P. M., Auel, H., & Laakmann, S. (2018). High-resolution community analysis of deep-sea copepods using MALDI-TOF protein fingerprinting. *Deep Sea Research Part I: Oceanographic Research Papers*, *138*, 122–130. https://doi.org/10.1016/j.dsr.2018.06.005

Karger, A., Bettin, B., Gethmann, J. M., & Klaus, C. (2019). Whole animal matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry of ticks – Are spectra of Ixodes ricinus nymphs influenced by environmental, spatial, and temporal factors? *PLoS One*, *14*(1), e0210590. https://doi.org/10.1371/journal.pone.0210590

Karger, A., Kampen, H., Bettin, B., Dautel, H., Ziller, M., Hoffmann, B., Süss, J., & Klaus, C. (2012). Species determination and characterization of developmental stages of ticks by whole-animal matrix-assisted laser desorption/ionization mass spectrometry. *Ticks and Tick-Borne Diseases*, *3*(2), 78–89. https://doi.org/10.1016/j.ttbdis.2011.11.002

Laakmann, S., Blanco-Bercial, L., & Cornils, A. (2020). The crossover from microscopy to genes in marine diversity: From species to assemblages in marine pelagic copepods. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *375*(1814), 20190446. https://doi.org/10.1098/rstb.2019.0446

Laakmann, S., Gerdts, G., Erler, R., Knebelsberger, T., Martínez Arbizu, P., & Raupach, M. J. (2013). Comparison of molecular species identification for North Sea calanoid copepods (Crustacea) using proteome fingerprints and DNA sequences. *Molecular Ecology Resources*, *13*(5), 862–876. https://doi.org/10.1111/1755-0998.12139

Lafri, I., Almeras, L., Bitam, I., Caputo, A., Yssouf, A., Forestier, C.-L., Izri, A., Raoult, D., & Parola, P. (2016). Identification of Algerian field-caught Phlebotomine sand Fly vectors by MALDI-TOF MS. *PLoS Neglected Tropical Diseases*, *10*(1), e0004351. https://doi.org/10.1371/journal.pntd.0004351

Lawrence, A. L., Batovska, J., Webb, C. E., Lynch, S. E., Blacket, M. J., Šlapeta, J., Parola, P., & Laroche, M. (2019). Accurate identification of Australian mosquitoes using protein profiling. *Parasitology*, *146*(4), 462–471. https://doi.org/10.1017/S0031182018001658

Lee, C. E., Posavi, M., & Charmantier, G. (2012). Rapid evolution of body fluid regulation following independent invasions into freshwater habitats: Evolution of body fluid regulation. *Journal of Evolutionary Biology*, *25*(4), 625–633. https://doi.org/10.1111/j.1420-9101.2012.02459.x

Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, *129*(2), 271–280. https://doi.org/10.1007/s004420100716

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. http://CRAN.R-project.org/doc/Rnews/

Maasz, G., Takács, P., Boda, P., Varbiro, G., & Pirger, Z. (2017). Mayfly and fish species identification and sex determination in bleak (*Alburnus alburnus*) by MALDI-TOF mass spectrometry. *Science of the Total Environment*, *601–602*, 317–325. https://doi.org/10.1016/j.scitotenv.2017.05.207

Maasz, G., Zrínyi, Z., Fodor, I., Boross, N., Vitál, Z., Kánainé Sipos, D. I., Kovács, B., Melegh, S., & Takács, P. (2020). Testing the applicability of MALDI-TOF MS as an alternative stock identification method in a cryptic species complex. *Molecules*, *25*(14), 3214. https://doi.org/10.3390/molecules25143214

Maltseva, A. L., Varfolomeeva, M. A., Lobov, A. A., Tikanova, P., Panova, M., Mikhailova, N. A., & Granovitch, A. I. (2020). Proteomic similarity of the Littorinid snails in the evolutionary context. *PeerJ*, *8*, e8546. https://doi.org/10.7717/peerj.8546

Müller, P., Engeler, L., Vavassori, L., Suter, T., Guidi, V., Gschwind, M., Tonolla, M., & Flacio, E. (2020). Surveillance of invasive Aedes mosquitoes along Swiss traffic axes reveals different dispersal modes for Aedes albopictus and ae. Japonicus. *PLoS Neglected Tropical Diseases*, *14*(9), e0008705. https://doi.org/10.1371/journal.pntd.0008705

Müller, P., Pflüger, V., Wittwer, M., Ziegler, D., Chandre, F., Simard, F., & Lengeler, C. (2013). Identification of cryptic *anopheles* mosquito species by molecular protein profiling. *PLoS One*, *8*(2), e57486. https://doi.org/10.1371/journal.pone.0057486

Murugaiyan, J., & Roesler, U. (2017). MALDI-TOF MS profiling-advances in species identification of pests, parasites, and vectors. *Frontiers in Cellular and Infection Microbiology*, *7*, 184. https://doi.org/10.3389/fcimb.2017.00184

Ngoy, S., Diarra, A. Z., Laudisoit, A., Gembu, G.-C., Verheyen, E., Mubenga, O., Mbalitini, S. G., Baelo, P., Laroche, M., & Parola, P. (2021). Using MALDI-TOF mass spectrometry to identify ticks collected on domestic and wild animals from The Democratic Republic of the Congo. *Experimental and Applied Acarology*, *84*(3), 637–657. https://doi.org/10.1007/s10493-021-00629-z

Niare, S., Almeras, L., Tandina, F., Yssouf, A., Bacar, A., Toilibou, A., Doumbo, O., Raoult, D., & Parola, P. (2017). MALDI-TOF MS identification of Anopheles gambiae Giles blood meal crushed on Whatman filter papers. *PLoS One*, *12*(8), e0183238. https://doi.org/10.1371/journal.pone.0183238

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2019). *Vegan: Community ecology package*. (version 2.5-6.) [R package]. https://CRAN.R-project.org/package=vegan

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Paulus, E., Brix, S., Siebert, A., Martínez Arbizu, P., Rossel, S., Peters, J., Svavarsson, J., & Schwentner, M. (2022). Recent speciation and hybridization in Icelandic deep-sea isopods: An integrative approach using genomics and proteomics. *Molecular Ecology*, *31*(1), 313–330. https://doi.org/10.1111/mec.16234

Rahlff, J., Peters, J., Moyano, M., Pless, O., Claussen, C., & Peck, M. A. (2017). Short-term molecular and physiological responses to heat stress in neritic copepods Acartia tonsa and Eurytemora affinis. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *203*, 348–358. https://doi.org/10.1016/j.cbpa.2016.11.001

Renz, J., Markhaseva, E. L., Laakmann, S., Rossel, S., Martinez Arbizu, P., & Peters, J. (2021). Proteomic fingerprinting facilitates biodiversity assessments in understudied ecosystems: A case study on integrated taxonomy of deep sea copepods. *Molecular Ecology Resources*, *21*(6), 1936–1951. https://doi.org/10.1111/1755-0998.13405

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. https://doi.org/10.1093/plankt/fbs022

Roddie, B. D., Leakey, R. J. G., & Berry, A. J. (1984). Salinity-temperature tolerance and osmoregulation in *syx* (Poppe) (Copepoda: Calanoida) in relation to its distribution in the zooplankton of the upper reaches of the forth estuary. *Journal of Experimental Marine Biology and Ecology*, *79*(2), 191–211. https://doi.org/10.1016/0022-0981(84)90219-3

Rossel, S., Barco, A., Kloppmann, M., Martínez Arbizu, P., Huwer, B., & Knebelsberger, T. (2021). Rapid species level identification of fish eggs by proteome fingerprinting using MALDI-TOF MS. *Journal of Proteomics*, *231*, 103993. https://doi.org/10.1016/j.jprot.2020.103993

Rossel, S., Kaiser, P., Bode-Dalby, M., Renz, J., Laakmann, S., Auel, H., Hagen, W., Arbizu, P. M., & Peters, J. (2023). Proteomic fingerprinting enables quantitative biodiversity assessments of species and ontogenetic stages in Calanus congeners (Copepoda, Crustacea) from the Arctic Ocean. *Molecular Ecology Resources*, *23*, 382–395. https://doi.org/10.1111/1755-0998.13714

Rossel, S., Khodami, S., & Martínez Arbizu, P. (2019). Comparison of rapid biodiversity assessment of Meiobenthos using MALDI-TOF MS and Metabarcoding. *Frontiers in Marine Science*, *6*, 659. https://doi.org/10.3389/fmars.2019.00659

Rossel, S., & Martínez Arbizu, P. (2018a). Effects of sample fixation on specimen identification in biodiversity assemblies based on proteomic data (MALDI-TOF). *Frontiers in Marine Science*, *5*, 149. https://doi.org/10.3389/fmars.2018.00149

Rossel, S., & Martínez Arbizu, P. (2018b). 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*(6), 1421–1434. https://doi.org/10.1111/2041-210X.13000

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. https://doi.org/10.1038/s41598-019-45718-7

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., & Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *34*(3), 396–402. https://doi.org/10.1016/0168-583X(88)90063-8

Ryzhov, V., & Fenselau, C. (2001). Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Analytical Chemistry*, *73*(4), 746–750. https://doi.org/10.1021/ac0008791

Semmouri, I., De Schamphelaere, K. A. C., Van Nieuwerburgh, F., Deforce, D., Janssen, C. R., & Asselman, J. (2020). Spatio-temporal patterns in the gene expression of the calanoid copepod *Temora longicornis* in the Belgian part of the North Sea. *Marine Environmental Research*, *160*, 105037. https://doi.org/10.1016/j.marenvres.2020.105037

Sinet-Mathiot, V., Smith, G. M., Romandini, M., Wilcke, A., Peresani, M., Jean-Jacques Hublin, J.-J., & Welker, F. (2019). Combining ZooMS and zooarchaeology to study late Pleistocene hominin behaviour at Fumane (Italy). *Scientific Reports*, *9*, 12350. https://doi.org/10.1038/s41598-019-48706-z

Sjöqvist, C., Godhe, A., Jonsson, P. R., Sundqvist, L., & Kremp, A. (2015). Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea-Baltic Sea salinity gradient. *Molecular Ecology*, *24*(11), 2871–2885. https://doi.org/10.1111/mec.13208

Suter, L., Polanowski, A. M., Clarke, L. J., Kitchener, J. A., & Deagle, B. E. (2021). Capturing open ocean biodiversity: Comparing environmental DNA metabarcoding to the continuous plankton recorder. *Molecular Ecology*, *30*(13), 3140–3157. https://doi.org/10.1111/mec.15587

Tandina, F., Niaré, S., Laroche, M., Koné, A. K., Diarra, A. Z., Ongoiba, A., Berenger, J. M., Doumbo, O. K., Raoult, D., & Parola, P. (2018). Using MALDI-TOF MS to identify mosquitoes collected in Mali and their blood meals. *Parasitology*, *145*(9), 1170–1182. https://doi.org/10.1017/S0031182018000070

Telleria, J., Biron, D. G., Brizard, J.-P., Demettre, E., Séveno, M., Barnabé, C., Ayala, F. J., & Tibayrenc, M. (2010). Phylogenetic character mapping of proteomic diversity shows high correlation with subspecific phylogenetic diversity in *Trypanosoma cruzi*. *Proceedings of the National Academy of Sciences*, *107*(47), 20411–20416. https://doi.org/10.1073/pnas.1015496107

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), 17. https://doi.org/10.4081/jlimnol.2012.e17

Wang, X., Wang, S., & Cai, Z. (2013). The latest developments and applications of mass spectrometry in food-safety and quality analysis. *Trends in Analytical Chemistry*, *52*, 170–185.

Williams, T. L., Andrzejewski, D., Lay, J. O., & Musser, S. M. (2003). Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. *Journal of the American Society for Mass Spectrometry*, *14*(4), 342–351. https://doi.org/10.1016/S1044-0305(03)00065-5

Yeom, J., Park, N., Jeong, R., & Lee, W. (2021). Integrative description of cryptic *Tigriopus species* from Korea using MALDI-TOF MS and DNA barcoding. *Frontiers in Marine Science*, *8*, 648197. https://doi.org/10.3389/fmars.2021.648197

Zurita, A., Djeghar, R., Callejón, R., Cutillas, C., Parola, P., & Laroche, M. (2019). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry as a useful tool for the rapid identification of wild flea vectors preserved in alcohol. *Medical and Veterinary Entomology*, *33*(2), 185–194. https://doi.org/10.1111/mve.1235

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: Peters, J., Laakmann, S., Rossel, S., Martínez Arbizu, P., & Renz, J. (2023). Perspectives of species identification by MALDI-TOF MS in monitoring—Stability of proteomic fingerprints in marine epipelagic copepods. *Molecular Ecology Resources*, *23*, 1077–1091. https://doi.org/10.1111/1755-0998.13779