A universal tool for marine metazoan species identification – Towards best practices in proteomic fingerprinting

Sven Rossel[1], Janna Peters[2], Nele Charzinski[3], Angelina Eichsteller[4], Silke Laakmann[5], Hermann Neumann[6], and Pedro Martinez[4]

[1]Senckenberg am Meer Deutsches Zentrum fur Marine Biodiversitatsforschung
[2]Senckenberg am Meer Deutsches Zentrum für Marine Biodiversitätsforschung
[3]Carl von Ossietzky Universitat Oldenburg
[4]Senckenberg am Meer
[5]Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg
[6]Johann Heinrich von Thünen-Institut für Seefischerei

June 14, 2023

Abstract

Proteomic fingerprinting using MALDI-TOF mass spectrometry is a well-established tool for identifying microorganisms and has shown promising results for identification of animal species, particularly disease vectors and marine organisms. And thus can be a vital tool for biodiversity assessments in ecological studies. However, few studies have tested species identification across different orders and classes. In this study, we collected data from 1,246 specimens and 198 species to test species identification in a diverse dataset. We also evaluated different specimen preparation and data processing approaches for machine learning and developed a workflow to optimize classification using random forest. Our results showed high success rates of over 90%, but we also found that the size of the reference library affects classification error. Additionally, we demonstrated the ability of the method to differentiate marine cryptic-species complexes and to distinguish sexes within species.

Introduction

Correct and cost-effective species identification is crucial in various research areas, including biodiversity assessments, where obtaining reliable information on species' occurrences and distributions is pivotal. If species cannot be morphologically identified to the species level, they are often assigned to higher taxonomic levels, leading to less detailed analyses and consequently imprecise conclusions (Bailey et al., 2001; Timms et al., 2013). However, identification of samples using COI-barcoding is expensive, time consuming (Rossel et al., 2019) and therefore not feasible in large biodiversity assessments including large numbers of specimens.

Matrix Assisted Laser Desorption/Ionization Time-of-flight mass spectrometry (MALDI-TOF MS) is a rapid species identification method that measures a proteome fingerprint to identify specimens using a reference library. With few preparation steps, peptides and proteins are extracted from tissue and embedded in a matrix absorbing laser radiation while measuring ionized, intact compounds in a mass spectrometer (Singhal et al., 2015). This method is routinely applied for the identification of microorganisms such as bacteria, viruses and fungi (Fenselau and Demirev, 2001; Sandrin et al., 2013; Calderaro et al., 2014). It was also used for food fraud detection (Mazzeo et al., 2008; Flaudrops et al., 2015) or to check food adulteration (Sassi et al., 2015). In pilot studies, it was successfully applied for identification of metazoans such as copepods (Laakmann et al., 2013; Kaiser et al., 2018; Rossel and Martínez Arbizu, 2019; Renz et al., 2021; Yeom et al., 2021; Rossel et al., 2022; Peters et al., 2023), isopods (Kürzel et al., 2022; Paulus et al., 2022), different groups of Cnidaria (Holst et al., 2019; Park et al., 2021; Korfhage et al., 2022), molluscs (Wilke et al., 2020), fish (Mazzeo et al., 2008; Volta et al., 2012) and especially disease vectors such as ticks, sandflies or mosquitoes (Yssouf et al., 2013; Chavy et al., 2019; Rakotonirina et al., 2020, 2022; Nabet et al., 2021). Most studies only analyzed a few species or were limited to a certain taxonomic group while studies across different classes and phyla are still missing. Also, no gold standard protocol for metazoan analytics has been established yet. Systematic tests, how data processing will affect the identification success and whether and how pipelines need to be adapted to higher-taxonomic-level identification are also missing.

For the first time, we present a generalized workflow for species identification of metazoans as well as the subsequent bioinformatics using a wide spectrum of marine taxa. We emphasize the importance of adjusting bioinformatics to the data set and finally prove the power of proteomic fingerprinting for differentiation of morphologically cryptic, closely related marine species and beyond mere species identification on sex level, making it a promising tool for ecological research.

Methods

Sample material

Tissue for measurements was taken mainly from the marine organisms tissue bank of the Senckenberg am Meer, German Centre for Marine Biodiversity Research, which was established using samples from numerous studies (Knebelsberger and Thiel, 2014; Knebelsberger et al., 2014; Markert et al., 2014; Gebhardt and Knebelsberger, 2015; Raupach et al., 2015; Barco et al., 2016; Laakmann et al., 2016; Rossel et al., 2020b) (supplementary table S1 for accession numbers) on North Sea metazoans. The material from this collection was taken from specimens processed for COI-barcoding to create reference libraries for a variety of marine animal groups. During this process, tissue samples of the respective specimens were stored in ethanol at -80°C. Tissue samples were available for Bivalvia (muscle, 18 species), Cephalopoda (muscle from arm, 12 species), Gastropoda (muscle from foot, 24 species), Polyplacophora (muscle from foot, 2 species), Ascidiacea (tissue, 1 species), Teleostei (muscle, 67 species), Elasmobranchii (muscle, 7 species), Malacostraca (muscle from foot or chelae, 39 species), Thecostraca (muscle from foot, 1 species), Pycnogonida (leg fragment, 1 species), Asteroidea (tube feet, 10 species), Ophiuroidea (tissue from arm, 10 species) and Echinoidea (tissue from the base of the tubercle, 6 species) ($n_{species}$= 198, $n_{specimens}$=1,246).

Sample preparation

The basic protocol of sample preparation was the same for all analyzed tissue samples. A very small tissue fragment (< 1 mm^3) was incubated for 5 minutes in HCCA as a saturated solution in 50% acetonitrile, 47.5% molecular grade water and 2.5% trifluoroacetic acid. Tissue from crustacean *Cancer pagurus* Linnaeus, 1758, the fish *Clupea harengus* Linnaeus, 1758, the cephalopod *Eledone cirrhosa* (Lamarck, 1798) and the echinoderm *Stichastrella rosea* (O.F. Müller, 1776) was used to find an optimal tissue to HCCA matrix ratio. Tissue was weighed on a METTLER TOLEDO XS3DU micro-balance and the amount of matrix was adjusted to tissue weight to obtain the desired ratios ranging from 0.012 µg µl^{-1} to 200 µg µl^{-1}. After incubation, 1.5 µl of the solution was transferred to 10 spots on a target plate, respectively. Mass spectra were measured with a Microflex LT/SH System (Bruker Daltonics) using method MBTAuto. Peak evaluation was carried out in a mass peak range between 2 k – 10 k Dalton (Da) using a centroid peak detection algorithm, a signal to noise threshold of 2 and a minimum intensity threshold of 600. To create a sum spectrum, 160 satisfactory shots were summed up.

Resulting from observations during this initial test, a fast applicable protocol was developed without the need to weigh each tissue sample. Matrix volume was added to tissue samples depending on tissue volume, i.e. tissue samples were always completely covered by HCCA matrix with a small layer (ca. 1 mm) of supernatant. Samples were incubated for 5 minutes and 1.5 µl of the solution were transferred to a single spot on a target plate for measurement. Each spot was measured between two to three times.

Mass spectra processing in R

Mass spectra data was imported to R (R-Core-Team, 2022) using MALDIquantForeign (Gibb, 2015) and further processed using MALDIquant (Gibb and Strimmer, 2012). Mass spectra were trimmed to an identical length from 2 to 20 kDa. Subsequently, spectra were square root transformed, smoothed using Savitzky Golay method (Savitzky and Golay, 1964), baseline corrected using SNIP approach (Ryan et al., 1988) and normalized using total ion current (TIC) method.

Spectra were quality controlled using the command 'screenSpectra' from the R-package MALDIrppa (Palarea-Albaladejo et al., 2017). Mass spectra with a notably high a-score were checked by eye and discarded if mass spectra were of bad quality. If due to this, only a single specimen for a certain species was retained, the remaining specimen was discarded from the data set.

Evaluation of random forest model for identification

Besides initial sample preparation and subsequent data processing, we tested how to improve a random forest (RF) model used for species identification. Optimal number of trees and variables was tested in a previous study (Rossel and Martínez Arbizu, 2018a). Here we assessed the effect of minimum number of specimens per species category on the resulting model power. We sampled the dataset using two to 11 specimens per species including only species with at least 11 specimens per class (n=20). For each minimum number of specimens, 100 data sets were sampled using 'sample_n' from R-package dplyr (Wickham et al., 2022), a RF model was created and the OOB errors assessed accordingly.

Standardization of data processing

Based on literature research and own observations, three data processing steps were identified, which may have a severe impact on data and the resulting quality of a random Forest (RF) classification model (Breimann, 2001). I) Iterations of baseline subtraction: this is a first manipulation step to reduce chemical noise and is carried out iteratively (Martínez Arbizu and Rossel, 2018). Increasing iterations will result in loss of low intensity peaks. II) Signal to noise ratio (SNR) during peak picking: an increase in SNR will exclude signals of low intensity. The higher the SNR value, the less peaks will be kept. III) Half window size (HWS) during peak picking: within the HWS the peak with the highest intensity will be chosen as the resulting peak during peak picking. The higher the HWS is chosen, the less peaks will be picked across an entire mass spectrum range.

Interactive effects of these data processing steps were tested using the classification success by a random forest model as target variable: iterations of baseline estimation and peak detection HWS were varied both between 5 to 30 and SNR from 3 to 20. In total, 12,186 analyses were carried out. In all cases, peak binning using 'binPeaks' from R-package MaldiQuant was repeated until the number of variables in the data did not further change. The RF model (ntree=2,000 and mtry=35) was trained on the Hellinger transformed peak intensities as suggested by Rossel and Martínez Arbizu (2018a). The RF out-of-box (OOB) error was used as measure for classification success. For these analyses, based on the results from RF-model evaluation, only species were included with at least six specimens. To investigate the main drivers of classification success, a generalized additive model (GAM, family: binomial; link function: logit) was calculated.

Testing the classification success

In concordance with the results from the previous tests, only species with at least six specimens were included in the model. Mass spectra from these species were processed according to the results from the test on variation of HWS (7), SNR (3) and baseline iteration (22). To test the overall classification success on species level, single specimens were separated from the RF training data set and subsequently identified using this model. After classification, the *post-hoc* test by Rossel and Martínez Arbizu (Martínez Arbizu and Rossel, 2018; Rossel and Martínez Arbizu, 2018a) using the R-package RFtools (*https://github.com/pmartinezarbizu/RFtools*) was applied to verify RF classification. This *post-hoc* test uses the empirical distribution of RF assignment probabilities from the RF model and compares the assignment probabilities of newly classified specimens to this distribution. Whereas classified specimens with assignment probabilities falling within this empirical distribution are considered true positive (tp), specimens with probabilities of assignment significantly different to this distribution considered false positive (fp).

Case studies

In order to show the applicability of MALDI-TOF MS, we present two model cases. First, data of the North Sea starfish *Astropecten irregularis* (Pennant, 1777) were investigated based on MALDI-TOF mass spectra. This species was found to be genetically divergent (Laakmann et al., 2016) while revealing a high morphological similarity. Differentiation of species was tested using RF models. Furthermore, data on the crustacean *Euterpina acutifrons* (Dana, 1847) from Rossel and Martínez Arbizu, 2019 was analyzed to show the applicability for sex level differentiation using hierarchical clustering and RF. Based on the Gini index the 30 most important peaks for species/sex differentiation in a RF model were extracted and investigated to show the expression within the respective groups.

Phyla and class models for identification

To test whether specimen can be identified on an above-species level, a RF model containing only class- and phylum-categories was applied. All spectra from species to be classified were excluded from the model to evaluate its use for specimen not included in a library. The respective specimens were identified using the model and the predicted class/phylum was tested with the RF *post-hoc* test. To test classification on phylum level, 1,246 specimens from 198 species were included. On class level, 1,227 specimens from 195 species were analyzed.

Results

The data set contained 1,246 specimens from 198 taxa including echinoderms (Asteroidea, Echinoidea and Ophiuroidea), molluscs (Bivalvia, Gastropoda, Polyplacophora and Cephalopoda), arthropods (Crustacea, Pantopoda) and chordates (Tunicata, Vertebrata: Teleostei, Elasmobranchii). For 1,139 specimens a published COI barcode or another molecular identifier is available (supplementary table S1). The remaining specimens were identified morphologically. For 226 specimens attempts to obtain mass spectra either failed or were of minor quality and discarded.

Sample preparation

To determine the concentration range for successful measurements, weighted tissue samples were mixed with varying amounts of α-cyano-4-hydroxycinnamic acid (HCCA). In total, 15 different tissue/matrix concentrations were tested ranging from 0.01 µg µl^{-1} to 200 µg µl^{-1} (Fig. 1A). Despite variations between samples, high quality mass spectra were generally assessed in a concentration range from to 3.1 µg µl^{-1} to 12.5 µg µl^{-1}. The largest concentration range for successful measurements was recorded for the echinoderm *Stichastrella rosea* (sample MT03612), with successful measurements across almost the entire concentration range. No measurements were obtained for concentrations of 200 µg µl^{-1}. It was only when concentrations reached 12.5 µg/µl or lower that results were obtained for all specimens.

Good measurements were obtained from small tissue samples, when these were completely submersed in the HCCA solution within a 1.5 ml microcentrifuge tube (Fig. 1B). Before samples/matrix ratio was too low to detect a signal, an increase in baseline height in the lower masses was recorded (Fig. 1C). When sample to matrix ratio was increased, an increase in noise was observed (Fig. 1D). Quality improvement of spectra from high tissue-to-HCCA matrix ratios was achieved by dilution. This was tested using tissue from the crustacean *Cancer pagurus*(sample MT01453). Concentrations were diluted from an initial concentration of 200 µg µl^{-1} that resulted in no mass spectra at all. The measurements from diluted preparations then showed similar results as measurements made with the respective concentrations from undiluted sample preparations (compare Fig. 1A brown and red results).

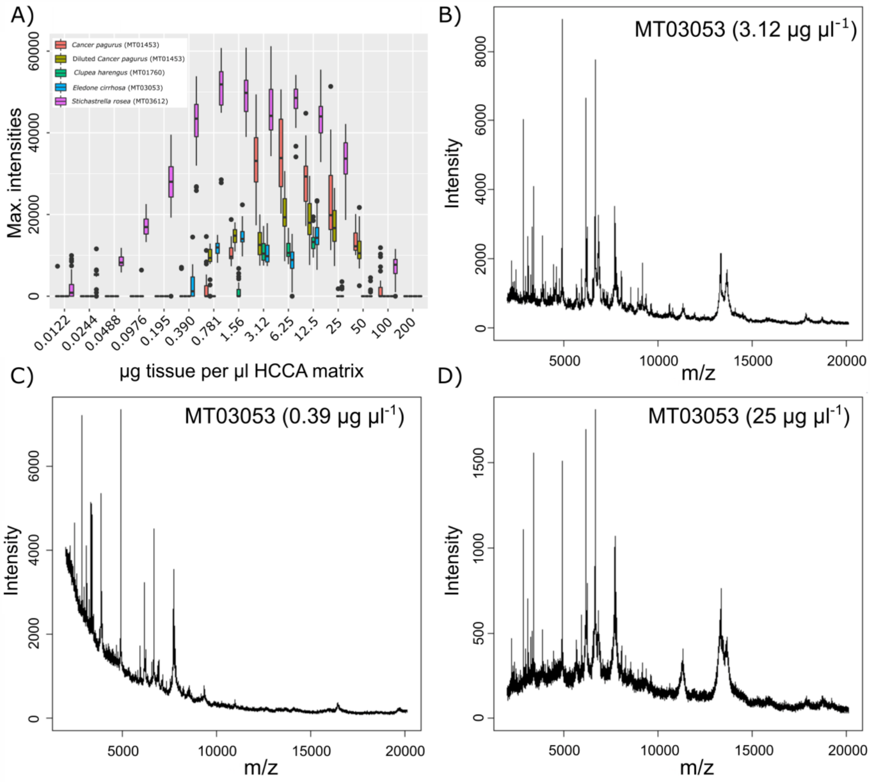

Fig. 1: Results of the sample preparation test. All graphs show peak intensity on the y-axis. In A, x-axis represents tissue:matrix ratio in µg per µl. In B C and D m/z values (ratio of molecule mass and loading) are depicted on the x-axis. A) Maximum intensities as a measure of quality for the different sample to HCCA matrix ratios assessed for four species. Additionally, for *Cancer pagurus* a dilution series (brown) was carried out. B) Good quality spectrum at a ratio of 3.12 µg tissue per µl matrix. C) Lower quality spectrum at 0.39 µg µl^{-1} showing a high baseline. D) Lower quality spectrum at 25 µg µl^{-1} showing stronger noise.

Optimize Random Forest model for classification

For application of RF as a method for classification, we evaluated how strongly the number of specimens per species influences model error. A repeated (n=100) random sampling of two to eleven specimens for species with at least 11 specimens in the data set (n=20) was carried out. This data was then used to create RF models and the OOB error was assessed as a quality criterion. Increasing the number of specimens per species resulted in a decrease of OOB error (Fig. 2). With only two specimens per species the OOB error ranges from 0 to 0.375 with a mean error of 0.18 (SD = 0.073). With eleven specimens per species, the error ranges from 0.005 to 0.036 with a mean error of 0.019 (SD = 0.008). The decrease in OOB error goes nearly into saturation for n >10. For further analyses, we chose n = 6 because the results show a strong decrease in OOB-error variability and a strong decrease in maximum OOB error at this point.

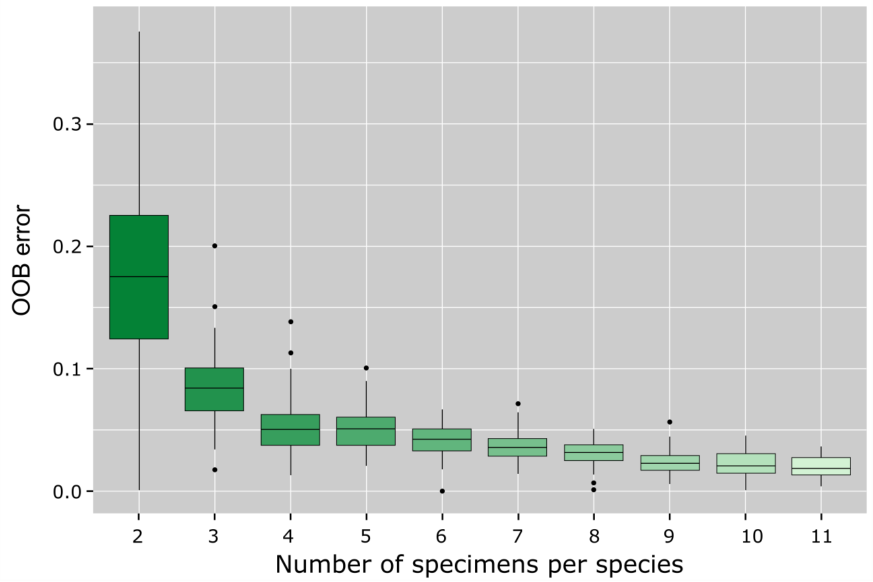

Fig. 2: Random forest model quality assessed with increasing number of specimens per species. For each number of specimens, 100 data sets were created by random sampling. The OOB error (y-axis) decreases with increasing number of specimens (x-axis) and starts going into saturation. Thus, around 10 specimens per species are generally recommended to obtain a high quality model.

Standardization of data processing

Different steps throughout data processing can have a severe impact on classification results. The effect of changing the different data processing steps was evaluated using the RF OOB error as an indicator. For each data set a RF model was trained and the OOB error recorded (supplementary figure 1). Whereas alteration of baseline subtraction iterations generally only had little impact on RF OOB error, changing HWS and SNR had greater effects (supplementary figure 1). The GAM shows that the OOB error is significantly influenced by alteration of the HWS (Table 1, p-value: 0.007) and SNR (Table 1, p-value: 0.001). A combination of 22 baseline estimation iterations, HWS of 7 and SNR of 3 resulted in the lowest OOB error of 0.032. These settings were used for further analyses.

Table. 1: Results of the GAM analyses to detect the most important variable for data processing optimization.

	Estimate	**Std.error**	**z value**	**P-value**
Intercept	-2.49	0.03	-72.72	$2e^{-16}$
	edf	**Ref.df**	**chi.sq**	**P-value**
Baseline iterations	1	1.001	0.125	0.72
Peak detection HWS	1	1.00	7.39	0.007
SNR	1	1.00	10.2	0.001
	R-sq.(adj) = 0.686	R-sq.(adj) = 0.686	Deviance explained = 68.7%	Deviance explained = 6

Classification success

Finally, we tested the identification success based on MALDI-TOF MS data for each specimen in the data set by excluding the respective specimen and using the remaining reference data to identify it.

Overall, 93% of the specimens (n=775) were identified correctly and 86% (n=721) were accepted as correctly classified by the*post-hoc* test (Fig. 3A). Identification for specimens of the classes Ascidacea, Teleostei, Elasmobranchii, Echinoidea, Ophiuroidea, Asteroidea, Bivalvia and Gastropoda resulted in success rates of more than 90%. For classes Cephalopoda and Thecostraca the identification success was still above 85%. Success rates lower than 80% were not recorded. Of the 61 misclassified specimens, 15 were assigned to the false species and recorded as correct identifications by the*post-hoc* test. Of all misclassified specimens, two were assigned to congeneric classes and rated as true positives by the *post-hoc*test, meaning these would have been misclassified and remain unrecognized.

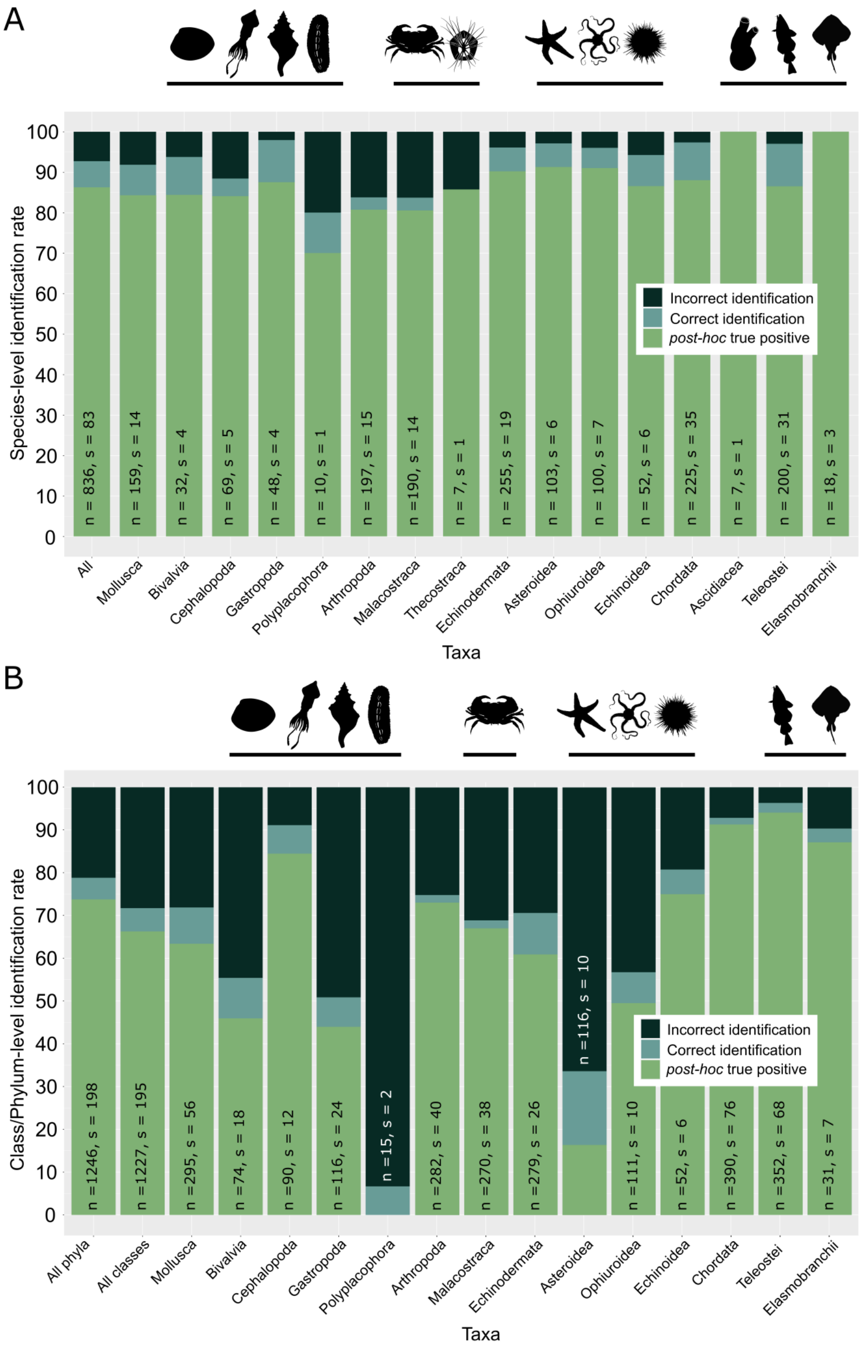

Fig. 3A : Identification rate on species level displayed for all included phyla and classes separately. **3B** : Results of RF specimen identification to phylum and class level. Bars are divided into three categories relevant to the identification success. The darkest color displays the fraction of incorrect identifications, the intermediate color displays correct random forest identifications and the light color represents the percentage of specimens recognized as correct identifications by the *post-hoc* test.

Case study - cryptic species

In the present data, the identification of the starfish*Astropecten irregularis* (Pennant, 1777) specimens from the North Sea serves as an example for closely related species that are still distinguishable by proteomic fingerprinting. In a previous study, this morphotype was found to consist of two major genetic clades with inter-clade distances in COI of up to 12%. Morphological differences were not determined so far. Both groups show different distribution patterns with overlaps (Laakmann et al., 2016). Our data included specimens of both clades, *A. irregularis* 1 (n=8) and *A. irregularis* 2 (n=27).

Data processing settings were optimized for the sub-set of data (HWS = 9 and SNR = 8). Within a RF model produced from the data, a clear distinction between the two genetic groups was possible. None of the specimens was misassigned to the respective other group. This RF model was also used to find the most important variables for differentiation of the two groups using the Gini index, which shows the degree of dissimilarity of the respective variables (Han et al., 2016). The 30 most important variables are given in Fig. 4A. Whereas all peaks can be found in specimens of both groups, the intensities differ strongly allowing a clear differentiation of *A. irregularis* clades using proteome fingerprinting.

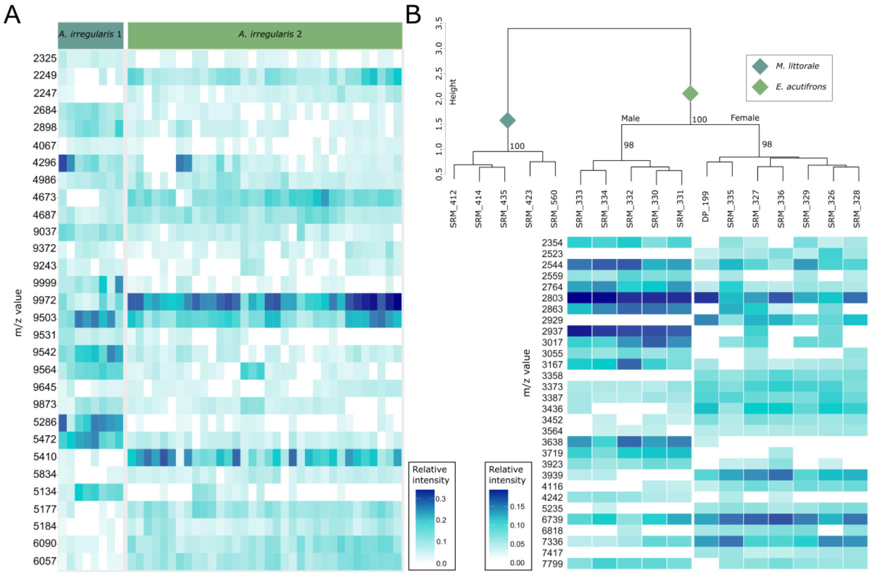

Fig. 4: **A** The 30 most important peaks for differentiation of the starfish *A. irregularis* groups within the random forest model. Species according to COI delimitation are given on top. Molecule masses sorted by size are given on the left hand side. **B**Hierarchical clustering depicts differentiation of the copepod *E. acutifrons* specimens on sex level. Nodal bootstrap support is displayed at the nodes of the tree. The heatmap below the clustering results depicts the 30 most important mass peaks for sex-differentiation using a random forest model with color-coded peak intensities. Data from the marine copepod *Microarthridion littorale* (Poppe, 1881) from the same study was used here as an outgroup species. Relative intensities are color coded.

Case study - sex determination

In previous research it was shown that sex determination may be possible in some species by analyzing the proteomic fingerprint (Rossel and Martínez Arbizu, 2019), however the data was not analyzed any further therein. In depth analyses emphasize these findings and show sex-specific protein patterns in the crustacean copepod *Euterpina acutifrons* (Fig. 4B). Mass peaks such as m/z 2523, 2929 and 7417 are female specific and not found in any of the male specimens. Others however, predominantly occur in male specimens (m/z 3638, 3719). Further mass peaks are evenly observed in measurements from both sexes but show intensity-pattern differences.

Phyla and class models for identification

If a species is not part of a reference library, it may be desirable to obtain a higher level classification. To test if this is possible based on MALDI-TOF mass spectra of metazoans, species were systematically taken out of the RF training data set and classified with a RF model that was trained on higher taxonomic level but does not include any information on the respective species to be classified. Regarding all phyla together, a classification success of 81% (77% true positive rate (tpr)) was achieved with phyla-wise success rates ranging from 73% (64 % tpr) in Echinodermata to 95% (92% tpr) in Chordata (Fig. 3B). On class level the combined success rate was 72% (66% tpr) ranging from 7% (0% tpr) in Polyplacophora, for which only two species were included in the data set, to 96% (94% tpr) in Teleostei.

For 31 taxa (n = 324), a congeneric species was included. Thus, it was tested if species have a higher affinity to be classified as a congeneric species in case the respective species is removed from the training data. Of these 31 taxa, 30% of specimens were classified as a congeneric species.

Discussion

The aims of this study were (1) to evaluate the wide applicability of proteomic fingerprinting for species identification in marine science across different metazoan phyla and classes, (2) to identify critical steps in sample preparation and data processing, and (3) to contribute to the development of standard procedures and best practices for MALDI-TOF MS based metazoan classification. The general applicability to metazoans has been proven before (Mazzeo et al., 2008; Dieme et al., 2014; Yssouf et al., 2014; Flaudrops et al., 2015; Mazzeo and Siciliano, 2016; Maász et al., 2017; Rossel and Martínez Arbizu, 2019; Rossel et al., 2020a). However, here we show for the first time the applicability of this method to a large taxonomic range using a comprehensive data set with an overall species identification success rate of 93%.

Similar high identification success rates on species level were observed for individual metazoan groups (Hynek et al., 2018; Vega-Rúa et al., 2018; Holst et al., 2019; Loaiza et al., 2019; Rakotonirina et al., 2020; Rossel et al., 2020a). Additionally, our results show that specimens absent from the reference library will be assigned to the correct phyla or class with a high probability implying some kind of phylogenetic signal on higher taxonomic level as was already reported for congeneric *Drosophila* before (Feltens et al., 2010). Testing if species would be classified as a congeneric species in the absence of the actual species was less promising in our study with only 30% of specimens being assigned to a congeneric species. This complies with other studies that only show occasional similarity of congeneric species e.g. in cluster analyses but without consistency across all congeneric species (Laakmann et al., 2013; Chavy et al., 2019; Rossel and Martínez Arbizu, 2019).

In closely related species, morphological identification can often be complicated. Using proteomic fingerprinting, these problems can however be resolved as indicated by the analysis of the *A. irregularis*complex. Even though mass spectra show high similarities, distinct patterns of peak presence and absence as well as pronounced differences in relative peak intensities serve as good markers for species identification. Beyond mere species identification, the example of*E. acutifrons* shows the power of the method to differentiate specimens even on a sex level. This has been shown before for e.g. the fish species *Alburnus alburnus* (Linnaeus, 1758) (Maász et al., 2017). Whereas authors focused on presence and absence of peaks, we were able to show that also relative intensities of certain mass peaks play an important role in differentiation of sexes. Prior studies on larger planktonic copepods have also shown a great potential for differentiation of developmental stages based on a proteomic fingerprint (Rossel et al., 2022).

Finally, we have shown the necessity of comprehensive reference libraries. Low numbers of specimens per species in reference libraries fail to provide sufficient information on species specific mass spectra features and intraspecific variability. Only with around nine to ten reference specimens per species, the identification error stabilizes on a constantly low level. This supports findings by Rakotonirina et al (2020) who found an increase of identification score with increasing numbers of available main spectrum patterns. In general we would recommend to use more than three specimens per species and preferably to include around ten specimens for every species in a reference library.

MALDI-TOF MS can be used as a universal method for species identification of metazoan species. Due to the short preparation time, low costs (Tran et al., 2015; Rossel et al., 2019) and high identification success it can be a valuable tool in biodiversity assessments replacing time-intense morphological identification or costly DNA barcoding. Especially in cases of closely related or very similar species it can foster a rapid identification. The applicability of proteome fingerprinting for the differentiation of cryptic species was already shown and even in cases of morphologically very similar species, still differences were found (Müller et al., 2013; Paulus et al., 2022).

Tissue samples used in this work were obtained from specimens stored between seven to 12 years under partly unknown storage conditions. We assume working with fresh or recently fixed material would have resulted in even higher identification success rates. This is supported by the high mass spectra quality obtained from fish species, which were extracted and put into freezer storage almost immediately after sampling (personal communication Knebelsberger). The adverse effect of fixation and storage on resulting mass spectra quality in metazoans was investigated several times and supports this assumption (Rossel and Martínez Arbizu, 2018b; Rakotonirina et al., 2020). We received good results for storage at -20°C and also for long-term storage at -80°C, thus we recommend cold storage of samples at -20°C, until further systematic analyses will specify threshold temperatures for short- (months) or long-term (years) storage.

Our tests have shown that sample concentration is pivotal to obtain good quality mass spectra. While too low sample/matrix ratios will result in lower intensities and a higher baseline, too much tissue will increase the noise in the data and result in unsuccessful measurements. For all investigated taxa, the same sample preparation method was used; however attention must be paid to the correct ratio of matrix and compound to be analyzed. This allows the wide application of this method without adaptation of the protocol to a certain species as it would be necessary for methods such as COI barcoding where certain groups would need highly specific sets of amplification primers (Lohman et al., 2009; Toumi et al., 2013) and adjustment of PCR settings.

Much effort is put into optimizing mass spectra quality by adjusting different preparation protocols (Jeverica et al., 2018; Wang et al., 2021) or developing methods for steps such as baseline correction, smoothing or peak picking (Ressom et al., 2007; Shin et al., 2010). Methods are adjusted either to increase classification success or to obtain better mass spectra reproducibility. Here, we tested the influence of certain steps during data processing on classification success focusing on the important steps for peak detection. Whereas baseline subtraction and adjustment of a SNR value both aim at reducing noise within the data, adjusting the HWS influences the peak picking resolution. Thus, by decreasing the HWS during peak detection, the number of peaks will increase as the highest peak within the HWS will be the detected. This will result in peaks of very similar size being recognized as distinct peaks, rather than being put together in a single bin. This does also explain the high effect of both parameters SNR and HWS compared to baseline subtraction. Baseline subtraction is constrained towards reducing instrument-dependent noise. Adjustment of the SNR value will however, like HWS alteration, affect the number of more dominant peaks and thus the general resolution of the mass spectra. Hence, more species-specific information is retained and more information is available for classification. Based on our results, rather than testing all variables, adjusting SNR and HWS should be adequate to optimize the data pipeline. However, it needs to be emphasized that this pipeline aims at optimizing species identification and may not be adequate for investigation of intraspecific variability as was shown elsewhere[16].

In summary, we propose a workflow applicable for any metazoan species or tissue sample to be identified: A comprehensive reference library is needed with species level identification by morphological or molecular approaches (Fig. 5a). In the lab, a small tissue (up to 1mm^3) is retrieved and incubated for at least 5 minutes in the HCCA-matrix solution. Of the resulting extract, 1 to 1.5 µl are transferred to a target plate for measurement. Data processing is carried out in R (Fig 5b). Mass spectra quality is done by eye and supported by R-packages such as MALDIrppa (Palarea-Albaladejo et al., 2017). Finally, based on previously assessed species identification, data processing can be optimized to obtain ideal settings for classification. Depending on our results this can be narrowed to adjustment of HWS- and SNR-value. Based on the reference library, a RF model can be calculated for specimen identification (Fig. 5c). Applying a *post-hoc* test will provide further support for the identification. If classification is not well supported, a RF model on class or phyla level can be applied to obtain higher-level classification.

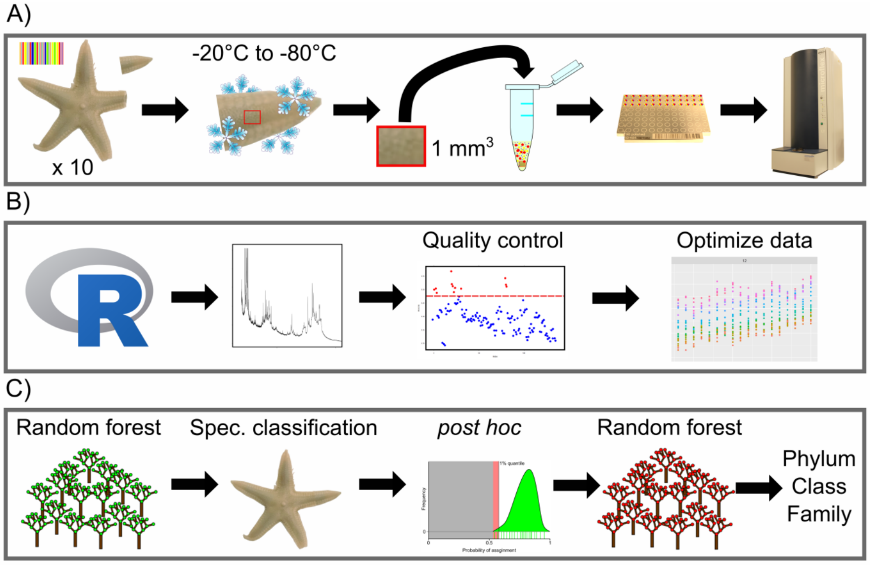

Fig. 5: Proposed workflow from specimen to classification. A) Sampling and sample preparation. A tissue sample from an identified specimen is stored at cold temperatures until further processing. Especially for long-term storage it should be aimed at low temperature storage. A smaller subsample is incubated in HCCA Matrix and the solution is transferred to target plate for measurement. B) Measurements are imported to R and mass spectra are quality controlled. The data then needs optimization for best model quality (e.g. adjust HWS, SNR, etc.). C) Classification using the optimized RF model. Application of the*post-hoc* test to verify classification. If classification is rejected, phylum/class/family level models may result in a higher taxonomic classification. The R logo is © 2016 The R Foundation and is used under the CC-BY-SA 4.0 license.

Conclusion

MALDI-TOF MS was proven an easy to apply, cost-effective and time-saving tool for identification across taxa. It is especially feasible in applications where mere species identification is desired, for example in biodiversity assessments. By the standardized workflow, a wide range of marine metazoan specimens can be identified quantitatively and effectively on species level thereby bypassing some of the high requirements associated with genetic methods, such as access to special laboratories, searching for primers etc. We want to highlight here that proteomic fingerprinting will be due to its simplicity, reliability and efficiency a valuable supplement to the molecular toolbox for taxonomy.

Acknowledgement

We thank Michael Raupach, Thomas Knebelsberger, Andrea Barco and their collaborators, students and technicians for bringing the tissue database into life and good documentation of their work as well as making their work publicly available. Respectively, this work was supported by the following grants: The Federal Ministry of Education and Research (Grant No. 03F0499A) and the Land Niedersachsen. This is publication 14 of the Senckenberg am Meer Proteomics Laboratory. This study was supported by the DFG initiative 1991 "Taxono-omics" (grant number RE2808/3-1/2). 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).

Author contributions

S.R, J.P, S.L and P.M.A conceived the study. H.N. carried out the majority of morphological species identifications. S.R., N. C. and A. E. carried out MALDI-TOF MS measurements. S.R. and J.P. analyzed the data and wrote a first manuscript draft. All authors significantly participated in critical revision of the manuscript draft.

Data accessibility

All mass spectra data will be made available at Data Dryad (DOI:XXXX) upon manuscript acceptance. Relevant R-Scripts will be stored alongside raw data.

Benefit-Sharing Statement

No data from foreign countries were acquired and thus, no benefit sharing as demanded by the Nagoya protocol is carried out.

Competing interests

The authors declare to not have any competing interests.

References

Bailey, R. C., Norris, R. H., and Reynoldson, T. B. (2001). Taxonomic resolution of benthic macroinvertebrate communities in bioassessments.*J. North Am. Benthol. Soc.* 20, 280–286.

Barco, A., Raupach, M. J., Laakmann, S., Neumann, H., and Knebelsberger, T. (2016). Identification of North Sea molluscs with DNA barcoding.*Mol. Ecol. Resour.* 16, 288–297.

Breimann, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32.

Calderaro, A., Arcangeletti, M.-C., Rodighiero, I., Buttrini, M., Gorrini, C., Motta, F., et al. (2014). Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry applied to virus identification. *Sci. Rep.* 4, 6803. doi: 10.1038/srep06803.

Chavy, A., Nabet, C., Normand, A. C., Kocher, A., Ginouves, M., Prévot, G., et al. (2019). Identification of French Guiana sand flies using MALDI-TOF mass spectrometry with a new mass spectra library. *PLoS Negl. Trop. Dis.* 13, e0007031. doi: 10.1371/journal.pntd.0007031.

Dieme, C., Yssouf, A., Vega-Rúa, A., Berenger, J.-M., Failloux, A.-B., Raoult, D., et al. (2014). Accurate identification of Culicidae at aquatic developmental stages by MALDI-TOF MS profiling. *Parasit. Vectors* 7, 544.

Feltens, R., Görner, R., Kalkhof, S., Gröger-Arndt, H., and 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 Evol. Biol.* 10, 1.

Fenselau, C., and Demirev, P. A. (2001). Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrom. Rev.*20, 157–171.

Flaudrops, C., Armstrong, N., Raoult, D., and Chabrière, E. (2015). Determination of the animal origin of meat and gelatin by MALDI-TOF-MS.*J. Food Compos. Anal.* 41, 104–112.

Gebhardt, K., and Knebelsberger, T. (2015). Identification of cephalopod species from the North and Baltic Seas using morphology, COI and 18S rDNA sequences. *Helgol. Mar. Res.* 69, 259.

Gibb, S. (2015). MALDIquantForeign: Import/Export routines for MALDIquant. A package for R. *HttpsCRANR-Proj.*

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

Han, H., Guo, X., and Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. in*2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (IEEE), 219–224.

Holst, S., Heins, A., and Laakmann, S. (2019). Morphological and molecular diagnostic species characters of Staurozoa (Cnidaria) collected on the coast of Helgoland (German Bight, North Sea).*Mar. Biodivers.* doi: 10.1007/s12526-019-00943-1.

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

Jeverica, S., Nagy, E., Mueller-Premru, M., and Papst, L. (2018). Sample preparation method influences direct identification of anaerobic bacteria from positive blood culture bottles using MALDI-TOF MS.*Anaerobe* 54, 231–235.

Kaiser, P., Bode, M., Cornils, A., Hagen, W., Martínez Arbizu, P., Auel, H., et al. (2018). High-resolution community analysis of deep-sea copepods using MALDI-TOF protein fingerprinting. *Deep-Sea Res. Part Oceanogr. Res. Pap.* 138, 122–130.

Knebelsberger, T., Landi, M., Neumann, H., Kloppmann, M., Sell, A. F., Campbell, P. D., et al. (2014). A reliable DNA barcode reference library for the identification of the North European shelf fish fauna.*Mol. Ecol. Resour.* 14, 1060–1071.

Knebelsberger, T., and Thiel, R. (2014). Identification of gobies (Teleostei: Perciformes: Gobiidae) from the North and Baltic Seas combining morphological analysis and DNA barcoding. *Zool. J. Linn. Soc.* 172, 831–845.

Korfhage, S. A., Rossel, S., Brix, S., McFadden, C. S., Ólafsdóttir, S. H., and Martínez Arbizu, P. (2022). Species Delimitation of Hexacorallia and Octocorallia Around Iceland Using Nuclear and Mitochondrial DNA and Proteome Fingerprinting. *Front. Mar. Sci.* 9. Available at: https://www.frontiersin.org/article/10.3389/fmars.2022.838201.

Kürzel, K., Kaiser, S., Lörz, A.-N., Rossel, S., Paulus, E., Peters, J., et al. (2022). Correct Species Identification and Its Implications for Conservation Using Haploniscidae (Crustacea, Isopoda) in Icelandic Waters as a Proxy. *Front. Mar. Sci.* 8. doi: doi: 10.3389/fmars.2021.795196.

Laakmann, S., Boos, K., Knebelsberger, T., Raupach, M. J., and Neumann, H. (2016). Species identification of echinoderms from the North Sea by combining morphology and molecular data. *Helgol. Mar. Res.* 70.

Laakmann, S., Gerdts, G., Erler, R., Knebelsberger, T., Martínez Arbizu, P., and Raupach, M. J. (2013). Comparison of molecular species identification for North Sea calanoid copepods (Crustacea) using proteome fingerprints and DNA sequences. *Mol. Ecol. Resour.* 13, 862–76. doi: 10.1111/1755-0998.12139.

Loaiza, J. R., Almanza, A., Rojas, J. C., Mejia, L., Cervantes, N. D., Sanchez-Galan, J. E., et al. (2019). Application of matrix-assisted laser desorption/ionization mass spectrometry to identify species of Neotropical Anopheles vectors of malaria. *Malar. J.* 18, 95.

Lohman, D. J., Prawiradilaga, D. M., and Meier, R. (2009). Improved COI barcoding primers for Southeast Asian perching birds (Aves: Passeriformes). *Mol. Ecol. Resour.* 9, 37–40.

Maász, G., Takács, P., Boda, P., Várbiró, G., and Pirger, Z. (2017). Mayfly and fish species identification and sex determination in bleak (Alburnus alburnus) by MALDI-TOF mass spectrometry. *Sci. Total Environ.* 601, 317–325.

Markert, A., Raupach, M. J., Segelken-Voigt, A., and Wehrmann, A. (2014). Molecular identification and morphological characteristics of native and invasive Asian brush-clawed crabs (Crustacea: Brachyura) from Japanese and German coasts: Hemigrapsus penicillatus (De Haan, 1835) versus Hemigrapsus takanoi Asakura & Watanabe 2005. *Org. Divers. Evol.* 14, 369–382.

Martínez Arbizu, P., and Rossel, S. (2018). RFtools: Miscellaneous Tools For Random Forest Models. doi: https://zenodo.org/record/118843.

Mazzeo, M. F., Giulio, B. D., Guerriero, G., Ciarcia, G., Malorni, A., Russo, G. L., et al. (2008). Fish authentication by MALDI-TOF mass spectrometry. *J. Agric. Food Chem.* 56, 11071–11076.

Mazzeo, M. F., and Siciliano, R. A. (2016). Proteomics for the authentication of fish species. *J. Proteomics* 147, 119–124.

Müller, P., Pflüger, V., Wittwer, M., Ziegler, D., Chandre, F., Simard, F., et al. (2013). Identification of cryptic Anopheles mosquito species by molecular protein profiling. *PLOS ONE* 8, e57486. doi: 10.1371/journal.pone.0057486.

Nabet, C., Kone, A. K., Dia, A. K., Sylla, M., Gautier, M., Yattara, M., et al. (2021). New assessment of Anopheles vector species identification using MALDI-TOF MS. *Malar. J.* 20, 1–16.

Palarea-Albaladejo, J., Mclean, K., Wright, F., and Smith, D. G. (2017). MALDIrppa: quality control and robust analysis for mass spectrometry data. *Bioinformatics* 34, 522–523.

Park, N., Yeom, J., Jeong, R., and Lee, W. (2021). Novel attempt at discrimination of a bullet-shaped siphonophore (Family Diphyidae) using matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-ToF MS). *Sci. Rep.* 11, 19077. doi: 10.1038/s41598-021-98724-z.

Paulus, E., Brix, S., Siebert, A., Martínez Arbizu, P., Rossel, S., Peters, J., et al. (2022). Recent speciation and hybridization in Icelandic deep-sea isopods: An integrative approach using genomics and proteomics. *Mol. Ecol.* 31, 313–330. doi: 10.1111/mec.16234.

Peters, J., Laakmann, S., Rossel, S., Martínez Arbizu, P., and Renz, J. (2023). Perspectives of species identification by MALDI-TOF MS in monitoring-Stability of proteomic fingerprints in marine epipelagic copepods. *Mol. Ecol. Resour.* doi: 10.1111/1755-0998.13779.

Rakotonirina, A., Pol, M., Kainiu, M., Barsac, E., Tutagata, J., Kilama, S., et al. (2020). MALDI-TOF MS: optimization for future uses in entomological surveillance and identification of mosquitoes from New Caledonia. *Parasit. Vectors* 13, 1–12.

Rakotonirina, A., Pol, M., Raharimalala, F. N., Ballan, V., Kainiu, M., Boyer, S., et al. (2022). MALDI-TOF MS: An effective tool for a global surveillance of dengue vector species. *PLOS ONE* 17, e0276488. doi: 10.1371/journal.pone.0276488.

Raupach, M. J., Barco, A., Steinke, D., Beermann, J., Laakmann, S., Mohrbeck, I., et al. (2015). The application of DNA barcodes for the identification of marine crustaceans from the North Sea and adjacent regions. *PLOS ONE* 10, e0139421.

R-Core-Team (2022). R: A language and environment for statistical computing. Available at: https://www.R-project.org/.

Renz, J., Markhaseva, E. L., Laakmann, S., Rossel, S., Martínez Arbizu, P., and Peters, J. (2021). Proteomic fingerprinting facilitates biodiversity assessments in understudied ecosystems: A case study on integrated taxonomy of deep sea copepods. *Mol. Ecol. Resour.*

Ressom, H. W., Varghese, R. S., Drake, S. K., Hortin, G. L., Abdel-Hamid, M., Loffredo, C. A., et al. (2007). Peak selection from MALDI-TOF mass spectra using ant colony optimization.*Bioinformatics* 23, 619–626.

Rossel, S., Barco, A., Kloppmann, M., Martínez Arbizu, P., Huwer, B., and Knebelsberger, T. (2020a). Rapid species level identification of fish eggs by proteome fingerprinting using MALDI-TOF MS. *J. Proteomics* , 103993.

Rossel, S., Deli, T., and Raupach, M. J. (2020b). First insights into the phylogeography and demographic history of the common hermit crab Pagurus bernhardus (Linnaeus, 1758)(Decapoda: Anomura: Paguridae) across the Eastern Atlantic and North Sea. *J. Crustac. Biol.* 40, 435–449.

Rossel, S., Kaiser, P., Bode-Dalby, M., Renz, J., Laakmann, S., Auel, H., et al. (2022). Proteomic fingerprinting enables quantitative biodiversity assessments of species and ontogenetic stages in Calanus congeners (Copepoda, Crustacea) from the Arctic Ocean. *Mol. Ecol. Resour.* n/a. doi: 10.1111/1755-0998.13714.

Rossel, S., Khodami, S., and Martínez Arbizu, P. (2019). Comparison of rapid biodiversity assessment of meiobenthos using MALDI-TOF MS and Metabarcoding. *Front. Mar. Sci.* 6, 659. doi: 10.3389/fmars.2019.00659.

Rossel, S., and Martínez Arbizu, P. (2018a). 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 Ecol. Evol.* 9, 1421–1434. doi: 10.1111/2041-210X.13000.

Rossel, S., and Martínez Arbizu, P. (2018b). Effects of Sample Fixation on Specimen Identification in Biodiversity Assemblies based on Proteomic Data (MALDI-TOF). *Front. Mar. Sci.* 5, 149. doi: 10.3389/fmars.2018.00149.

Rossel, S., and Martínez Arbizu, P. (2019). Revealing higher than expected diversity of Harpacticoida (Crustacea: Copepoda) in the North Sea using MALDI-TOF MS and molecular barcoding. *Sci. Rep.* 9, 9182. doi: 10.1038/s41598-019-45718-7.

Ryan, C., Clayton, E., Griffin, W., Sie, S., and Cousens, D. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* 34, 396–402.

Sandrin, T. R., Goldstein, J. E., and Schumaker, S. (2013). MALDI TOF MS profiling of bacteria at the strain level: a review. *Mass Spectrom. Rev.* 32, 188–217.

Sassi, M., Arena, S., and Scaloni, A. (2015). MALDI-TOF-MS platform for integrated proteomic and peptidomic profiling of milk samples allows rapid detection of food adulterations. *J. Agric. Food Chem.* 63, 6157–6171.

Savitzky, A., and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639.

Shin, H., Sampat, M. P., Koomen, J. M., and Markey, M. K. (2010). Wavelet-based adaptive denoising and baseline correction for MALDI TOF MS. *Omics J. Integr. Biol.* 14, 283–295.

Singhal, N., Kumar, M., Kanaujia, P. K., and Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.* 6. doi: 10.3389/fmicb.2015.00791.

Timms, L. L., Bowden, J. J., Summerville, K. S., and Buddle, C. M. (2013). Does species-level resolution matter? Taxonomic sufficiency in terrestrial arthropod biodiversity studies. *Insect Conserv. Divers.* 6, 453–462.

Toumi, F., Waeyenberge, L., Viaene, N., Dababat, A., Nicol, J. M., Ogbonnaya, F., et al. (2013). Development of two species-specific primer sets to detect the cereal cyst nematodes Heterodera avenae and Heterodera filipjevi. *Eur. J. Plant Pathol.* 136, 613–624.

Tran, A., Alby, K., Kerr, A., Jones, M., and Gilligan, P. H. (2015). Cost Savings Realized by Implementation of Routine Microbiological Identification by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. *J. Clin. Microbiol.* 53, 2473–2479. doi: 10.1128/JCM.00833-15.

Vega-Rúa, A., Pagès, N., Fontaine, A., Nuccio, C., Hery, L., Goindin, D., et al. (2018). Improvement of mosquito identification by MALDI-TOF MS biotyping using protein signatures from two body parts.*Parasit. Vectors* 11, 574.

Volta, P., Riccardi, N., Lauceri, R., and 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. *J. Limnol.* 71, e17.

Wang, J., Wang, H., Cai, K., Yu, P., Liu, Y., Zhao, G., et al. (2021). Evaluation of three sample preparation methods for the identification of clinical strains by using two MALDI-TOF MS systems. *J. Mass Spectrom.* 56, e4696.

Wickham, H., Francois, R., Henry, L., and Müller, K. (2022). dplyr: A Grammar of Data Manipulation. Available at: https://CRAN.R-project.org/package=dplyr.

Wilke, T., Renz, J., Hauffe, T., Delicado, D., and Peters, J. (2020). Proteomic Fingerprinting Discriminates Cryptic Gastropod Species.*Malacologia* 63, 131–137.

Yeom, J., Park, N., Jeong, R., and Lee, W. (2021). Integrative Description of Cryptic Tigriopus Species From Korea Using MALDI-TOF MS and DNA Barcoding. *Front. Mar. Sci.* 8, 495.

Yssouf, A., Flaudrops, C., Drali, R., Kernif, T., Socolovschi, C., Berenger, J.-M., et al. (2013). Matrix-assisted laser desorption ionization–time of flight mass spectrometry for rapid identification of tick vectors. *J. Clin. Microbiol.* 51, 522–528.

Yssouf, A., Socolovschi, C., Leulmi, H., Kernif, T., Bitam, I., Audoly, G., et al. (2014). Identification of flea species using MALDI-TOF/MS.*Comp. Immunol. Microbiol. Infect. Dis.* 37, 153–157.

Supplementary Fig. 1: Out of box error of the random forest model (OOB error on y-axis) for the different combinations of baseline subtraction iterations, signal to noise ratio (SNR) and half window size (HWS) during peak picking. Each box represents the number of baseline iteration steps ranging from 5 to 30. The x-axis displays the SNR value ranging from 3 to 20. Colors indicate different HWS ranging from 5 to 30. The results are shown for all 12,186 variable combinations.