LIMNOLOGY
and
OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 18, 2020, 183–195

doi: 10.1002/lom3.10358

Unsupervised biodiversity estimation using proteomic fingerprints from MALDI-TOF MS data[a]

Sven Rossel[1], *** Pedro Martínez Arbizu**[1,2]

[1]Senckenberg am Meer, German Centre for Marine Biodiversity Research (DZMB), Wilhelmshaven, Germany
[2]AG Marine Biodiversität, FKV-IBU, Universität Oldenburg, Oldenburg, Germany

Abstract

Species identification using matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) data strongly relies on reference libraries to differentiate species. Because comprehensive reference libraries, especially for metazoans, are rare, we explored the accuracy of unsupervised diversity estimations of communities using MALDI-TOF MS data in the absence of reference libraries to provide a method for future application in ecological research. To discover the best analysis strategy providing high congruence with true community structures, we carried out a simulation with more than 30,000 analyses using different combinations of data transformations, dimensionality reductions, and cluster algorithms. Species profile, Hellinger, and presence/absence transformations were applied to raw data and dimensions were reduced using principal component analysis (PCA), t-distributed stochastic neighbor embedding, and uniform manifold approximation and projection. To estimate biodiversity, data were clustered making use of partitioning around medoids, model-based clustering, and K-means clustering. The analyses were carried out on published mass spectrometry data of harpacticoid copepods. Most successful combinations (Hellinger transformation + PCA or raw data + partitioning around medoids) returned good values even for difficult species distributions containing numerous singleton species. Nevertheless, errors occurred most frequently because of such singleton taxa. Hence, replicative sampling in wide sampling areas for analysis is emphasized to increase the minimum number of specimens per species, thus reducing putative sources of errors. Our results demonstrate that MALDI-TOF MS data can be used to accurately estimate the biodiversity of unknown communities using unsupervised learning methods. The provided approach allows the biodiversity comparison of sampled regions for which no reference libraries are available. Hence, especially data on groups which demand a time-consuming identification or are highly abundant can be analyzed within short working time, accelerating ecological studies.

Matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a tool widely used in microbiology to identify bacteria, viruses, or fungi (Majchrzykiewicz-Koehorst et al. 2015; Singhal et al. 2015; Normand et al. 2017). It can be used as a rapid, cost-effective alternative to specimen by specimen DNA barcoding (Hebert et al. 2003). That is why recently, mainly in pilot studies, this technique was successfully applied for species identification of metazoans (Dvorak et al. 2014; Yssouf et al. 2014; Mazzeo and Siciliano 2016). To identify specimens based on a proteomic fingerprint, often company supplied supervised identification software solutions such as the MALDI Biotyper by Bruker are used. These find the most similar spectra from a reference library and return a value of certainty for the resulting identification. Similar to this, an open source R-based random forest (Breimann 2001) approach was introduced using machine learning to classify species (Rossel and Martínez Arbizu 2018*a*) according to the available reference library. Applying a post hoc test evaluates if the identification can be considered correct or false positive. Some studies employed techniques such as hierarchical clustering (Kaiser et al. 2018) or principal component analysis (PCA; Hynek et al. 2018) to discriminate species. However, all these techniques rely on reference libraries to assess species diversity and some fail to detect false-positive classifications. Hence, identifying new species in biodiversity assessments is difficult.

*Correspondence: sven.rossel@senckenberg.de

Author Contribution Statement: S.R. carried out the analyses. P.M.A. designed the study and contributed to the writing of the manuscript and gave final approval for publication.

Additional Supporting Information may be found in the online version of this article.

However, assessing biodiversity is crucial in ecological studies to understand the interaction of a community to the surrounding ecosystem but also to infer effects of a changing environment on species' diversity. Quantifying biodiversity is necessary to allow comparisons of different ecosystems and environments. Nevertheless, assessing biodiversity using proteomic fingerprinting in areas for which no MALDI-TOF reference libraries are available is difficult. Supervised tools such as the Bruker MALDI-TOF Biotyper or the random forest approach cannot provide identifications without a library. Hierarchical clustering and PCA on the other hand will fail to provide species margins and thus depend on researchers to recognize these by themselves subjectively. In molecular analyses, tools such as the Automatic Barcode Gap Discovery (ABGD; Puillandre et al. 2012*a*) or the Generalized Mixed Yule Coalescent approach (Pons et al. 2006) are frequently used to automatically delimit species. However, to date there is no comparable tool for MALDI-TOF MS data to delimit species.

By having tested various combinations of data transformations, dimensionality reduction methods, and different clustering algorithms, we provide a workflow to unsupervised biodiversity estimation based on MALDI-TOF MS data without the need for reference libraries. The workflow can easily be applied by using the R function provided in the Supporting Information S1.

Materials and methods

Data set

MALDI-TOF MS data from published mass spectra (Rossel and Martínez Arbizu 2018*b*,*c*, 2019*a*,*b*; Supporting Information Table S1) on harpacticoid copepods were pooled and screened for species with at least 10 specimens. Only specimens reported to be obtained from appropriately stored samples were included and combined into the data set used for this study.

By repetitive sampling from this data set, 660 different data sets were constructed containing between 2 and 34 different species. In total, 20 data sets per n species were generated (33×20). For each species, only 75% of the 10 specimens available per species were included in the simulated data set (Fig. 1) by random sampling.

The data set is a matrix containing specimens as rows and m/z values (molecule mass/charge) as columns. Values in these columns indicate peak intensities after standard workflow consisting of data trimming, smoothing, normalization, noise reduction, peak detection, and peak binning (Bode et al. 2017; Kaiser et al. 2018; Holst et al. 2019). In the following, these data will be referred to as raw data.

Data analyses

To find the best method for unsupervised biodiversity estimation, different combinations of data transformation, dimensionality reduction, and cluster algorithms were tested (Fig. 1). Data were either analyzed with presence/absence (p/a), species profile (total), or Hellinger transformation (Legendre and Gallagher 2001) using "decostand" from the R-package vegan (Oksanen et al. 2013) or without data transformation (raw). Transformed data were tested without dimensionality reduction (none) or after applying PCA (Pearson 1901), uniform manifold approximation and projection (UMAP; McInnes and Healy 2018) from the R-package "umap"

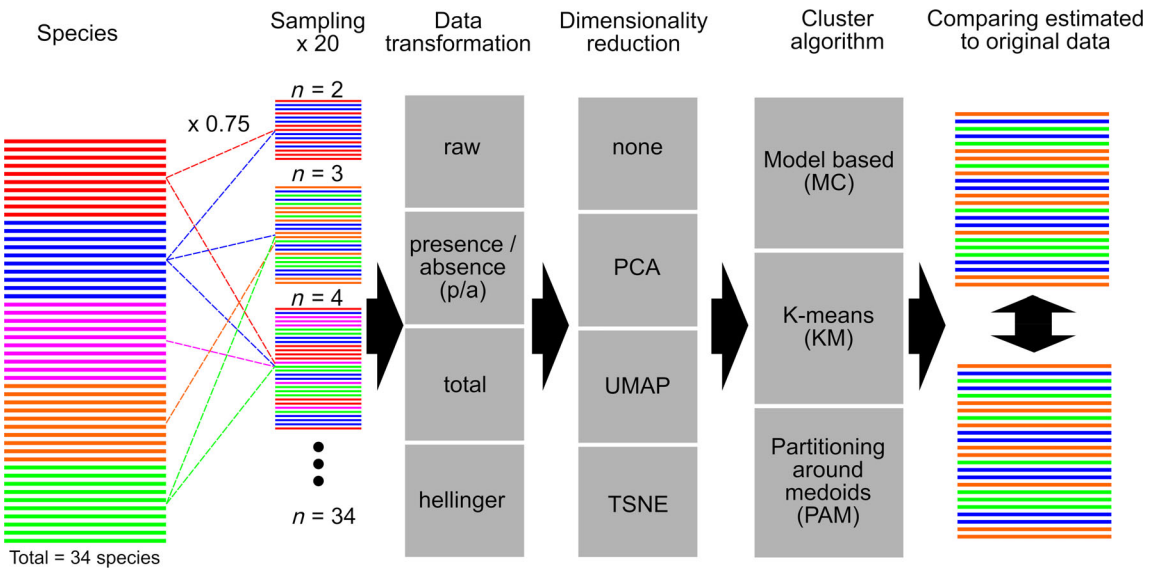

Fig. 1. Workflow for the test of unsupervised diversity estimation. From the initial 34 species, specimens were sampled with a frequency of 0.75 into new data sets containing between two and 34 species. For each of these species numbers, 20 data sets were simulated. Different data transformations, dimensionality reductions, and cluster algorithms were applied to these data sets. Finally, the resulting estimated biodiversity was compared to the correct data.

(Konopka 2018) or t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton 2008) from the R-package "Rtsne" (Krijthe 2015). The resulting data sets were analyzed with model-based clustering (MC; Fraley and Raftery 2002) using the command "Mclust" from the R-package "mclust" (Scrucca et al. 2016), K-means clustering (Lloyd 1982; Fraley and Raftery 2002) using the command "cascadeKM" from the R-package "vegan" or with partitioning around medoids (PAM; Kaufman and Rousseeuw 1990) from the R-package "cluster" (Maechler et al. 2018). While for the majority of methods the most commonly applied R-packages were used, the "Rtsne" package was chosen because it uses a faster t-SNE implementation compared to other t-SNE packages.

When analyzing p/a data in mass spectrometry, weight is put on the presence of a mass peak and not on the intensity thus, giving importance also to less intense peaks. This may particularly be relevant in studies searching for biomarkers to differentiate between closely related species (Carrera et al. 2013). However, unlike raw data, p/a transformation disregards information provided through signal intensities. That is why, additional to raw data, species profile (total) and Hellinger data transformations (Legendre and Gallagher 2001), which were already shown to notably reduce identification errors in machine learning applications (Rossel and Martínez Arbizu 2018*a*), were tested. These data transformations are originally recommended for use in ecological studies because of giving low weights to rare species (Legendre and Gallagher 2001), reducing the effect of many zeros in a data set. In ecological research, many zeroes may occur especially in studies analyzing gradient data comprising sites with different species compositions (Legendre and Gallagher 2001). Because of mass spectra differences between species, MALDI-TOF MS data sets also do often contain numerous zeros. Therefore, these transformations were found suitable for this kind of data. Like with raw data, one advantage of these relative transformations over p/a transformation is the additional information obtained through species-specific peak intensities considered during analyses.

When using species profiles transformation, relative abundances are obtained row wise by division through margin totals. Hellinger transformation is based on these relative abundances and is calculated by applying a further square root transformation to total transformed data (see command decostand from R-package "vegan").

Our workflow for unsupervised biodiversity estimation demands reanalyzing the data set as many times as there are specimens in the data set. Computational effort can be very high because data sets may contain several hundred m/z values (Rossel and Martínez Arbizu 2018*c*, 2019*b*). Therefore, dimensionality reduction is an important step to reduce computational resources needed. However, chosen dimensionality reduction still needs to preserve the actual data structure as much as possible. Here, we chose PCA because the majority of information is kept while reducing the length of the data set to the number of specimens included and hence the demand for computational power. If the data contain fewer specimens (rows) than mass peaks (columns), the number of dimensions is automatically reduced by this PCA implementation to the number of specimens contained in the data set. t-SNE and UMAP are especially designed to deal with complex, high dimensional data, preserving local or global structures in the data (McInnes and Healy 2018), which is the key premise to find clusters. Because t-SNE is designed to reduce dimensionality for visualization purposes, we chose dimensionality reduction to three dimensions. UMAP on the other hand generally reduces the data to two dimensions. That is why, in contrast to the application of PCA in our workflow, t-SNE and UMAP reduce the number of dimensions more vastly saving additional computational power during data clustering.

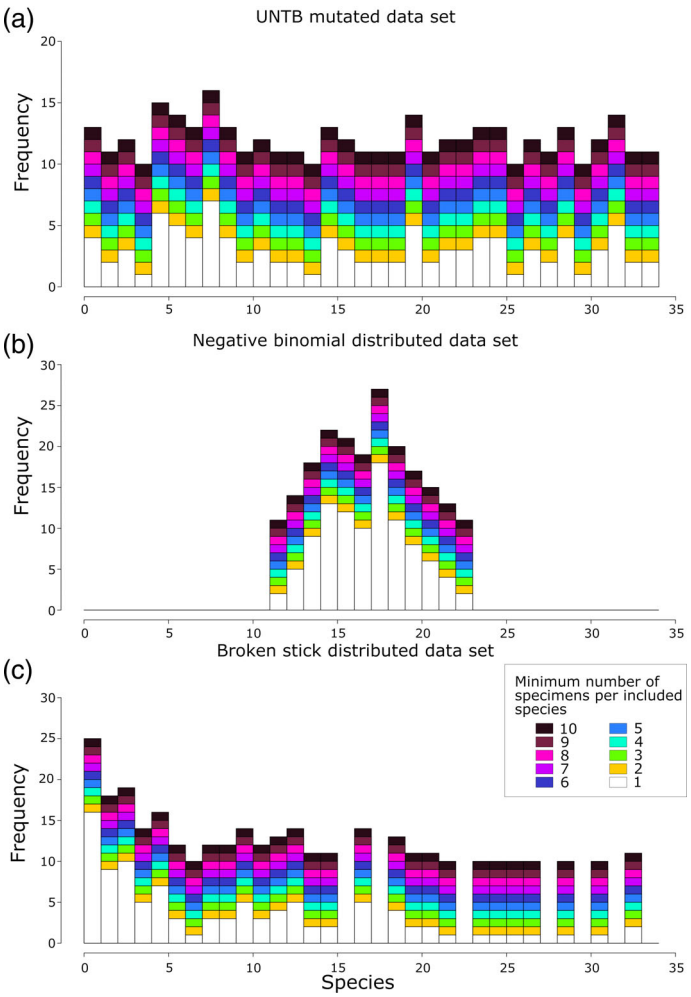

Fig. 2. Different species distribution patterns were used for the species dominance simulation. The colors indicate distributions that were used to test the influence of minimum number of specimens per species on the resulting biodiversity estimates. (**a**) UNTB distribution, (**b**) negative binomial distribution, and (**c**) broken stick distribution.

Table 1. Average percentage difference of all 48 combinations for $n_{species}$ ranging from 2 to 12, for $n_{species} > 12$ and for the entire data set.

Data transformation_dimensionality reduction_clustering method	% difference from correct value for $n_{species}$ = 2–12	% difference from correct value for $n_{species}$ = 13–34	Total % difference from correct value
Hellinger_PCA_PAM	5.546	3.342	4.08
Hellinger_none_PAM	5.546	3.342	4.08
Hellinger_TSNE_PAM	7.959	2.352	4.22
Hellinger_UMAP_PAM	11.981	1.633	5.10
Total_TSNE_PAM	9.840	3.435	5.57
p/a_TSNE_PAM	10.307	3.215	5.58
p/a_PCA_PAM	11.263	3.740	6.25
p/a_none_PAM	11.324	3.774	6.29
raw_TSNE_PAM	12.239	4.834	7.30
Total_UMAP_PAM	20.062	12.835	15.24
p/a_UMAP_PAM	20.944	12.477	15.30
raw_UMAP_PAM	19.894	24.551	23.00
raw_PCA_PAM	34.305	46.572	42.48
raw_none_PAM	34.305	46.572	42.48
Total_PCA_PAM	28.195	59.091	48.79
Total_none_PAM	28.195	59.100	48.80
Hellinger_PCA_KM	34.666	68.016	56.90
Hellinger_none_KM	34.666	68.042	56.92
Hellinger_PCA_MC	63.760	62.005	63.85
p/a_PCA_KM	47.414	79.831	69.03
p/a_none_KM	48.152	79.721	69.20
Total_PCA_MC	96.792	61.065	72.97
raw_PCA_MC	100.858	61.834	74.84
raw_none_KM	90.715	68.484	75.89
raw_PCA_KM	90.732	68.560	75.95
Hellinger_UMAP_MC	94.596	67.679	76.65
Total_none_KM	91.869	69.087	76.68
Total_PCA_KM	91.884	69.131	76.71
Hellinger_UMAP_KM	95.279	67.687	76.88
p/a_UMAP_KM	95.459	67.685	76.94
p/a_UMAP_MC	98.269	67.668	77.87
Total_UMAP_KM	98.251	67.686	77.87
raw_UMAP_KM	99.338	67.685	78.24
Total_UMAP_MC	101.631	67.664	78.99
p/a_PCA_MC	92.172	72.816	79.27
raw_UMAP_MC	104.253	67.661	79.86
Hellinger_TSNE_MC	112.703	67.633	82.66
Total_TSNE_MC	114.007	67.622	83.08
p/a_TSNE_MC	115.606	67.602	83.60
Total_TSNE_KM	115.465	67.680	83.61
Hellinger_TSNE_KM	115.703	67.678	83.69
raw_TSNE_MC	120.905	67.680	84.28
p/a_TSNE_KM	118.270	67.683	84.55
raw_TSNE_KM	117.732	67.553	85.42
Total_none_MC	128.582	67.692	87.99
raw_none_MC	130.214	67.690	88.53

(Continues)

Table 1. Continued

Data transformation_dimensionality reduction_clustering method	% difference from correct value for $n_{species}$ = 2–12	% difference from correct value for $n_{species}$ = 13–34	Total % difference from correct value
Hellinger_none_MC	130.257	67.692	88.55
p/a_none_MC	130.257	67.692	88.55

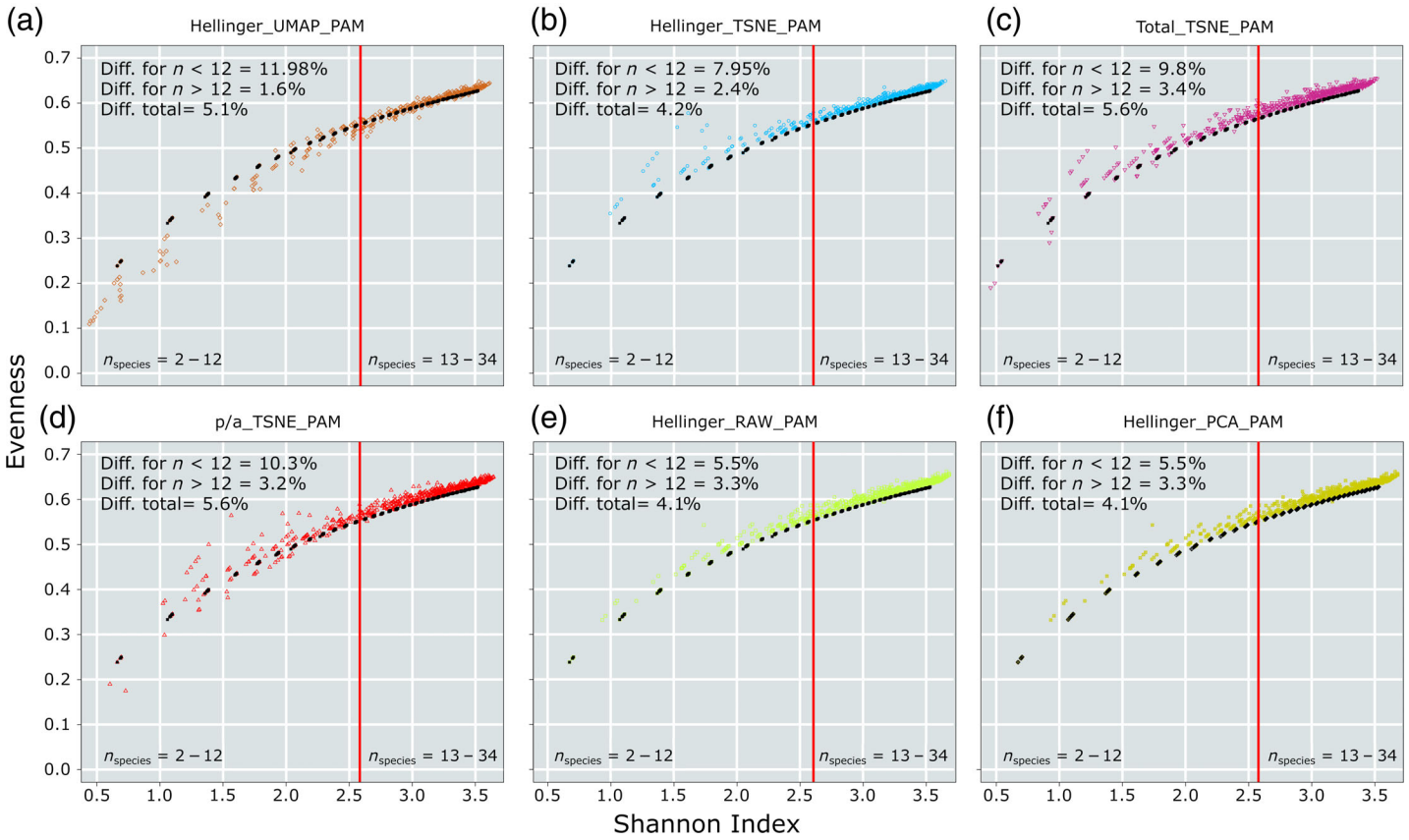

Fig. 3. Species evenness plotted against Shannon index for the six best performing combinations. Colored symbols display diversity estimations in comparison to the correct diversity (black). The average percentage difference to the correct diversity is displayed for all combinations of $n_{species}$ between 2 and 12, for $n_{species}$ larger than 12 and for the entire data set.

To estimate the number of species in the data set, three different cluster algorithms were applied. K-means was used as one of the most widely used cluster algorithms. Because in contrast to K-means clustering K-medoids is less sensitive to noise and outliers (Park and Jun 2009), PAM clustering was chosen as the most common K-medoids implementation. Both these methods depend on distances of data points to a suspected cluster mean or medoid. MC on the other hand uses Bayesian information criterion to find the best underlying model parameters for a Gaussian Mixture Model to assign clusters (Scrucca et al. 2016).

In total, 48 combinations of data transformations, dimensionality reductions and clustering methods were applied (data transformation × dimensionality reduction × clustering: 4 × 4 × 3) resulting in 31,680 analyses of the generated data sets. The number of clusters to test was chosen between 2 and the number of specimens –1.

While "MClust" and "cascadeKM" provided functions to assign the optimal number of clusters, for PAM clustering the average silhouette width after analyses for each number of proposed clusters was saved. The result showing the highest average silhouette width was chosen as the best number of clusters (Rousseeuw 1987).

From all clustering approaches, the best number of estimated clusters was used to calculate the Shannon index and species evenness. To find the best performing combinations, the mean percentage deviation from the correct diversity was calculated for the entire range of species but also for all species numbers individually.

Species dominance simulation

To test the best six approaches for "real world scenarios", data sets containing 100 specimens were generated according to three typical species distributions. From the original data set, simulated mass spectra (n = 3,400) using "smooth.data" from the R-package "RFtools" (Martínez Arbizu and Rossel 2018; Rossel and Martínez Arbizu 2018*a*) were generated to increase the number of spectra, from which a population can be simulated.

To create a data set with data distributed according to the Unified Neutral Theory of Biodiversity and Biogeography (UNTB), the 3400 mass spectra (100 mass spectra per species) were used as a metacommunity which was mutated using the command "untb" from the R-package "untb" (Hankin 2007). The mutation was carried out with a probability of new organisms not being a descendant of an existing individual of zero, five organisms that die in each time step and 10,000 simulated generations. The results were adjusted to a community of 100 specimens (Fig. 2a). The Negative binomial (Nb) distribution (Fig. 2b) was generated in R using "rbinom" with a probability of success of 0.5. The command "rrbs" from the R-package "sads" (Prado et al. 2018) was used to generate a Broken stick (BS) distribution (Fig. 2c). According to these distributions, communities were sampled out of the 3,400 mass spectra using the command "stratsample" from the R-package "survey" (Lumley 2004).

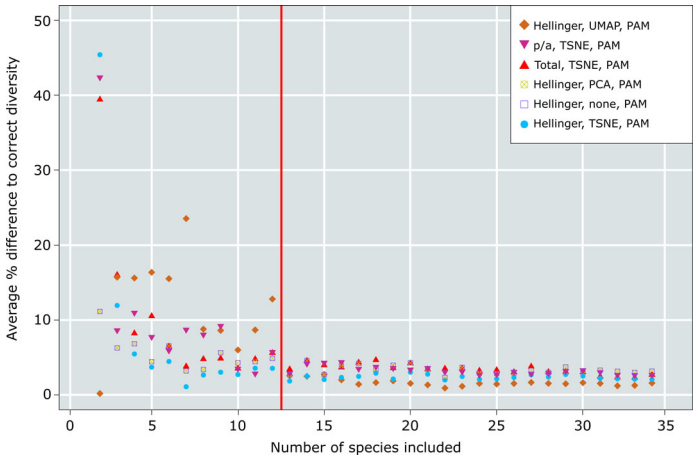

Fig. 4. Average percentage difference to correct diversity for all $n_{species}$. With $n_{species} > 12$ the variance of percentage difference and the difference itself decreases. While Hellinger transformed data with UMAP dimensionality reduction and PAM clustering performs worst with $n_{species} < 12$, it performs best with $n_{species} > 12$.

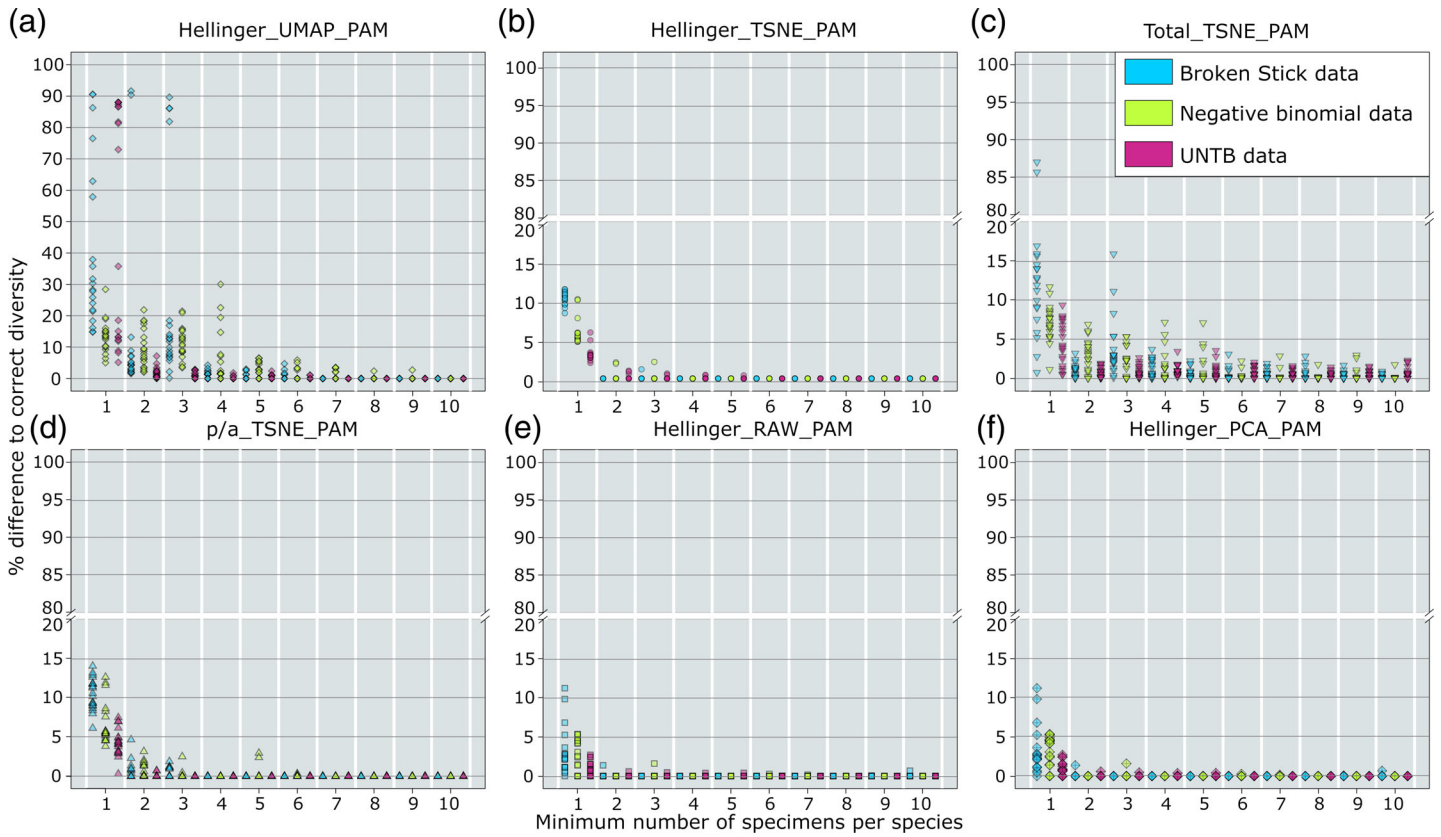

Fig. 5. Average percentage difference to correct diversity for the three different tested species distribution patterns. It is displayed how the difference changes from a minimum number of specimens per species of one to 10.

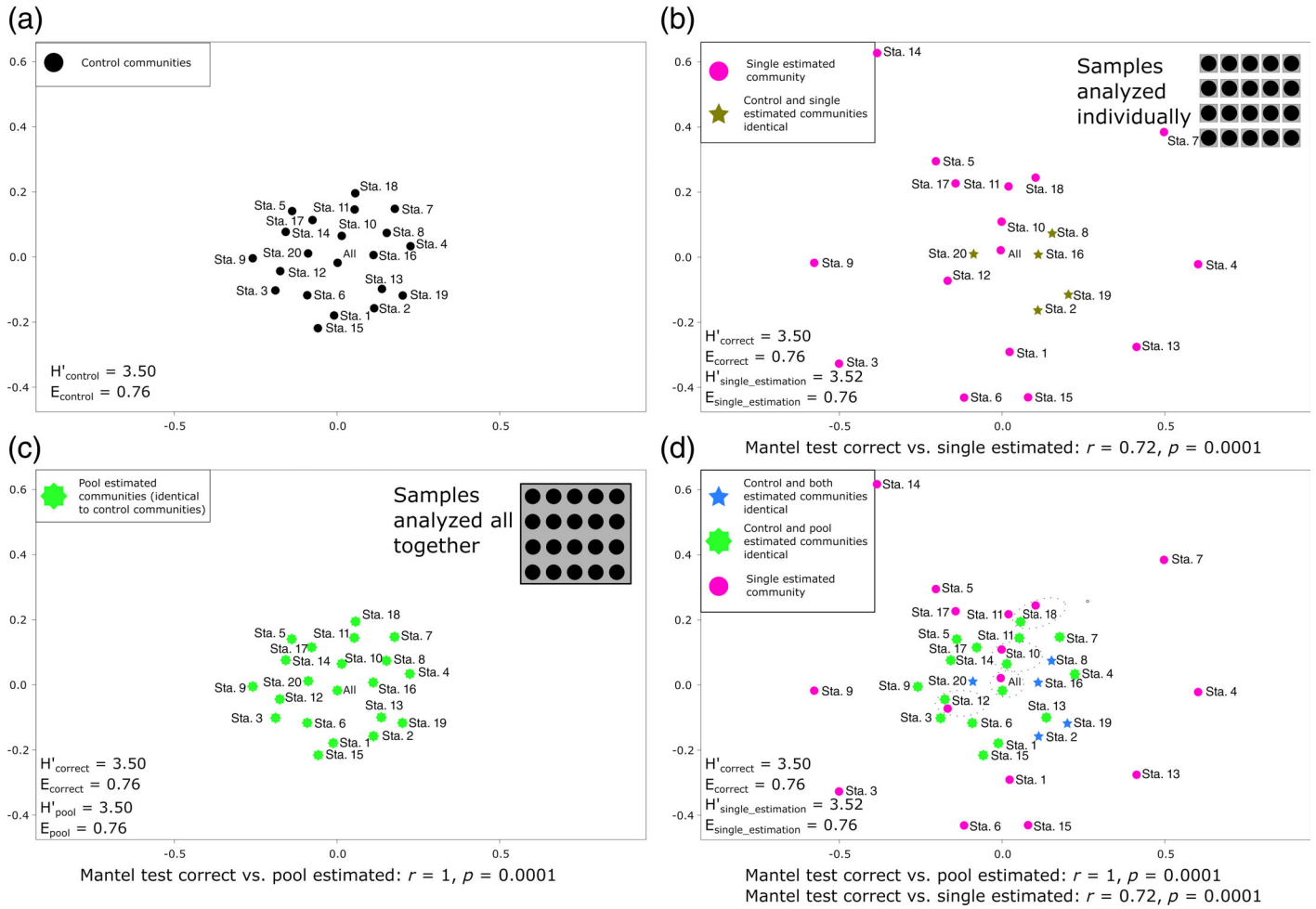

Fig. 6. MDS plots displaying analyzed stations in the same coordinate system. (**a**) Control data, (**b**) individually estimated data, and (**c**) pool estimated data. (**d**) Displays the different communities plotted onto each other. In (d), simulated stations which were identical in all three data sets are marked with a blue star. Stations in which only control and pool analyzed data were identical are marked in green. Marked as "all" is the community from all stations combined. Stations from different analyses that are very similar but still show different positions within the MDS plot are encircled by dotted lines. In (b) and (c), the analysis setup is displayed. In (b), simulated stations were analyzed individually while in (c) all were analyzed together. Stations occurring twice in (d) are stations that deviate from the control communities.

To test the influence of the minimum number of specimens per species (n_{min}) on the resulting diversity estimates, n_{min} was varied between 1 and 10. For each n_{min}, 20 data sets were generated. For n_{min} = 1, the distributions were generated as described above. To keep the distribution pattern for tests of n_{min} from 2 to 10, in each step one specimen per included species was added (Fig. 2, colored distributions). Shannon index was assessed for all data sets and the percentage deviation from the correct values was calculated.

Diversity estimated individually vs. pooled estimation

Finally, a scenario was simulated in which 20 stations with varying combinations of 29 different species were generated. All stations contained 100 specimens distributed according to BS distribution of which 10 were singleton species (number of specimens per species equals one). Species clusters were estimated either individually (100 specimens per analysis) or all stations pooled together (2,000 specimens per analysis). Communities resulting from both analyses were Hellinger transformed and plotted together with control communities in an MDS plot to check for congruence. Furthermore, a Mantel test for both communities was carried out to compare it to control communities using 9,999 permutations.

Results

In total, 31,680 analyses were carried out and estimated species diversities for all 48 combinations of data transformation, dimensionality reduction and clustering methods were evaluated (Table 1; Fig. 3). The nine best estimates with a percentage difference from the correct diversity of less than 10% used PAM clustering. These were followed by seven further

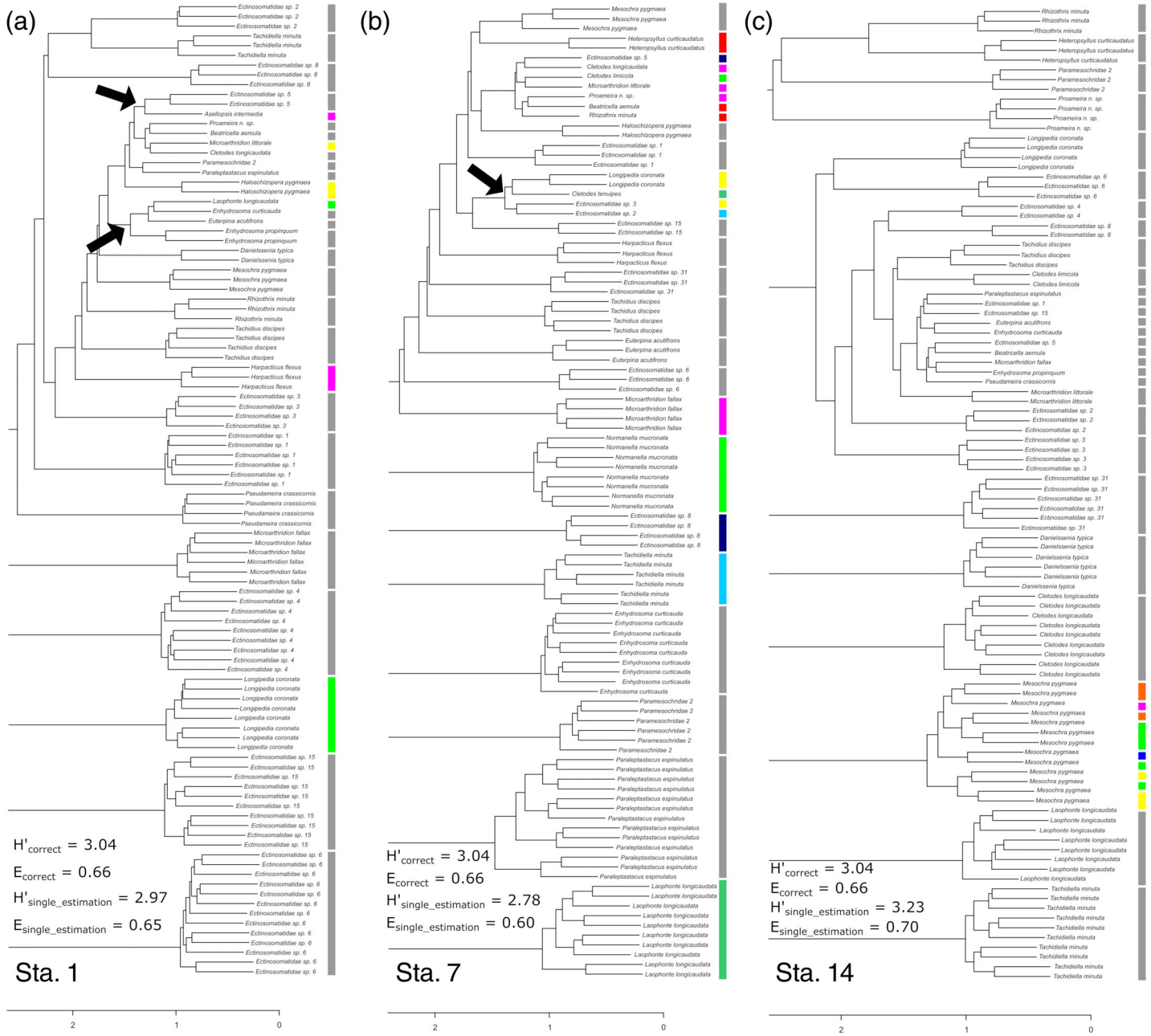

Fig. 7. Hierarchical cluster analyses of simulated sta. 1, 7, and 14 of the community analyses displayed in figure 6. To the right of each tree, the results of individual cluster estimations are plotted. The colored bars display the errors that occurred during cluster assignment. Gray bars indicate clusters that were estimated correctly. Ranging from only a few singletons that were assigned to the wrong clusters (**a**) to all singletons that were classified erroneously (**b**). Finally (**c**) depicts a case in which during individual diversity estimation a single species was split into five different clusters. In the lower left hand corner, the Shannon index and species evenness are displayed. Arrows highlight clusters for which delimitation based on expert opinion can be considered difficult.

estimates including PAM as clustering algorithm (difference < 50%). Percentage difference of all diversity estimates including K-means or MC clustering ranged from 56.90% to 88.55%.

Notably, most combinations tend to overestimate diversity (Fig. 3b–f) while only the combination of Hellinger transformation with UMAP and PAM underestimates diversity within the range of 2 to 12 analyzed species (Fig. 3a). The best estimations with only 5.55% difference from the correct values were made with Hellinger transformed data analyzed without dimensionality reduction using PAM clustering or with PCA for dimensionality reduction. These two

combinations also showed the lowest overall difference from the correct diversity. However, in the range from 13 to 34 analyzed species in the data set, a combination of Hellinger transformation, UMAP, and PAM performed best with only 1.63% difference, followed by Hellinger transformation, t-SNE, and PAM with 2.35% difference. With increasing number of species included in the estimation, the average deviation to the correct diversity decreased (Fig. 4). The variance in average differences reduced when the number of included species exceeded 12 (Fig. 4). However, the best six estimates (Table 1; Fig. 3) generally showed quite similar results. Hence, these were used for further tests including dominance of species as it would be expected in actual samples.

Species dominance simulation

As species in real samples are typically not evenly distributed but follow different distribution models, communities according to three different species distributions were generated. The six best performing combinations were used to estimate diversity of communities containing a certain number of minimum specimens per species (n_{min}) ranging from 1 to 10 (Fig. 2). For each n_{min}, 20 data sets were analyzed resulting in 3,600 analyses. For each analysis, the percentage difference of the estimation to the correct diversity was evaluated and plotted separately in Fig. 5. All combinations failed to consistently provide estimations that are congruent with the correct diversities when singletons were included in the analyzed communities (Fig. 5). However, even when $n_{min} = 1$, it was possible to obtain some correct estimations from Hellinger transformed data using either PCA or no dimensionality reduction with PAM clustering (Fig. 5e,f). The differences of estimations to the correct diversity were generally higher for BS distributed data sets and lower in the UNTB data sets. Most approaches performed almost flawless when $n_{min} > 3$ (Fig. 5b, d–f). Percentage differences were higher for analyses using PAM clustering with Hellinger transformed data after UMAP was applied (Fig. 5a) and PAM clustering of Species profile transformed data after t-SNE dimensionality reduction (Fig. 5c).

Individual estimation vs. pooled estimation

For the final test of applicability in a "real world scenario", 20 stations with BS distributed species (Fig. 6a) were simulated and either analyzed all together or individually (Fig. 6b,c, upper right hand corner). By creating 20 stations with different distributions of the same 29 species, we wanted to simulate sampling within a wider study area, where chances of reoccurring species in different samples is high. Results showed good concordance of species communities as displayed in Fig. 6d. This also resembled by high *r*-values of the Mantel tests. Individually estimated communities pooled together still showed an *r*-value of 0.72 (Fig. 6b). The pooled estimated stations showed an *r*-value of 1 compared to the correct communities (Fig. 6c). Here, biodiversity and species evenness were estimated 100% correctly. The estimated Shannon index for pool estimation and control communities was 3.50. Of the 20 individually estimated communities, 5 were absolutely congruent to the correct data (Fig. 6b, yellow stars). Deviating species compositions for some communities were often caused by singleton specimens included in the analysis. However, not all singletons were automatically assigned to other clusters. Sta. 1 for instance contained 10 singleton species of which only 3 were assigned to other clusters. Nevertheless, the majority of false cluster assignments were caused by singleton species and sometimes all singletons were assigned to other species (Fig. 7b). In one case, a single species was split into numerous clusters. At sta. 14, the species *Mesochra pygmaea* (Claus 1863) ($n = 13$) was split into five different clusters (Fig. 7c). This resulted in an all overestimated Shannon index slightly higher than the correct distribution ($H = 3.52$).

Discussion

Accuracy across the different setups

The initial tests clearly showed that several combinations of data transformations and dimensionality reduction together with PAM clustering provided diversity measures highly congruent to the correct values.

However, the clustering success differed strongly when only few species were included ($n_{species} < 12$). That is why, even though Hellinger transformation with UMAP dimensionality reduction and PAM clustering only showed lower percentage difference from the correct diversity when $n_{species} > 12$ (1.6%), we recommend using either Hellinger transformed data with PAM clustering without applying a dimensionality reduction or using PCA. These approaches showed the lowest deviation from the correct value on average for the entire examined range of species numbers (4.08%). This is also supported by the results of the species dominance simulation analyzing specimen numbers ranging from 1 to 10 for each species. Here, the aforementioned methods showed only low deviation from the correct diversity even though singletons were included in the analyses. The number of singletons also affected the differences in estimations for the different species distribution patterns. Deviations of BS distributed data were always found to be a bit higher compared to other distributions as these included more singleton species than the others. However in a test with 20 simulated stations, species distributions of 5 stations were perfectly in congruence with the correct data when estimations were carried out for each station individually. Moreover, for some stations, distribution patterns deviating only slightly from the correct distributions were estimated. This is also resembled by the high Mantel test *r*-value when comparing the allover result of the single estimated communities to control communities.

Finally, the test on 20 simulated stations showed that including several stations from a certain geographic area and repetitive sampling at certain stations, which show

overlapping species occurrences, should be favored over analyzing single stations individually. Repetitive sampling and sampling of different stations in a geographic area will likely increase number of specimens for each species and hence result in a more robust result for this unsupervised clustering approach. This follows the general recommendations to increase sample size to increase power in statistical tests (Fairweather 1991). However, when including additional stations, care must be taken to include stations which will likely include similar species as doing otherwise may again lead to inclusion of further singleton species. Regarding Harpacticoida for instance, deep-sea studies frequently deal with low-frequency occurrences of species and numerous singletons (Schmidt et al. 2019). In such cases, analyzing stations individually would inevitably lead to high errors that have to be taken into account when working with the results. That is why analyzing samples from repeated sampling at a station conjointly increases chances of obtaining good biodiversity measures. In contrast to deep-sea studies, shallow-water studies often deal with lower species diversities and higher specimen numbers per species (Packmor and George 2018). That is why applying the described workflow should result in receiving good biodiversity measures.

Data transformation, dimensionality reduction, and clustering methods

Throughout the tests, it was shown that an additional relative data transformation vastly improve the results. In contrast to using p/a data, data from an additional relative data transformation also take peak intensities into account. Due to Hellinger transformation, peak intensities are more similar between specimens from the same species with smaller intraspecific distances during clustering, resulting in more distinct, species-specific clusters compared to total data transformation. The additional relative data transformations give less weight to less intense mass peaks (Legendre and Gallagher 2001), thus reducing the influence of frequent zeros making it feasible for MALDI-TOF MS data sets. Moreover, applying an additional transformation including a square root transformation probably diminishes the influence of outliers compared to total data transformation. However, relative intensities of peaks as an important factor for grouping mass spectra into clusters are neglected when p/a transformation is carried out. Increased interspecific difference as a result of varying abundances of the same peptide or protein between species does not influence groupings because peaks are considered as present without providing further information about species specificity of molecule abundances.

When PCA is used for dimensionality reduction of data sets containing fewer rows than columns, the applied PCA implementation reduces the number of dimension to the number of specimens in the data set. Hence, the number of analyzed parameters equals the number of specimens in the data set. Thus, this dimensionality reduction results in the loss of only a negligible part of the information. That is why applying PCA and working without dimensionality reduction resulted in the same clusters after Hellinger transformation. We recommend using PCA only when the former number of traits is larger than the number of specimens included. Otherwise the calculation time is artificially delayed. If very large data sets in terms of included specimens have to be analyzed, using Hellinger transformation with t-SNE dimensionality reduction and PAM could be an alternative to using raw data. Because the number of dimensions is reduced to three, the computing time for clustering is also highly reduced. Deduced from the results we obtained, the estimated biodiversity should only deviate slightly more from the correct values than for the aforementioned approaches as this approach showed comparatively good results also when singletons were included.

In contrast to PAM clustering, both K-means clustering and MC largely failed to produce biodiversity measures closely resembling correct values. While PAM clustering uses a hypothetical point centroid within the data as starting point to find clusters, K-means uses an actual data point to begin with. This makes K-means clustering, for instance, more sensitive to outliers (Park and Jun 2009), resulting in differing number of cluster. The problem with MC on the other hand may be that this method assumes an underlying distribution of the data which may be different from the models included in this methods implementation.

Potential pitfalls in biodiversity assessments

When applying the workflow presented here, attention has to be paid to data origin and quality. For instance, joining data from different studies may cause problems as various factors such as storage and fixation were shown to influence mass spectra quality (Yssouf et al. 2014; Rossel and Martínez Arbizu 2018*c*). This may affect the species-specific fingerprint in such a way that single species could be recognized as different species due to mass spectra alteration. Also, prior to application in cross laboratory studies, it needs to be tested if in inter laboratory analyses the species-specific fingerprint is retained or if mass spectra obtained from different instruments display different complexities causing the same abovementioned outcome. Nevertheless, for bacteria, cross laboratory tests have already been carried out successfully (Mellmann et al. 2009). Hence, this may not be a problem during unsupervised biodiversity estimation. Moreover, accurate binning of homologous mass peaks is crucial to reduce artificial variability within a species that might cause splitting single species into several very similar species. Peak binning is carried out to merge homologous data into discrete bins. To date, different methods for peak binning such as using predefined bins of a certain width (Laakmann et al. 2013; Kehrmann et al. 2016), clustering of peaks to group into certain m/z values (Chen et al. 2007) or using other cross spectra algorithms (Gibb and Strimmer 2012). However, sometimes homologous peaks are binned into different m/z values resulting in artificially

divergent peaks within a single species. Hence, attention must be paid to peak binning and alignment to obtain reliable results.

The same is true for different developmental stages. For calanoid copepods it was shown by different authors that, within species clusters from hierarchical clustering, different developmental stages can be found in separated, stage-specific clusters (Laakmann et al. 2013; Bode et al. 2017; Kaiser et al. 2018). Thus, it should be paid attention to developmental stages of analyzed specimens when preparing such microscopic species.

Comparison to DNA-based delimitation methods

With this study, we try to provide a method for unsupervised biodiversity estimation. This includes unsupervised species delimitation based on mass spectrometry data as congruent to real species (according to COI and morphological investigations) as possible. To date, analyzing mass spectrometry data for species identification is mainly carried out in supervised approaches (Yssouf et al. 2013). Reference libraries are searched and a species assignment is returned with a certain probability of identification correctness. However, if species are not part of such a reference library, supervised methods cannot delimit these. To differentiate between species, clustering can be applied. However, hierarchical clustering itself does not provide information on cluster margins and hence, researchers themselves have to evaluate species boundaries from cluster analyses. Arrows in Fig. 7 emphasize clusters, in which it is at least difficult to delimit species based on, for instance, branch lengths.

In DNA barcoding (Hebert et al. 2003), several methods are already available to delimit species. In contrast to our approach that tries to find cluster boundaries, ABGD searches for a barcoding gap in pairwise distances to successfully delimit species. GMYC on the other hand delimits species based on a change in evolutionary speed among branches of an ultrametric tree (Puillandre et al. 2012*a*). This however demands a phylogenetic tree which cannot be computed based on mass spectrometry data, as for metazoans MALDI-TOF MS data were shown only once to reflect evolutionary relationships (Feltens et al. 2010). Both techniques provide good results on species boundaries and are used frequently to assess species diversity in numerous studies (Puillandre et al. 2012*b*; Lin et al. 2018). This supports the demand for such delimitation methods also in mass spectrometry applications for biodiversity studies. Especially in research fields and taxonomic groups with high number of undescribed species and difficult morphological identifications, such a method can help to accelerate biodiversity assessments. Instead of demanding morphological identification of all analyzed specimens, only a few specimens for every cluster have to be examined to be able to assign clusters to morphological working species.

Conclusion

Our unsupervised biodiversity estimation workflow allows comparison of sampling sites in ecological studies without prior knowledge of the species occurring in the working area and without the need for complete reference libraries. As was shown, the average difference to the correct biodiversity measures is little and often estimates are perfectly congruent to the actual distribution of species. However, caution must be taken on data quality as this will likely affect results.

Data availability statement

No new data were collected for this study. The sources of the used data are given in Supplementary Table S1.

References

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

Breimann, L. 2001. Random forests. Mach. Learn. **45**: 5–32.

Carrera, M., B. Cañas, and J. M. Gallardo. 2013. Proteomics for the assessment of quality and safety of fishery products. Food Res. Int. **54**: 972–979.

Chen, S., D. Hong, and Y. Shyr. 2007. Wavelet-based procedures for proteomic mass spectrometry data processing. Comput. Stat. Data Anal. **52**: 211–220.

Dvorak, V., P. Halada, K. Hlavackova, E. Dokianakis, M. Antoniou, and P. Volf. 2014. Identification of phlebotomine sand flies (Diptera: Psychodidae) by matrix-assisted laser desorption/ionization time of flight mass spectrometry. Parasit. Vectors **7**: 21.

Fairweather, P. G. 1991. Statistical power and design requirements for environmental monitoring. Mar. Freshw. Res. **42**: 555–567.

Feltens, R., R. Görner, S. Kalkhof, H. Gröger-Arndt, and M. von Bergen. 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.

Fraley, C., and A. E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**: 611–631.

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

Hankin, R. K. 2007. Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. J. Stat. Softw. **22**: 1–15.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. **270**: 313–321.

Holst, S., A. Heins, and S. Laakmann. 2019. Morphological and molecular diagnostic species characters of Staurozoa (Cnidaria) collected on the coast of Helgoland (German Bight, North Sea). Mar. Biodivers. **49**: 1775–1797. doi:10.1007/s12526-019-00943-1

Hynek, R., S. Kuckova, P. Cejnar, P. Junková, I. Pvrikryl, and J. Rihová Ambrovzová. 2018. Identification of freshwater zooplankton species using protein profiling and principal component analysis. Limnol. Oceanogr.: Methods **16**: 199–204.

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

Kaufman, L., and P. J. Rousseeuw. 1990. Finding groups in data: An introduction to cluster analysis. New York: Wiley.

Kehrmann, J., S. Wessel, R. Murali, A. Hampel, F.-C. Bange, J. Buer, and F. Mosel. 2016. Principal component analysis of MALDI TOF MS mass spectra separates *M. abscessus* (sensu stricto) from *M. massiliense* isolates. BMC Microbiol. **16**: 24.

Konopka, T. 2018. UMAP: Uniform manifold approximation and projection. Available from https://CRAN.R-project.org/package=umap

Krijthe, J. H. 2015. Rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation. Available from https://github.com/jkrijthe/Rtsne

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

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia **129**: 271–280.

Lin, X.-L., E. Stur, and T. Ekrem. 2018. Exploring species boundaries with multiple genetic loci using empirical data from non-biting midges. Zool. Scr. **47**: 325–341.

Lloyd, S. 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**: 129–137.

Lumley, T. 2004. Analysis of complex survey samples. J. Stat. Softw. **9**: 1–19.

van der Maaten, L., and G. Hinton. 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. **9**: 2579–2605.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2018. Cluster: Cluster analysis basics and extensions. R Package Version 2.0.1.

Majchrzykiewicz-Koehorst, J. A., E. Heikens, H. Trip, A. G. Hulst, A. L. de Jong, M. C. Viveen, N. J. Sedee, J. van der Plas, F. E. Coenjaerts, and A. Paauw. 2015. Rapid and generic identification of influenza A and other respiratory viruses with mass spectrometry. J. Virol. Methods **213**: 75–83.

Martínez Arbizu, P., and S. Rossel. 2018. RFtools: Miscellaneous tools for random forest models. Available from https://zenodo.org/record/118843

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

McInnes, L., and J. Healy. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mellmann, A., F. Bimet, C. Bizet, A. Borovskaya, R. Drake, U. Eigner, A. Fahr, Y. He, E. Ilina, M. Kostrzewa, et al. 2009. High interlaboratory reproducibility of matrix-assisted laser desorption ionization-time of flight mass spectrometry-based species identification of nonfermenting bacteria. J. Clin. Microbiol. **47**: 3732–3734.

Normand, A.-C., C. Cassagne, M. Gautier, P. Becker, S. Ranque, M. Hendrickx, and R. Piarroux. 2017. Decision criteria for MALDI-TOF MS-based identification of filamentous fungi using commercial and in-house reference databases. BMC Microbiol. **17**: 25.

Oksanen, J. and others. (2013). Package "vegan." Community ecology package, version 2.

Packmor, J., and K. H. George. 2018. Littoral Harpacticoida (Crustacea: Copepoda) of Madeira and Porto Santo (Portugal). J. Mar. Biol. Assoc. U.K. **98**: 171–182.

Park, H.-S., and C.-H. Jun. 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. **36**: 3336–3341.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Phil. Mag. J. Sci. **2**: 559–572.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. **55**: 595–609.

Prado, P. I., M. D. Miranda, and A. Chalom. 2018. sads: Maximum likelihood models for species abundance distributions. Available from https://CRAN.R-project.org/package=sads

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012a. ABGD, automatic barcode gap discovery for primary species delimitation. Mol. Ecol. **21**: 1864–1877. doi:10.1111/j.1365-294X.2011.05239.x

Puillandre, N., M. Modica, Y. Zhang, L. Sirovich, M.-C. Boisselier, C. Cruaud, M. Holford, and S. Samadi. 2012b. Large-scale species delimitation method for hyperdiverse groups. Mol. Ecol. **21**: 2671–2691.

Rossel, S., and P. Martínez Arbizu. 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. **00**: 1–14.

Rossel, Sven., Martínez Arbizu, Pedro. 2018b. Data from: Effects of sample fixation on specimen identification in biodiversity assemblies based on proteomic data (MALDI-TOF), Dryad, Dataset. https://doi.org/10.5061/dryad.1md2jq1

Rossel, S., and P. Martínez Arbizu. 2018c. Effects of sample fixation on specimen identification in biodiversity assemblies based on proteomic data (MALDI-TOF). Front. Mar. Sci. **5**: 149.

Rossel, Sven., Martínez Arbizu, Pedro. 2019a. Revealing higher than expected diversity of *Harpacticoida* (Crustacea: Copepoda) in the North Sea using MALDI-TOF MS and molecular barcoding. Sci. Rep. **9**: 9182. https://doi.org/10.5061/dryad.f8s1f6m

Rossel, S., and P. Martínez Arbizu. 2019b. Data from: Revealing higher than expected diversity of Harpacticoida (Crustacea: Copepoda) in the North Sea using MALDI-TOF MS and molecular barcoding, v2, Dryad, Dataset. https://doi.org/10.5061/dryad.f8s1f6m.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**: 53–65.

Schmidt, C., V. V. Sattarova, L. Katrynski, and P. M. Arbizu. 2019. New insights from the deep: Meiofauna in the Kuril-Kamchatka trench and adjacent abyssal plain. Prog. Oceanogr. **173**: 192–207.

Scrucca, L., M. Fop, T. B. Murphy, and A. E. Raftery. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. **8**: 289.

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

Yssouf, A., C. Socolovschi, C. Flaudrops, M. O. Ndiath, S. Sougoufara, J.-S. Dehecq, G. Lacour, J.-M. Berenger, C. S. Sokhna, D. Raoult, and P. Parola. 2013. Matrix-assisted laser desorption ionization-time of flight mass spectrometry: An emerging tool for the rapid identification of mosquito vectors. PLoS One **8**: e72380.

Yssouf, A., C. Socolovschi, H. Leulmi, T. Kernif, I. Bitam, G. Audoly, L. Almeras, D. Raoult, and P. Parola. 2014. Identification of flea species using MALDI-TOF/MS. Comp. Immunol. Microbiol. Infect. Dis. **37**: 153–157.

Acknowledgments

We thank Katja Uhlenkott for her introduction to cloud computing that helped accelerating the analyses carried out in this study. And we thank the anonymous reviewer for helpful comments improving the manuscript quality. This is publication no 9 of Senckenberg am Meer Proteome Laboratory.

Conflict of Interest

None declared.

Submitted 18 June 2019
Revised 03 March 2020
Accepted 31 March 2020

Associate editor: Michael Beman