Received: 2 May 2023 | Accepted: 24 October 2023

DOI: 10.1111/een.13297

ORIGINAL ARTICLE

Ecological Entomology

DNA metabarcoding of light trap samples vs. morphological species identification

Patrick Strutzenberger | **Brigitte Gottsberger** | **Florian Bodner** | **Fabian Bartusel** | **Daniel Jerga** | **Konrad Fiedler**

Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

Correspondence
Patrick Strutzenberger, Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria.
Email: patrick.strutzenberger@univie.ac.at

Associate Editor: Christopher Hassall

[Correction added on 11 December 2023, after first online publication: The first and last name of the authors: Brigitte Gottsberger, Fabian Bartusel and Daniel Jerga have been corrected.]

Abstract

1. DNA metabarcoding has developed into a commonly used tool for biodiversity assessment and monitoring. How results from DNA metabarcoding are compared with studies based on 'classic', in most cases morphological species identification, is still unclear. Studies investigating species detection against a known baseline are virtually non-existent.
2. In this study, we used light trap samples collected in eastern Austria to investigate the concordance between morphological species lists and results obtained from COI metabarcoding using the Illumina MiSeq platform. Two primer combinations of different lengths (313 and 205 bp) were compared to assess the influence of amplicon length.
3. Species detection rates ranged between 0.38 and 0.69; the shorter amplicon had on average higher species detection rates compared with the longer amplicon. Singleton species were less likely to be detected through metabarcoding.
4. The major determinant for a species to be detected was its biomass, viz. smaller species had a lower chance to be detected. However, there is also evidence of taxonomic bias on the level of superfamilies. While the influence of biomass is to be expected, the presence of taxonomic bias gives reason for concern and requires further studies. Such a bias can be of significance when metabarcoding is used to determine conservation measures. Ordination analyses of all sampling sites showed that as far as community ecology is concerned, the overall pattern obtained from the full species list was mostly preserved in our metabarcoding results.

KEYWORDS
arthropods, biodiversity, community ecology, monitoring, moths

INTRODUCTION

DNA-metabarcoding approaches have emerged to become a standard tool in biodiversity monitoring and discovery, serving as a versatile way to support community-wide assessments especially in the fields of ecology and biodiversity research (Liu et al., 2020). Such approaches have proven to be successful over a wide range of taxa from bacteria to vertebrates in all kinds of habitats, from aquatic and soil biota to terrestrial ecosystems (Abdelfattah et al., 2018; Semenov, 2021; van der Loos & Nijland, 2021).

Most applications of DNA-metabarcoding can be classified into two major categories: (1) studies on already known assemblages (e.g., biodiversity monitoring and biosecurity: Piper et al., 2019) and (2) studies on mostly unknown assemblages (e.g., species discovery), especially in tropical systems or microbial biota (Abdelfattah

et al., 2018). Both of these applications have different requirements and challenges. Monitoring applications imply some degree of previous knowledge of the species assemblages to be studied. Missing or unexpectedly detected species will likely be noticed as such and can be properly addressed. Species discovery applications on the other hand have no or very little means to adequately deal with improper recovery of operational taxonomic units (OTUs).

Morphological identification of organisms, especially in speciose target groups, is time-consuming and requires taxonomic expertise, which depending on the target taxon can be rare or even completely unavailable. The lack of taxonomic expertise for a certain taxon can be caused by phenomena commonly described as the 'taxonomic impediment' (Giangrande, 2003) or might pertain to 'dark biodiversity' (Lewis et al. (2017), Page (2016)). DNA barcoding of individual specimens is very time-consuming and processing of large samples requires considerable effort and resources. Metabarcoding holds the promise to reduce the effort that comes along with individual-based barcoding (Yu et al., 2012). Furthermore, metabarcoding provides efficient means to measure biodiversity in taxa where means for morphological identification are not available. On the other hand, metabarcoding is so far unable to provide quantitative data, although first steps have been made towards quantitative metabarcoding (Shelton et al., 2023).

A central concern with all metabarcoding applications is the achieved detection rate of OTUs. In real-world samples, the actual number of OTUs present in the sample is a priori unknown. Estimations on the efficacy of such approaches to detect all species in a bulk sample are commonly based on the sequencing of mock communities (Elbrecht et al., 2019). While those mock communities are selected to include a broad taxonomic range as well as represent variations in body size, such mock communities usually include only a single individual per OTU (Elbrecht et al., 2019). Studies using mock communities are, therefore, unable to estimate the effects of varying OTU abundances on the level of detection. The relevance of mock community sequencing on the coverage achieved in real samples is, therefore, limited by design. Some studies (Mata et al., 2021) have investigated species detection rates of morphological determination versus metabarcoding, but data on detection rates of a defined set of species is still lacking. Studies investigating detection rates in real-world samples using a baseline of morphologically determined species are very rare (Elbrecht et al. 2017a). It is well established that small organisms are less likely to be detected in metabarcoding approaches. This has different implications for species discovery as opposed to studies of an already well-known assemblage. This problem is exacerbated if many species in a community are small-sized and at the same time also rare, that is, represent an overall low fraction of assemblage biomass, but make up a sizeable fraction of biodiversity in terms of species numbers as well as functional attributes. Comparisons based on real-world samples have so far relied on relative comparisons only; absolute comparisons against a known baseline are still lacking. Various strategies besides primer affinity to increase OTU recovery rates are being employed in metabarcoding studies. Most of those instances are concerned with different strategies of sample preparation. Some have argued for extensive pre-sorting (Braukmann et al., 2019; Elbrecht et al. 2017b; Elbrecht et al., 2021), whereas other studies on terrestrial arthropods did not perform any pre-sorting (Hausmann et al., 2020). Sequencing strategies that encompass pre-sorting involve the separation of the specimens contained in the sample into a number of size classes. Most applications of pre-sorting employ separate DNA extraction of each fraction; DNA extracts are then pooled with ratios adjusted to the biomass contained in each fraction (Elbrecht et al. 2017b). Semi-automated strategies employing sample fractionation using a selection of subsequently more selective sieves in an electric sieve-shaker have also been proposed (Elbrecht et al., 2021). Regardless of the strategy used, pre-sorting comes with a substantial and often prohibitive effort both in time and the required personnel, alleviating a considerable part of the alleged practical advantages of metabarcoding approaches (Elbrecht et al., 2021). Pre-sorting also introduces additional sources of sample contamination as well as observer error when manual sorting is used.

In this study, we use Cytochrome oxidase subunit I (COI) metabarcoding and take a nested calibration approach to estimate OTU detection rates in temperate-zone terrestrial insect assemblages. To do so, we manually identified a subset of our real-world samples before sequencing. This subset serves as calibration enabling us to estimate the OTU recovery rate in a real-world scenario.

The aims of this study are as follows:

1. Explore detection rates of a select set of species as a part of complex light trap samples.
2. Compare the efficacy of two different amplicon lengths. Longer amplicons facilitate sequence matching, whereas shorter amplicons are easier and cheaper to sequence.
3. Investigate potential bias in detection rates and explore the underlying factors.
4. Explore the practical consequences of incomplete species detection for the study of species assemblages.

MATERIALS AND METHODS

Field work

Samples were collected in two forest stands in the Danube river flood plains in eastern Austria, situated in the National Park Donau-Auen about 15 km east of Vienna and separated from another by a levee. The forest that stands north of the levee is decoupled from inundation dynamics for 150 years, whereas the forest parts south of the levee still are subject to riverine floods almost every year. Accordingly, local insect communities differ substantially in terms of species composition, species and functional diversity between these two forest fractions (Guariento et al., 2020; Truxa & Fiedler, 2012). Insect assemblages are also characterised by large fractions of 'rare' species that show up only as singletons or doubletons in samples (Truxa & Fiedler, 2016). Within each forest stand, five automatic light traps were put up about 1.5 m above ground, spaced 100 m apart, for one night per month between March and August 2020 (see Bartusel, 2021 for detailed locations). Light traps were constructed as in Axmacher and Fiedler (2004) and equipped with a multi-spectral LED light source (Brehm, 2017). Because of their mode of construction and placement, they targeted flying

insects, whereas flightless invertebrates might only be trapped by chance. Samples were stored at −20°C after transfer to the laboratory.

Sample preparation

For the purposes of this study, we pooled all five traps per site to create one aggregate sample per forest stand. Five of those aggregate samples were selected for presentation in this study. The selected samples for March, May and August were collected on the south of the levee (seasonally flooded habitats), whereas samples for April and June were collected north of the levee (non-flooded habitats). Sample preparation was kept to an absolute minimum. Macrolepidoptera, Pyraloidea and a few other non-macrolepidoptera (Limacodidae, Cossidae) were initially separated from the samples for manual species identification and later recombined to make up the total original sample. Manual species determination was performed by experts (FB) or under expert supervision (FB and KF). Large amounts of very large insects, for example, stag beetles (*Lucanus cervus*) and hornets (*Vespa crabro*), were removed except for a single individual per sample. Lepidoptera were not removed regardless of size or number.

Samples were dried at 50°C for 48 h. The dried samples were ground using an IKA Tube Mill 100 at 25,000 min^{-1} for 2 × 10 s in a 100 mL single use a grinding chamber. As the volume of all samples exceeded the maximum usable volume of the grinding chamber, we ground samples in multiple runs. Individual grounding runs were pooled in a 1 L flask (approximately 5–8 times sample volume) to allow for proper mixing of separate grinds. Ground samples were stored at −20°C before DNA extraction.

The amount of ground tissue used was standardised across all samples. DNA extraction was performed using the Qiagen DNA Blood & Tissue Kit. Digestion with proteinase K was performed at 56°C for 150 min. We used 1260 uL of buffer ATL and 140 uL proteinase solution for each sample, and the remaining extraction process was performed according to the kit's manual.

PCR amplification and sequencing

We tested two primer combinations for their suitability to detect OTUs. The selected primer combinations are fwhF2/fwhR2n (205 bp target region, EL2) and mICOIntF/Fol-degen-rev (313 bp target region, EL8), respectively. Those two primer combinations showed the highest OTU recovery rate in a comparison of 21 primer combinations using a mock community (Elbrecht et al., 2019). Primer combinations with different amplicon lengths were selected to assess the performance of shorter versus longer amplicons. All five monthly samples were sequenced using both primer combinations resulting in a total of 10 samples being analysed.

All Polymerase Chain Reaction (PCR) reactions were performed using the Thermo Scientific Phusion Green Hot Start II High-Fidelity PCR Master Mix. In the first step, the target region was amplified using the above-mentioned primers. PCR reactions were set up as follows: 12.5 uL Phusion Hot Start II Master Mix, 6.5 µL ddH2O, 1.25 µL of each primer and 1 µL DNA. The cycler program was as follows: 98°C for 30 s; 10 cycles of 98°C for 10 s, 67–58°C (decreasing for 1°C per cycle) for 30 s, and 72°C for 15 s; 30 cycles of 98°C for 10 s, 61°C for 30 s, and 72°C for 15 s; and 5 min at 72°C.

In the second step, the PCR product obtained from step one was amplified with fusion primers containing the original primer, four to seven base barcodes for sample discrimination, Illumina sequencing primers and adapters. The entire primer construct used in Elbrecht et al., 2019 was adopted for this study. See Table S1 for fusion primer sequences. PCR reactions were set up as follows: 12.5 µL Phusion Hot Start II Master Mix, 6.5 µL ddH2O, 1.25 µL of each primer and 1 µL DNA. The cycler program was as follows: 98°C for 30 s; 10 cycles of 98°C for 10 s, 67–58°C (decreasing for 1°C per cycle) for 30 s, and 72°C for 15 s; 9 cycles of 98°C for 10 s, 61°C for 30 s, and 72°C for 15 s; 5 min at 72°C; or 98°C for 30 s; 10 cycles of 98°C for 10 s, 67–58°C (decreasing for 1°C per cycle) for 30 s, and 72°C for 15 s; 20 cycles of 98°C for 10 s, 61°C for 30 s, 72°C for 15 s; 5 min at 72°C; or 98°C for 30 s; 19 cycles of 98°C for 10 s, 58°C for 30 s, 72°C for 15 s, and 72°C for 15 s; and 5 min at 72°C.

PCR products were purified using the Qiagen QiaQuick PCR purification kit according to the manufacturers' instructions. Samples were pooled, and samples with primer combination EL8 were introduced at a 10% higher concentration than EL2 to compensate for sequencing bias of the Illumina platform. Sequencing was performed on the Illumina MiSeq v3 600 paired end platform, and 5% PhiX was injected. The samples reported in this study were sequenced along with other samples.

Sequence processing

The Illumina bcl file was converted into fastq format using bcl2fastq provided by Illumina. Paired-end reads were assembled using PEAR (Zhang et al., 2013) with the minimum overlap set to 150 base pairs for EL2 and 200 base pairs for EL8 and a minimum PHRED base quality score of 30. Demultiplexing was done using demultiplex (Laros, 2021). Primers and tags were removed using the 'fastx_truncate' command, 'fastx_revcomp' was applied to all reversed sequences, and both are provided by usearch v 11 (Edgar, 2010). Assembled reads deviating from the expected length by more than one base pair were discarded. The number of reads per sample can vary considerably; to obtain a standardised sequencing depth, all samples were down-sampled to 100,000 and 10,000 reads, respectively. Downsampling was performed using the 'fastx_subsample' command provided by usearch.

Manually determined species set

The reference database to identify species contained in the nested calibration species set was constructed to include all species detected in the respective samples by manual determination (226 species, see Table S2). COI barcode sequences were obtained through the BOLD public data portal in October 2021 (Ratnasingham & Hebert, 2007). All public

sequences were used, and identical sequences were removed from the reference data using the 'fastx_uniques' command, part of usearch v11 (Edgar, 2010). Reference sequences were checked for errors using a sequence alignment. An NJ tree was calculated to identify potential erroneous sequences. Sequences less than 200 bp in length were excluded, corresponding to the read length of the shorter amplicon. We were able to obtain at least one full-length (658 bp) sequence for all species in our nested calibration set. The final reference database contained 6509 sequences. We used the free version of usearch v11 to perform an exhaustive search of the obtained sequence reads for each sample against the reference database using the 'usearch_local' command. Parameters were set as follows: minimum identity = 0.98, evalue = 10e-6, maxaccepts = 0, and maxrejects = 0. Sequence reads were matched to the reference database based on species names as practiced in manual identification. The output obtained from usearch was aggregated according to species names using a custom R script (supplement). Reads with a query coverage of less than 90% were removed and a minimum evalue threshold of 10e-100 was applied. The proportion of successfully detected species was then calculated and visualised. Sequencing depth was rarefied in 10,000 read steps from 100,000 reads to 0 to determine the necessary sequencing depth for future applications.

A size/weight regression approach was applied to estimate the mass of all 228 moth species in the nested calibration set. We weighed specimens of 75 species found in our samples (*n* per species between 1 and 5) and obtained wingspan data from the literature (Guariento et al., 2020). Wingspan and mass were log-transformed. Specimens used for size/weight regression were obtained from samples not part of this study. To achieve the best possible estimation of biomass, we split those 75 species into two categories based on robust and slender stature of adult moths. Hereby, 'robust' relates to moths with a broad thorax and abdomen, and thus a rather high wing loading, like typical representatives of European Noctuidae, Notodontidae, Sphingidae, Lymantriinae, Arctiini and alike. Representatives of Geometridae, Pyraloidea and some clades within Erebidae (e.g., Herminiinae, Hypeninae, Hypenodinae and Lithosiini) were scored as 'slender' moths. We used a linear model based on those 75 species in order to predict the mass of all 228 target species (see Table S3). To explore the influence of biomass and taxonomy on the level of superfamilies, we built a logistic general linear mixed model using 'glmer' with family set to 'binomial' using a 'logit' link function. Species detection (0 = not detected, 1 = detected) was the dependent variable, and taxonomy on the level of superfamilies was the independent variable. Biomass and the number of individuals sampled per species were included as random effects. The MuMin R package (Burnham & Anderson, 2002) was used to calculate Nagelkerke's (Nagelkerke, 1991) and Nakagawa & Schielzeth's R^2 (Nakagawa & Schielzeth, 2013). Effect sizes of fixed effects were calculated using the package effect size (Ben-Shachar et al., 2020).

To estimate the implications of incomplete species recovery on ecological characterisation of species assemblages, an Non-metric multidimensional scaling (NMDS) ordination using 'metaMDS' with Bray–Curtis distances, contained in the vegan R package (Oksanen et al., 2022), was performed. The full species assemblages per sampling site as well as the subsets recovered by primer combinations, and species assemblages recovered by EL2 and EL8 were treated as separate samples. All other settings were left on default.

Total OTU assignment

For species assignment of the total sequence reads to the manually determined set of species, we used BINs (Ratnasingham & Hebert, 2013) as provisional taxonomic units. BOLD was queried in October 2021 for sequences of all arthropods collected in the following countries in the order of number of sequences: Germany, Norway, Finland, Russia, Italy, France, Austria, United Kingdom, Belarus, Sweden, Switzerland, Czech Republic, Romania, Netherlands, Croatia, Slovakia, Slovenia, Belgium, Hungary, Denmark, Ukraine, Serbia, Estonia, Latvia, Ireland, Lithuania, Luxembourg, Moldova, Liechtenstein, San Marino and Monaco (no terrestrial arthropod sequences available). Major crustacean orders were removed to reduce the size of the reference data set. Identical sequences were removed from the reference data using the 'fastx_uniques' command, part of usearch v11. Sequences less than 200 bp in length were also excluded. The final reference database contained 312,643 sequences. No further quality checks were performed on the sequences.

We used the free version of usearch version 11 to perform an exhaustive search of the data against the database using the 'usearch_local' command. Parameters were set as follows: minimum identity = 0.98, evalue = 10e-6, maxaccepts = 0 and maxrejects = 0. Some usearch runs were performed on the Vienna Scientific Cluster (VSC3) computing cluster. The output obtained from usearch was aggregated according to BIN:URI using a custom R script. Reads with a query coverage of less than 90% were removed and an e-value threshold of 10e-100 was applied. The obtained hits were manually inspected for matches to potential erroneous database sequences; no such cases were detected.

RESULTS

All 10 samples combined amounted to 2,165,809 paired reads conforming to our quality criteria. Raw sequence reads are available at https://phaidra.univie.ac.at/o:1646993.

Nested calibration

In our nested calibration set of species comprised of our selection of manually determined species, a total of 228 species of Lepidoptera were detected by means of morphological identification. Metabarcoding analyses recovered 158 of those species in total across both primer combinations.

Results indicate different performances of the two primer combinations. The shorter amplicon, EL2, had in general a higher detection rate compared with EL8 and showed a more consistent pattern across the five samples (Figure 1a). Detection rates ranged from 0.46 to 0.69. Sub-sampling to 10,000 reads leads to a notable decrease in detection

rates to 0.31–0.44. When singleton species were excluded (Figure 1c), detection rates increased to 0.65–0.72 for 100,000 reads and 0.38–0.63 for 10,000 reads, respectively. Species detection rates for EL8 ranged from 0.37 to 0.72 (Figure 1b), and sub-sampling to 10,000 reads resulted in a decrease to 0.13–0.44 (Figure 1d). When excluding singleton species detection, rates increased to 0.47–0.76 for 100,000 reads and to 0.14–0.53 for 10,000 reads (Figure 1d). See Table S2 for the detailed results on species level.

Rarefied species detection rates (Figure 2) reveal that depending on the sample, the proportion of species recovered with only a single read ranged between 0.04 and 0.25. While performance varied between samples, the overall pattern was very similar for both primer combinations.

An NMDS ordination (Figure 3) of the full moth community data compared to the subsets recovered with both primer combinations, respectively, showed that the overall pattern among the five sampling months was preserved with all three species sets. The first ordination axis largely depicts seasonal progress in moth assemblages from early spring (March) to summer, whereas summer samples from May, June and August are separated from another along the second ordination axis, but not in temporal sequence. Samples taken in May and June showed a wider spread when the full sample is considered in comparison to the subsets recovered by primer combinations EL2 and EL8. Results obtained from EL2 and EL8 clustered closer to each other than to the full, manually identified, sample with April being the only exception.

Taxon-specific performance summarised across all samples differed among the four major lepidopteran superfamilies considered in our nested calibration (Figure 4). Both primer combinations showed a very similar pattern with high detection rates for Noctuoidea and considerably lower detection rates for Pyraloidea and Geometroidea. Downsampling to 10,000 reads further reduced the detection rate by a considerable margin. Lasiocampoidea, Cossoidea and Zygaenoidea with only a single representative species each in our samples were not considered in this comparison. The single lasiocampoid species was never detected through metabarcoding despite its large size while the representatives of the other two clades were detected at least once.

Results obtained from logistic regression of species biomass and taxonomy versus detectability indicated that while body mass was a strong predictor for detectability (standard regression coefficient = 0.66, $p = 0.00045$), taxonomy on superfamily level was another significant predictor (standard regression coefficient = 0.32, $p = 0.04576$). Overall model fit was assessed using Nagelkerke's

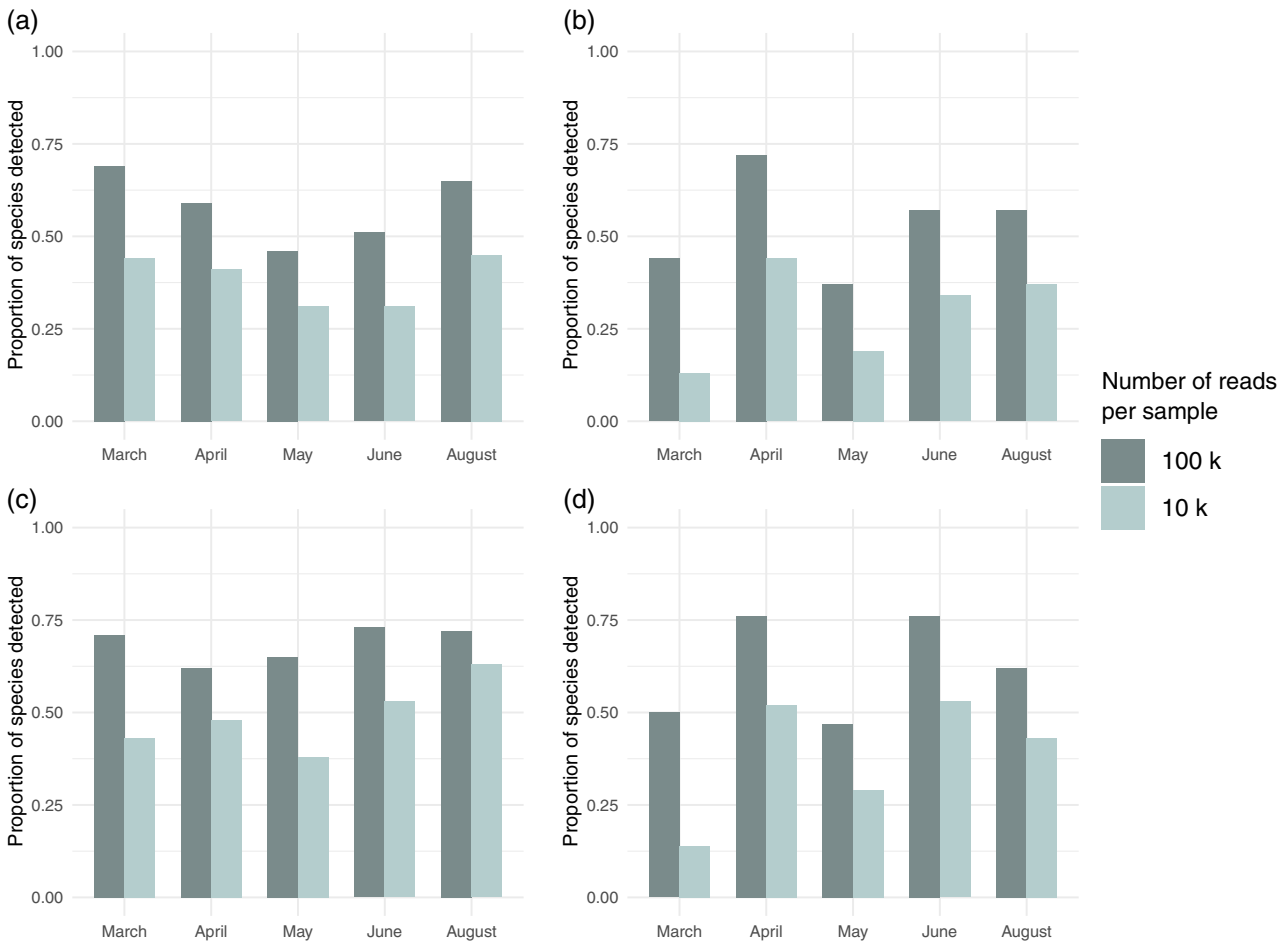

FIGURE 1 Detection rates of moth species by metabarcoding, relative to morphological species sorting, in the nested calibration set. (a) Primer combination EL2, (b) primer combination EL8, (c) primer EL2 without singleton species, and (d) EL8 without singleton species. Detection rates for 100,000 and 10,000 reads per bulk sample are indicated.

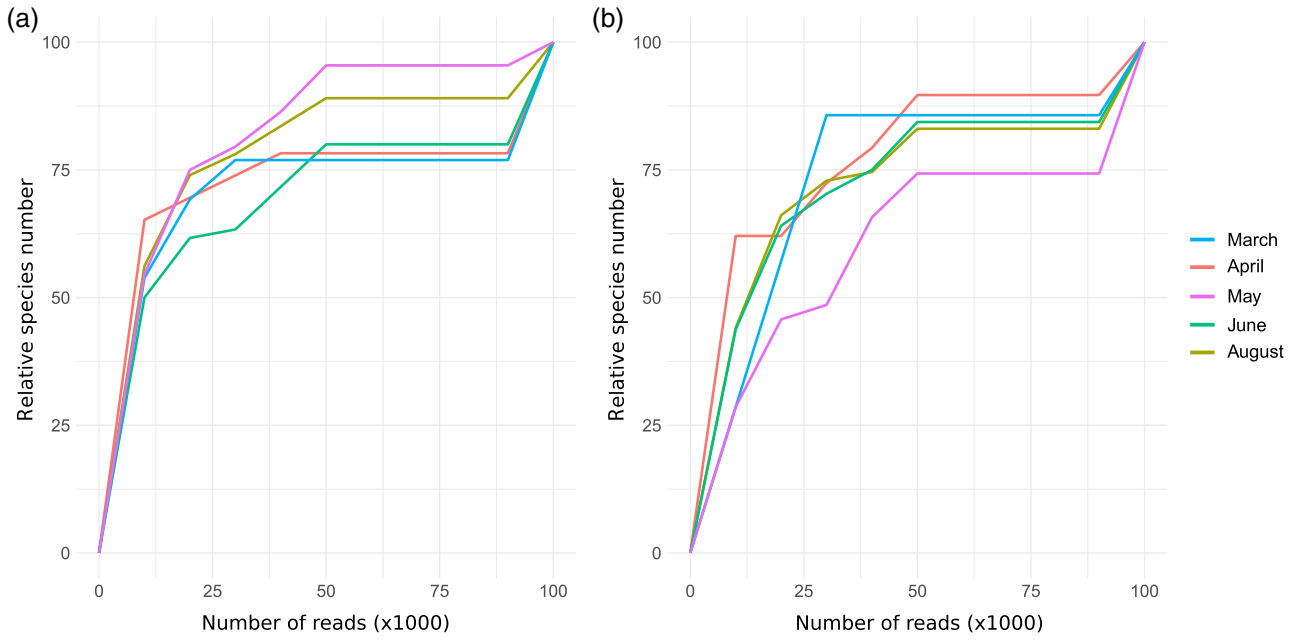

FIGURE 2 Rarefied species detection rates for 0 to 100,000 reads in steps of 10,000 reads, (a) primer combination EL2 and (b) primer combination EL8.

adjusted R^2 ($R^2 = 0.1516557$) and Nakagawa & Schielzeth's R^2 ($R^2m = 0.09026883$, $R^2c = 0.1372625$). These results show that smaller species are less likely to be detected, and this effect is also visually apparent in a histogram of species detection in relation to biomass (Figure 5).

Within the manually identified set of species, we found 22 cases of 20 different species where the DNA meta-barcode data showed the presence of species not detected during morphological determination of the samples.

Total OTU assignment

The total number of arthropod BINs detected in the bulk light trap samples ranged from 24 to 131 per sample for EL2 and between 13 and 148 BINs for EL8 (Figure 6a,b). The number of detected BINs increased from spring to summer, peaking in June. A notable exception was the sample taken in May, where EL8 detected only one more BIN compared with April. Lepidoptera was the most detected order making up more than 50% of OTUs in each sample. Diptera and Coleoptera were the second and third most species-rich orders, respectively. Other orders were present in low quantities. Additional BINs of Lepidoptera not covered by our manually determined set of species, so-called 'Microlepidoptera', for example, of the families Gracillariidae and Tortricidae, were detected. This way, we were able to detect 40 additional BINs for EL2 and 43 additional BINs for EL8. See Table S4 for detailed results.

Both EL2 and EL8 detected representatives of a total of 11 arthropod orders (Figure 7). All orders except for Blattodea, Entomobryomorpha, Ixodida and Mesostigmata were detected by both primer combinations. The unusual presence of Odonata was confirmed by visual observation prior to sample processing. While winged insects made up the vast majority of OTUs in the light trap samples, also some representatives of flightless arthropods were detected, such as mites, ticks and springtails, albeit at very low numbers.

DISCUSSION

Species detection (manually identified species set)

In accordance with the results obtained by Elbrecht et al. (2019), primer combination EL2 consistently showed the best performance in recovering species detected by morphological identification. EL8 performed slightly worse than EL2. This slight difference can likely be attributed to the bias towards short reads inherent to the Illumina platform (Schirmer et al., 2015). Down-sampling to 10,000 reads reduced the OTU detection rate to well below 0.5 in most cases, and 10,000 reads can, therefore, be considered insufficient to achieve a satisfactory OTU coverage. Rarefaction of sequence reads showed that species detection rate started to drop off at around 40,000 reads. Reduced detection rates for the sample taken in May might indicate suboptimal sample homogeneity in this case. The fact that removing singleton species eliminated this marked difference corroborates this suspicion. Tissue from single individuals is more likely to improperly mix into a bulk sample than is the case with two or more individuals present.

Singleton species showed lower detection rates compared with more abundant species (Figure 1). This is not unexpected as species present with only a single individual are much more likely to be affected by suboptimal sample homogeneity. Furthermore, small-sized

Ecological Entomology

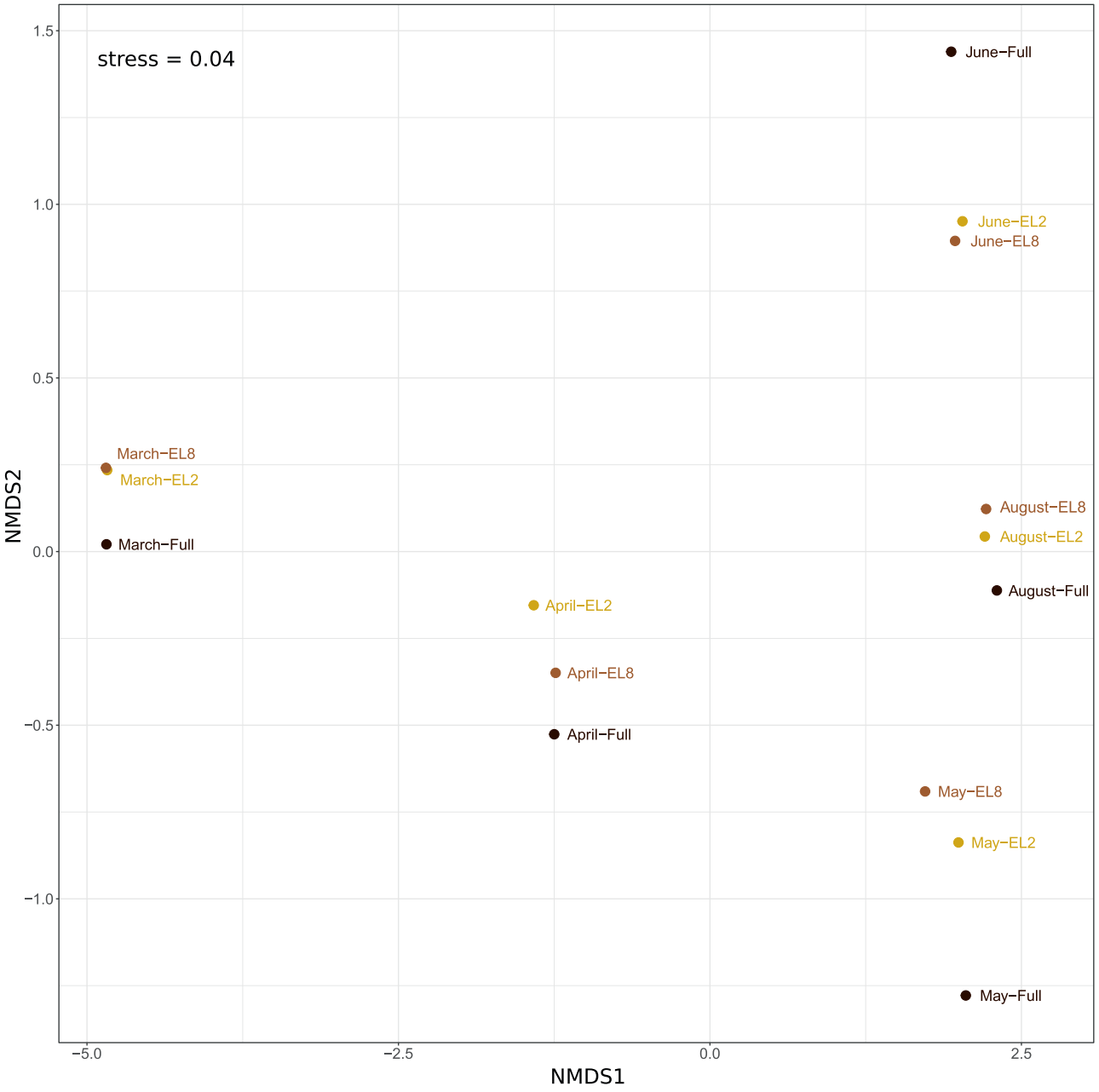

FIGURE 3 NMDS ordination of the full community as obtained by manual species determination compared with the subsets recovered by metabarcoding of bulk samples using primer combinations EL2 and EL8.

species will have an even lower chance of detection if only a single individual is present due to the dilution of their DNA past a detectable level. The importance of singleton species for community ecology and related fields is well explored in tropical systems (Coddington et al., 2009) but can also be relevant in temperate systems (Matos da Costa & Sielezniew, 2023). The latter study on light-trap samples of moths in eastern central Europe demonstrated that bias against singleton species can substantially change conclusions based on the data. Truxa and Fiedler (2016) found that considerable information on community structure can be gained from rare species in the very same study system in eastern Austria where the samples for this study were collected. In the present study, the effect caused by incompletely sampled assemblages was not apparent. An NMDS ordination comparing the manually determined set of species against assemblages recovered by metabarcoding showed that the overall pattern was preserved in both primer combinations in spite of low detection rates in some samples.

Benchmarks for detection rates in real-world samples are almost non-existent. Elbrecht et al. (2017) took an approach similar to this present study in a limnetic system, although taxonomic units above species level were used there for a considerable proportion of OTUs. Average detection rates ranged from 0.6 to 0.7 depending on the primer combination. Later on, Elbrecht et al. (2019) sequenced a single Malaise trap sample with 21 primer different combinations. This

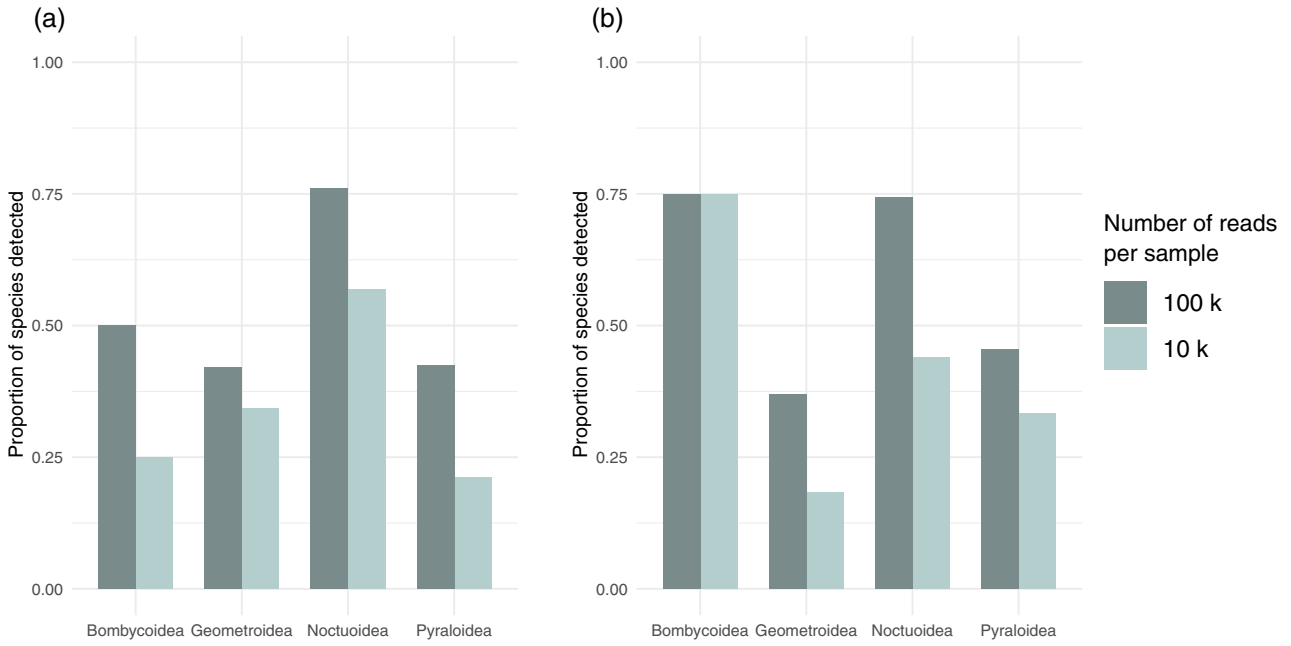

FIGURE 4 Detection rates of moths in metabarcoding of bulk samples by superfamily. (a) Primer combination EL2 and (b) primer combination EL8. Sample sizes (i.e., morphologically identified species numbers) are as follows: Bombycoidea: 4, Geometroidea: 76, Noctuoidea: 109 and Pyraloidea: 33. Three superfamilies represented by only a single species are not shown (see text).

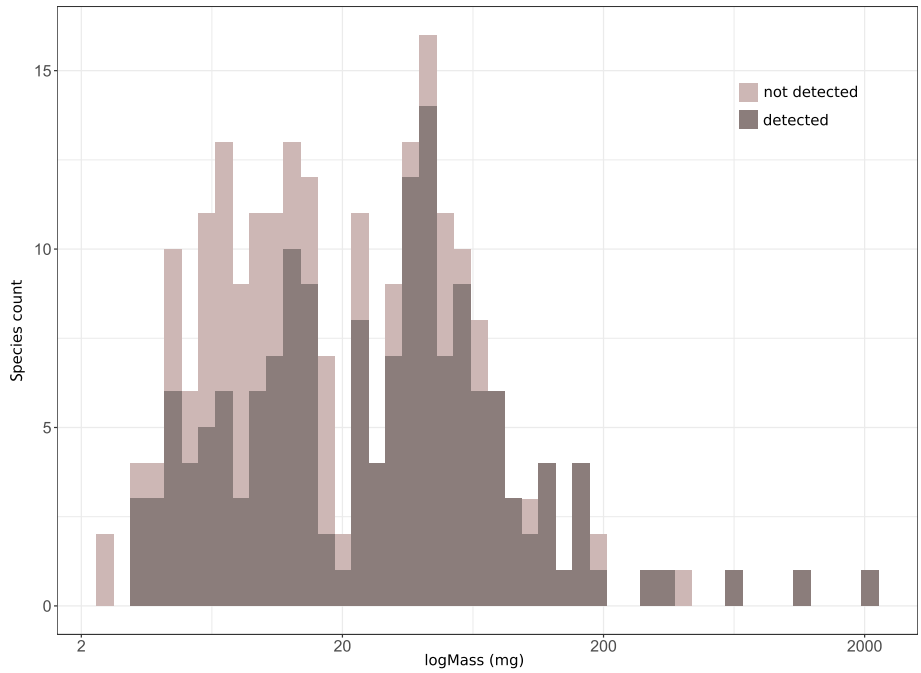

FIGURE 5 Detection of species in relation to their body mass. X-axis is in log scale and labels are in real space. The gap at 20 mg closely corresponds to the distinction between the slender and robust species in our size/weight regression model.

sample can, therefore, be considered to be unusually well explored even though no manual determination of species was performed. The best performing primer combination (EL2) recovered 911 (70.3%) out of a total of 1295 OTUs detected across all primer combinations. Hence, the level of detection obtained in those studies is very similar to the level observed in our results.

Our results revealed a pronounced bias against smaller moth species. This was to be expected and cannot be fully alleviated without either using very high sequencing depths, thereby massively increasing sequencing cost or by employing elaborate (and resource demanding) sample fractionation schemes as described by Elbrecht, Peinert, and Leese (2017).

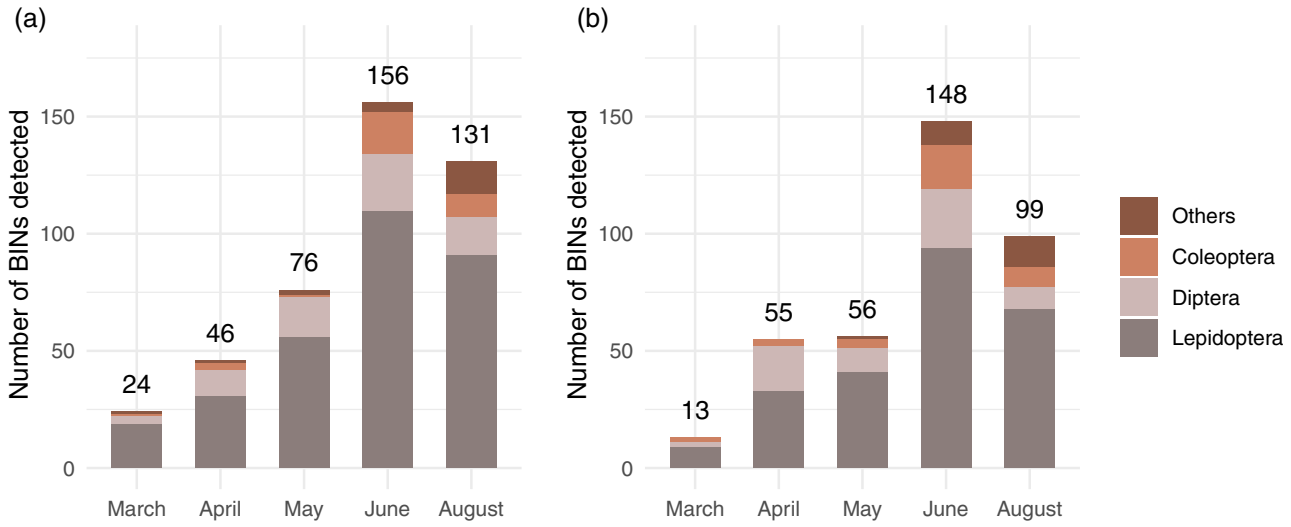

FIGURE 6 Total OTU assignment per light trap bulk sample with major insect orders indicated. (a) Primer EL2 and (b) primer EL8. Numbers above bars indicate the total number of BINs detected.

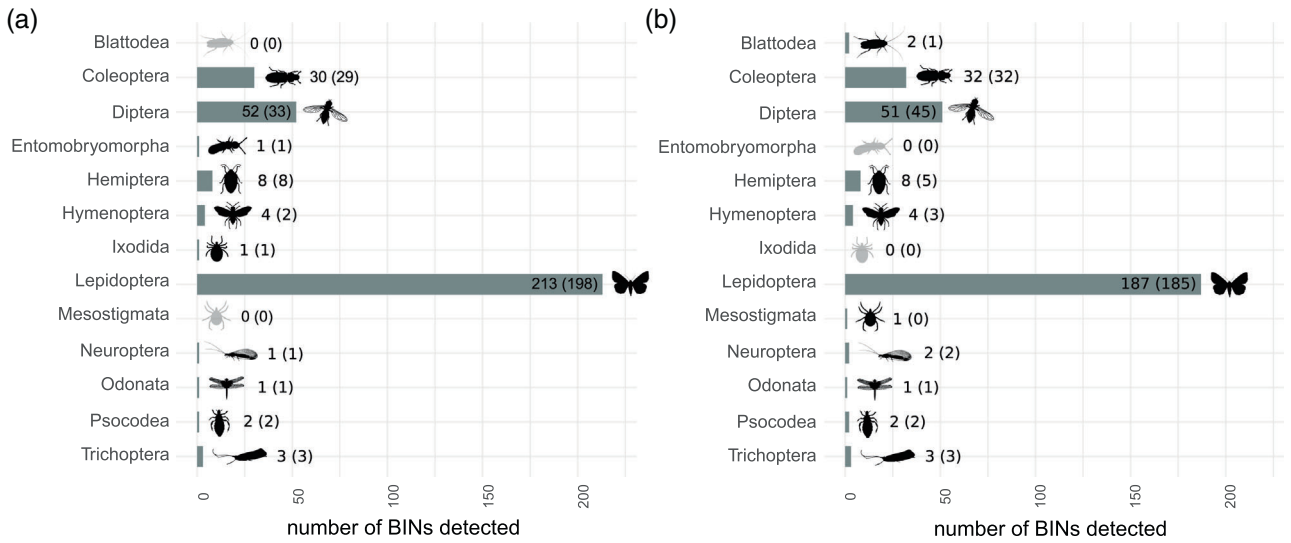

FIGURE 7 Frequency of all arthropod orders detected in total OTU assignment. (a) Primer combination EL2 and (b) primer combination EL8. Numbers indicate the number of detected BINs, and the numbers in bracket indicate BINs with species-level names. Silhouettes were obtained from PhyloPic.com; all images are in the public domain.

In addition, we were able to detect taxonomic bias on the level of superfamilies. Such a bias can easily be introduced by taxon-dependent primer affinity. Although anecdotal reports of numerous taxonomic biases in commonly used insect primers abound, no rigorous study has ever been performed on that subject in Lepidoptera. Combined with the observed bias against smaller species, this is likely to put detection of Geometroidea and Pyraloidea at a considerable disadvantage compared with Noctuoidea. Even though in our case the influence on overall community ecological patterns was minor (Figure 3), the effect on detection rates was apparent (Figure 4). Some conclusions drawn from metabarcoding data, for instance, the recommendation of conservation measures, can be very sensitive to taxonomic bias. Taxon-specific comparisons of different primer combinations would be required to discern those effects in detail. At any rate, the observation that two species-rich clades that make up a sizeable fraction of temperate-zone forest moth assemblages suffered from low detection rates is a source of concern, when it comes to inferences about the richness of local communities, for example, along ecological gradients. In addition, it is important to note that our manually identified species set contained mostly medium-to-large-sized moths in relation to the full body size spectrum seen in Lepidoptera. Including more clades of 'Microlepidoptera' (e.g., Gelechioidea, Tineoidea, Tortricoidea) would almost certainly result in even lower detection rates.

Apart from the non-detection of a range of moth species that were definitively part of the assemblages (see above), a few species

that belonged to the target groups of morphological sorting were detected through metabarcoding but apparently were not recorded during morphology-based sample analysis. These observed discrepancies between morphological determination and results obtained from metabarcoding can be split into three categories. In seven cases, we are confident that this can be attributed to instances of 'cryptic diversity' where closely related and morphologically similar species were confused or not recognised as distinct on the grounds of external wing patterns in larger samples of similar (and in some instances worn) individuals. In eight other cases, however, we are uncertain as to why those species were detected in the samples. It is conceivable that these species were misidentified or overlooked, but we consider sample contamination as an equal possibility. In another seven cases, we are confident that sample contamination has occurred, because it is not conceivable that these easy-to-identify species might have gone unnoticed through morphological sample analysis. There were a number of steps in our processing pipeline where contamination could have been introduced. Already in the field, contamination could have been introduced through the trap vanes or sampling containers. In the lab, cross contamination was possible while drying the samples as several samples had to be dried at once in the same drying cabinet due to time constraints. Contamination could also have occurred during DNA extraction and at both PCR stages. Specimens of all additionally detected species were collected and processed in our lab within 12 months of processing the samples in this study. Cross-contamination in our lab is, therefore, a possibility.

Contamination is frequently observed in metabarcoding studies, usually as an incidental observation that is not explored any further. Sources of contamination and strategies to minimise sample contamination, on the other hand, are well explored (Liu et al., 2020). Precise data on the prevalence of contamination in metabarcoding approaches are virtually non-existent as most studies are unable to quantify contamination due to the lack of prior knowledge on OTU composition. The level of contamination observed in this study is likely of no concern in biodiversity monitoring applications but might be of concern in species discovery applications.

Total OTU assignment

Results from the total OTU assignment showed an expected composition of insect orders attracted to light traps and are in accordance with visual observations made during emptying of the traps and sample preparation. Representatives of a number of orders not usually attracted to light were detected at low numbers. Those orders likely entered the trap by chance, for example, via the trees, the traps were suspended from. Mites (Mesostigmata) might have been introduced as ectoparasites on winged insects. When comparing our results obtained from automatic light traps with other sampling methods, it is important to note that automatic light traps tend to be biased against small-sized moths (Axmacher & Fiedler, 2004). There is good evidence that this bias is caused by different settling behaviours of large- versus small-sized moths at the trap (Wölfling et al., 2016). Samples obtained by different means, for example, manual collection on a sheet, are, therefore, likely to show a different composition especially at the lower end of the body-size spectrum. The diversity of non-Lepidopteran OTUs and Lepidopteran OTUs not part of the nested calibration ('Microlepidoptera') showed that detection of the species in the nested calibration set was possible while being embedded in a complex sample. In addition, the metabarcoding approach facilitated the analysis of additional 'Microlepidoptera' assemblages where morphological identification would have been very time-consuming, would have called for routine microscopic preparation of genitalic slides and might sometimes require the consultation of taxonomic specialists for the respective family. Given that reference sequences of an increasing number of Central European 'Microlepidoptera' are now available in the BOLD database, for these insects, metabarcoding promises to be a cost-effective means of using the full potential of light trap samples for inferences on biodiversity.

CONCLUSIONS

Based on our results, we recommend primer combination EL2 for future studies on samples of flying insects. The shorter length of EL2 enables the use of cheaper sequencing options saving approximately 30% of sequencing costs. As suggested by the results of previous studies (e.g., Elbrecht et al., 2019; Elbrecht & Steinke, 2019), 100,000 reads per sample turned out to be a good balance between the cost and quality of results. In regard to sequencing depth, our data show a marked drop in species detection rates only at 40,000 or 50,000 reads (Figure 2). This indicates that for samples comparable with the ones in this study, sequencing depths substantially below 100,000 can still yield acceptable results.

Until there are good baseline data on taxon recovery and taxonomic biases, we suggest to apply a nested calibration approach whenever possible. This approach can provide a very accurate quality control for metabarcoding studies and possibly predict overall detection rates. Ideally, a subset of species representing a large range of body sizes should be selected and manually identified. Within the same Illumina run, it is likely that applying nested calibration to only one or a few samples will be sufficient to get a useful estimate of overall detection rates. We recognise that the required effort will still not be feasible in all cases. In future studies, it would also be desirable to investigate how large the calibrated subset needs to be in order to serve as a predictor for overall detection rates. A significant degree of taxonomic bias was detected on the level of moth superfamilies, which might distort relationships between the samples in regard to species richness, whereas patterns of species composition remained largely unchanged. More in-depth investigation of those biases is required.

Further studies are required to determine whether the OTU recovery rates observed in this study are typical for similar metabarcoding applications. Strategies to improve OTU recovery rates can involve the use of order-specific PCR primers (e.g., Šigut et al., 2017). Alternatively, 16S rRNA metabarcoding has been demonstrated to be a valuable supplement to COI metabarcoding (Elbrecht et al., 2016). However, the lack of available reference

sequences severely limits the use of 16S rRNA metabarcoding in practice. As of August 28rd 2023, 16S rRNA sequences for a mere 1373 species of Lepidoptera are available on Genbank, and only 9 of the 228 manually identified species in this study have publicly available 16S rRNA sequences. All approaches to improve detection rates come with trade-offs in cost and time required for sequencing. For some applications, perhaps in tropical systems with very high species numbers and comparatively little prior knowledge of species composition, adopting one or more of those proposed strategies might be indispensable. We do, however, think that once we have a good understanding of expected detection rates single marker, single amplicon approaches without any pre-sorting/fractionation can still be a valuable and cost-efficient way to study biodiversity. Using the nested-calibration approach might in some cases be more suitable as opposed to the substantial increase in cost and time required for the alternatives proposed above.

AUTHOR CONTRIBUTIONS

Patrick Strutzenberger: Conceptualization; investigation; writing – original draft; methodology; validation; visualization; writing – review and editing; formal analysis; supervision; data curation. **Brigitte Gottsberger:** Methodology; writing – review and editing; data curation; resources. **Florian Bodner:** Data curation; writing – review and editing; resources. **Fabian Bartusel:** Writing – review and editing; resources. **Daniel Jerga:** Resources; writing – review and editing. **Konrad Fiedler:** Methodology; validation; writing – review and editing; writing – original draft; investigation; formal analysis; resources; data curation.

ACKNOWLEDGEMENTS

We are grateful to the staff of the National Park Donau-Auen, in particular Christian Baumgartner and Karoline Zsak, for granting access to the study sites and for manifold logistic support during field work. Computation was in part performed on the Vienna Scientific Cluster (VSC). Illumina sequencing was performed at the VBCF sequencing service. Sequencing cost was subsidised by the Faculty of Life Sciences, University of Vienna.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Phaidra at https://phaidra.univie.ac.at/o:1646993.

ORCID

Patrick Strutzenberger https://orcid.org/0000-0003-3278-5693
Brigitte Gottsberger https://orcid.org/0000-0002-6317-306X
Konrad Fiedler https://orcid.org/0000-0002-4789-3753

REFERENCES

Abdelfattah, A., Malacrinò, A., Wisniewski, M., Cacciola, S.O. & Schena, L. (2018) Metabarcoding: a powerful tool to investigate microbial communities and shape future plant protection strategies. *Biological Control*, 120, 1–10. Available from: https://doi.org/10.1016/j.biocontrol.2017.07.009

Axmacher, J.C. & Fiedler, K. (2004) Manual versus automatic moth sampling at equal light sources—a comparison of catches from Mt. Kilimanjaro. *Journal of the Lepidopterists' Society*, 58(4), 196–202.

Bartusel, F. (2021) Possible change in species diversity and species composition of nocturnal Lepidoptera in the Danubian floodplain after over ten years. MSc thesis, University of Vienna. https://doi.org/10.25365/thesis.70990

Ben-Shachar, M.S., Luedecke, D. & Makowski, D. (2020) Effectsize: estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56), 2815. Available from: https://doi.org/10.21105/joss.02815

Braukmann, T.W.A., Ivanova, N.V., Prosser, S.W.J., Elbrecht, V., Steinke, D., Ratnasingham, S. et al. (2019) Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, 19, 711–727. Available from: https://doi.org/10.1111/1755-0998.13008

Brehm, G. (2017) A new LED lamp for the collection of nocturnal lepidoptera and a spectral comparison of light-trapping lamps. *Nota Lepidopterologica*, 40(1), 87–108. Available from: https://doi.org/10.3897/nl.40.11887

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edition. New York: Springer-Verlag.

Coddington, J.A., Agnarsson, I., Miller, J.A., Kuntner, M. & Hormiga, G. (2009) Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. *Journal of Animal Ecology*, 78(3), 573–584.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460–2461. Available from: https://doi.org/10.1093/bioinformatics/btq461

Elbrecht, V., Bourlat, S.J., Hörren, T., Lindner, A., Mordente, A., Noll, N.W. et al. (2021) Pooling size sorted Malaise trap fractions to maximize taxon recovery with metabarcoding. *PeerJ*, 9, e12177. Available from: https://doi.org/10.7717/peerj.12177

Elbrecht, V., Braukmann, T.W.A., Ivanova, N.V., Prosser, S.W.J., Hajibabaei, M., Wright, M. et al. (2019) Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7, e7745. Available from: https://doi.org/10.7717/peerj.7745

Elbrecht, V., Peinert, B. & Leese, F. (2017) Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, 7, 6918–6926. Available from: https://doi.org/10.1002/ece3.3192

Elbrecht, V. & Steinke, D. (2019) Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. *Freshwater Biology*, 64, 380–387. Available from: https://doi.org/10.1111/fwb.13220

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J. et al. (2016) Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, 4, e1966. Available from: https://doi.org/10.7717/peerj.1966

Elbrecht, V., Vamos, E.E., Meissner, K., Aroviita, J. & Leese, F. (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, 8, 1265–1275. Available from: https://doi.org/10.1111/2041-210X.12789

Giangrande, A. (2003) Biodiversity, conservation, and the 'Taxonomic impediment'. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 13, 451–459. Available from: https://doi.org/10.1002/aqc.584

Guariento, E., Strutzenberger, P., Truxa, C. & Fiedler, K. (2020) The trinity of ecological contrasts: a case study on rich insect assemblages by means of species, functional and phylogenetic diversity measures. *BMC Ecology*, 20(1), 1–14. Available from: https://doi.org/10.1186/s12898-020-00298-3

Hausmann, A., Segerer, A.H., Greifenstein, T., Knubben, J., Morinière, J., Bozicevic, V. et al. (2020) Toward a standardized quantitative and qualitative insect monitoring scheme. *Ecology and Evolution*, 10, 4009–4020. Available from: https://doi.org/10.1002/ece3.6166

Laros, J.F.J. (2021) Demultiplex. Available from: https://github.com/jfjlaros/demultiplex

Lewis, R.J., de Bello, F., Bennett, J.A., Fibich, P., Finerty, G.E., Götzenberger, L. et al. (2017) Applying the dark diversity concept to nature conservation. *Conservation Biology*, 31(1), 40–47.

Liu, M., Clarke, L.J., Baker, S.C., Jordan, G.J. & Burridge, C.P. (2020) A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, 45, 373–385. Available from: https://doi.org/10.1111/een.12831

Mata, V.A., Ferreira, S., Campos, R.M., da Silva, L.P., Veríssimo, J., Corley, M.F.V. et al. (2021) Efficient assessment of nocturnal flying insect communities by combining automatic light traps and DNA metabarcoding. *Environmental DNA*, 3, 398–408. Available from: https://doi.org/10.1002/edn3.125

Matos da Costa, J. & Sielezniew, M. (2023) The contribution of singletons and doubletons captured using weak light heath traps for the analysis of the Macroheteroceran assemblages in forest biotopes. *Diversity*, 15, 508. Available from: https://doi.org/10.3390/d15040508

Nagelkerke, N.J.D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4, 133–142. Available from: https://doi.org/10.1111/j.2041-210x.2012.00261.x

Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P. et al. (2022) _vegan: Community Ecology Package_. R package version 2.6-2. Available from: https://CRAN.R-project.org/package=vegan

Page, R.D. (2016) DNA barcoding and taxonomy: dark taxa and dark texts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702), 20150334.

Piper, A.M., Batovska, J., Cogan, N.O.I., Weiss, J., Cunningham, J.P., Rodoni, B.C. et al. (2019) Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, 8(8), giz092. Available from: https://doi.org/10.1093/gigascience/giz092

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: the barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7, 355–364. Available from: https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S. & Hebert, P.D.N. (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One*, 8(7), e66213. Available from: https://doi.org/10.1371/journal.pone.0066213

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T. & Quince, C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6), e37. Available from: https://doi.org/10.1093/nar/gku1341

Semenov, M.V. (2021) Metabarcoding and metagenomics in soil ecology research: achievements, challenges, and prospects. *Biology Bulletin Reviews*, 11, 40–53. Available from: https://doi.org/10.1134/S2079086421010084

Shelton, A.O., Gold, Z.J., Jensen, A.J., D'Agnese, E., Andruszkiewicz, A.E., Van Cise, A. et al. (2023) Toward quantitative metabarcoding. *Ecology*, 104(2), e3906. Available from: https://doi.org/10.1002/ecy.3906

Šigut, M., Kostovčík, M., Šigutová, H., Hulcr, J., Drozd, P. & Hrček, J. (2017) Performance of DNA metabarcoding, standard barcoding, and morphological approach in the identification of host–parasitoid interactions. *PLoS One*, 12(12), e0187803. Available from: https://doi.org/10.1371/journal.pone.0187803

Truxa, C. & Fiedler, K. (2012) Down in the flood? How moth communities are shaped in temperate floodplain forests. *Insect Conservation and Diversity*, 5(5), 389–397. Available from: https://doi.org/10.1111/j.1752-4598.2011.00177.x

Truxa, C. & Fiedler, K. (2016) Massive structural redundancies in species composition patterns of floodplain forest moths. *Ecography*, 39(3), 253–260.

van der Loos, L.M. & Nijland, R. (2021) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecololgy*, 30, 3270–3288. Available from: https://doi.org/10.1111/mec.15592

Wölfling, M., Becker, M.C., Uhl, B., Traub, A. & Fiedler, K. (2016) How differences in the settling behaviour of moths (Lepidoptera) may contribute to sampling bias when using automated light traps. *European Journal of Entomology*, 113, 502–506.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. et al. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3, 613–623. Available from: https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2013) PEAR: a fast and accurate Illumina paired-end reAd mergeR. *Bioinformatics*, 30, 614–620. Available from: https://doi.org/10.1093/bioinformatics/btt593

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

TABLE S1. Stage one PCR primers, fusion primers and barcodes used for Illumina sequencing are indicated for each sample in this study.

TABLE S2. Species/site matrix for each month sampled. Detected species are indicated for morphological identification (no suffix), primer combination EL2 (suffix -EL2), and primer combination EL8 (suffix -EL8). For species detected in morphological identification the number of individuals is indicated. For meta-barcode samples the number of obtained reads is indicated for each species.

TABLE S3. Raw data used for size/weight regression. Weights for each species and assignment to the 'slender' and 'robust' category are indicated. Detection in meta-barcode samples is coded as presence (1), absence (0). 'Frequency' refers to the number of individuals detected by morphological identification. Taxonomy is given on the level of superfamilies, GEO (Geometroidea), NOC (Noctuoidea), PYR (Pyraloidea), BOM (Bombycoidea), OTH (others).

TABLE S4. Raw results of the 'total OTU assignment' approach. Sequence matches are indicated for each sample and amplicon along with their respective e-values and other statistics.

How to cite this article: Strutzenberger, P., Gottsberger, B., Bodner, F., Bartusel, F., Jerga, D. & Fiedler, K. (2023) DNA metabarcoding of light trap samples vs. morphological species identification. *Ecological Entomology*, 1–12. Available from: https://doi.org/10.1111/een.13297