Upscaling biodiversity monitoring: Metabarcoding estimates 31,846 insect species from Malaise traps across Germany

Dominik Buchner[1*], James S. Sinclair[2], Manfred Ayasse[3], Arne Beermann[1,4], Jörn Buse[5], Frank Dziock[6], Julian Enss[4,7,8], Mark Frenzel[9], , Thomas Hörren[7], Yuanheng Li[1], Michael T. Monaghan[12,13], Carsten Morkel[14], Jörg Müller[15,16], Steffen U. Pauls[17,18,19], Ronny Richter[20,21], Tobias Scharnweber[22], Martin Sorg[7], Stefan Stoll[8,23], Sönke Twietmeyer[24], Wolfgang W. Weisser[25], Benedikt Wiggering[26], Martin Wilmking[22], Gerhard Zotz[27], Mark O. Gessner[10,11], Peter Haase [2,4,8$], Florian Leese[1,4$]

*Corresponding author; dominik.buchner@uni-due.de

1 Aquatic Ecosystem Research, University of Duisburg Essen, Universitätsstraße 5, 45141 Essen, Germany

2 Senckenberg Research Institute and Natural History Museum Frankfurt, Clamecystraße 12, 63571 Gelnhausen, Germany

3 Institute of Evolutionary Ecology and Conservation Genomics, University of Ulm, Albert-Einstein Allee 11, 89069 Ulm, Germany

4 Centre for Water and Environmental Research, Universitätsstraße 2, 45141 Essen, Germany

5 Black Forest National Park, Kniebisstraße 67, 72250 Freudenstadt, Germany

6 University of Applied Sciences HTW Dresden, Pillnitzer Platz 2, 01326 Dresden

7 Entomological Society Krefeld, Magdeburger Straße 38-40, 47800 Krefeld, Germany

8 Faculty of Biology, University of Duisburg Essen, Universitätsstraße 5, 45141 Essen, Germany

9 Helmholtz Centre for Environmental Research – UFZ, Department of Community Ecology, Theodor-Lieser-Straße 4, 06120 Halle, Germany

10 Department of Plankton and Microbial Ecology, Leibniz Institute of Freshwater Ecology & Inland Fisheries (IGB), Zur alten Fischerhütte 2, 16775 Stechlin, Germany

11 Department of Ecology, Berlin Institute of Technology (TU Berlin), Ernst-Reuter-Platz 1, 10587 Berlin, Germany

12 Department of Evolutionary and Integrative Ecology, Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany

13 Institute of Biology, Freie Universität Berlin, Königin-Luise-Straße 2/4, Gartenhaus 14195 Berlin, Germany

14 Kellerwald-Edersee National Park, Laustraße 8, 34537 Bad Wildungen, Germany

15 Field Station Fabrikschleichach, Department of Animal Ecology and Tropical Biology, Julius-Maximilians-Universität Würzburg, 97070 Würzburg, Germany

16 Bavarian Forest National Park, 94481 Grafenau, Germany

17 Senckenberg Research Institute and Natural History Museum Frankfurt, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

18 LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

19 Institute for Insect Biotechnology, Justus-Liebig-University Gießen, Ludwigstraße 23, 35390 Gießen, Germany

20 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany

21 Systematic Botany and Functional Biodiversity, Institute for Biology, Leipzig University, Johannisallee 21, 04103, Leipzig, Germany

22 Institute for Botany and Landscape Ecology, Greifswald University, Soldmannstraße 15, 17487 Greifswald, Germany

23 University of Applied Sciences Trier, Umwelt-Campus Birkenfeld, Campusallee, 55768 Hoppstädten-Weiersbach

24 Eifel National Park, 53937 Schleiden-Gemünd, Germany

25 Chair for Terrestrial Ecology, Department of Life Science Systems, School of Life Sciences, Technische Universität München, Hans-Carl-von-Carlowitz-Platz 2, 85354 Freising-Weihenstephan, Germany

26 Lower Saxon Wadden Sea National Park Authority, Virchowstraße 1, 26382 Wilhelmshaven, Germany

27 Institute of Biology and Environmental Sciences, Carl von Ossietzky Universität Oldenburg, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany

$ equally contributing senior authors

Abstract

Mitigating ongoing losses of insects and their key functions (e.g., pollination) requires accurately tracking large-scale and long-term community changes. However, doing so has been notoriously hindered by uniquely high insect species diversity that requires prohibitively high investments of time, funding, and taxonomic expertise. Here, we show that these concerns can be addressed through a comprehensive, scalable and cost-efficient DNA metabarcoding workflow. We use 1,815 samples from 75 Malaise traps across Germany from 2019 and 2020 to demonstrate how metabarcoding can be incorporated into large-scale insect monitoring networks for less than 50 € per sample, including supplies, labor and maintenance. With on average 1.4M sequence reads per sample we uncovered 10,803 validated insect species, of which 83.9% were represented by a single OTU. We estimated another 21,403 plausible species, which likely either lack a reference barcode or are undescribed. The total of 31,846 species is similar to the number of insect species known for Germany (~35,500). Because Malaise traps capture only a subset of insects, our approach identified many species likely unknown from Germany or new to science. Our reproducible workflow (~80% OTU-similarity among years) provides a blueprint for large-scale biodiversity monitoring of insects and other biodiversity components in near real time.

Introduction

Insects are the most diverse taxonomic group on Earth and contribute to essential ecosystem processes and services, such as pollination, nutrient cycling and organic matter decomposition (Cardoso et al. 2020). However, insect populations are declining (Wagner 2020) and our ability to mitigate these declines is hindered by poor understanding of spatial distributions, habitat requirements, biotic interactions, dynamics, and even the overall number of extant species. About 1 million insect species have been described to date, but recent estimates of total species numbers stand at about 5.5 million (Stork 2018) or even more (IPBES 2019). Surprisingly, many new species are being reported even for extremely well-studied and generally species-poor areas with a long history of entomological research. For example, based on 4,000 species caught with Malaise traps in Sweden, Karlsson et al. (2020) reported almost 700 insect species new to science and ~1,300 new to Sweden. Similarly, in a DNA barcode analysis of about 62,000 specimens collected with Malaise traps, Chimeno et al. (2022) estimated over 2,000 dipteran species new to Germany, raising the total number of 33,341 known insect species (Klausnitzer 2005) to approximately 35,500. Thus, while many new insect species are being continuously reported, the vast majority still remain unknown.

A key constraint to closing the current information gap on insects is the vast number of species and samples that must be processed. Terrestrial and aquatic monitoring methods, like Malaise, canopy, or light traps, as well as sediment (benthos) samples from freshwater, can collect thousands of specimens (e.g., Resh and Jackson 1993, Karlsson et al. 2020, Habel et al. 2023), many of which are small and difficult to identify even for experts, resulting in taxonomic neglect (Srivathsan et al. 2023) and extensive time needed for identification that by far outweighs the time required for collection. National and international monitoring networks, such as the Global Malaise Trap

program (Geiger et al. 2016) , BioScan (Hobern 2021), LifePlan (www.helsinki.fi/en/projects/lifeplan), or the Swedish Malaise Trap program (Karlsson et al. 2020), have started coordinated large-scale insect sampling initiatives based on standardized traps and procedures (e.g., Hallmann et al. 2017). However, classical taxonomic analyses of these samples cannot keep pace with the rate of collection, particularly given the low and globally declining number of taxonomic experts (European Commission 2022). While the importance of taxonomic expertise remains undisputed, there is a clear need for alternative methods to assess insect species diversity both efficiently and reliably (van Klink et al. 2020, Chua et al. 2023).

Following more than a decade of research and trial applications, DNA metabarcoding has now reached a high technology readiness level, presenting a promising solution for examining a fuller range of insect biodiversity in large-scale monitoring programs. A range of suitable protocols for specimen collection, processing and data analysis are now available (Buchner et al. 2021c, Montgomery et al. 2021), as are suitable primer pairs (Braukmann et al. 2019, Elbrecht et al. 2019) and an accessible platform for insect barcode reference data (Ratnasingham and Hebert 2007). The key strength of DNA metabarcoding is that it rapidly delivers taxonomically highly resolved taxa lists, whereas quantitative information on species abundance or biomass remains a challenge (Sickel et al. 2023).

Despite its potential, DNA metabarcoding has not yet been implemented in large-scale, long-term insect biodiversity monitoring programs. There are several likely reasons. In particular sample throughput is still constrained by the need for significant manual labor input, expensive DNA kits and reagents, and difficulties in accessing information on lab and analysis procedures (McGee et al. 2019), which all present roadblocks to large-scale implementation. Furthermore, incomplete and partly inconsistent reference databases still limit the accuracy and quantity of species-level assignments and thus the completeness and validity of the resulting species lists. This problem remains particularly pervasive for highly diverse groups like Diptera and Hymenoptera where species that are often difficult to distinguish based on morphological criteria and thus tend to be underrepresented in reference libraries. The unknown (i.e., undescribed) species in these poorly explored groups are often referred to as “dark taxa” (Hartop 2021). However, in the absence of reference barcodes for species, or even when formal species descriptions are lacking, DNA metabarcoding can still overcome these limitations using genetic distance thresholds to approximate entities that roughly reflect species, such as molecular Operational Taxonomic Units (OTUs) or Barcode Index Numbers (Ratnasingham and Hebert 2013) (BINs). Even in the absence of species in databases, these distance-based entities can approximate species numbers and have been applied in ecological and ecotoxicological research for many years (Sturmbauer et al. 1999, Hoppeler et al. 2016, Beermann et al. 2018).

DNA metabarcoding analyses to date have been limited to either a specific region within a country (e.g., Geiger et al. 2016, Uhler et al. 2021, Habel et al. 2023), short time spans (e.g., Huang et al. 2022, Li et al. 2023) or specific taxonomic groups (e.g., Chimeno et al. 2022, Huang et al. 2022). None has quantified the full extent of insect diversity – including dark taxa – at the whole-country scale and through time. Thus, the present study fills this gap by presenting a robust DNA metabarcoding workflow for application to large-scale insect monitoring programs, combined with a new multi-level

procedure to assess the validity of species records. The underlying dataset consists of 1,815 Malaise trap samples collected in 2019 and 2020 from 75 individual traps across Germany (Figure 1). Our aims were to: (1) present the workflow and provide detailed and publicly available laboratory procedures and programs; (2) quantify the time and cost investment required per sample; and (3) evaluate the reliability of the approach in terms of number of known species detected, number of dark taxa and their likely validity, and variation in species detected between years. The results obtained with the new workflow show that DNA metabarcoding is feasible for large-scale and long-term insect monitoring, and providing insight into insect diversity at scales that have been challenging to study so far.

Materials and Methods

Sampling

Sampling was conducted as part of the nationwide German Malaise trap monitoring program (Welti et al. 2022) comprising forests, grassland, agricultural and urban areas (https://www.ufz.de/lter-d/index.php?de=46285). Most of the 31 sites in which the Malaise traps were placed belong either to the German LTER-D network (Haase et al. 2016, Mirtl et al. 2018) (Long-Term Ecological Research) or to the network of national natural landscapes (https://nationale-naturlandschaften.de). At each site, one to six Malaise traps have been operated to monitor biomass and species composition of flying insects in different habitats. In the present study, we used a total of 1,815 Malaise trap samples from 56 locations across Germany during 2019, with 19 locations added in 2020 (Figure 1). Traps were emptied every two weeks from the beginning of April until the end of October in both years (approx. 15 samples per year, depending on site-specific climatic conditions). Insects were caught in 80% ethanol and their wet biomass was measured following Welti et al.(2022). Samples were kept in 96% ethanol and protected from light until later genetic analysis.

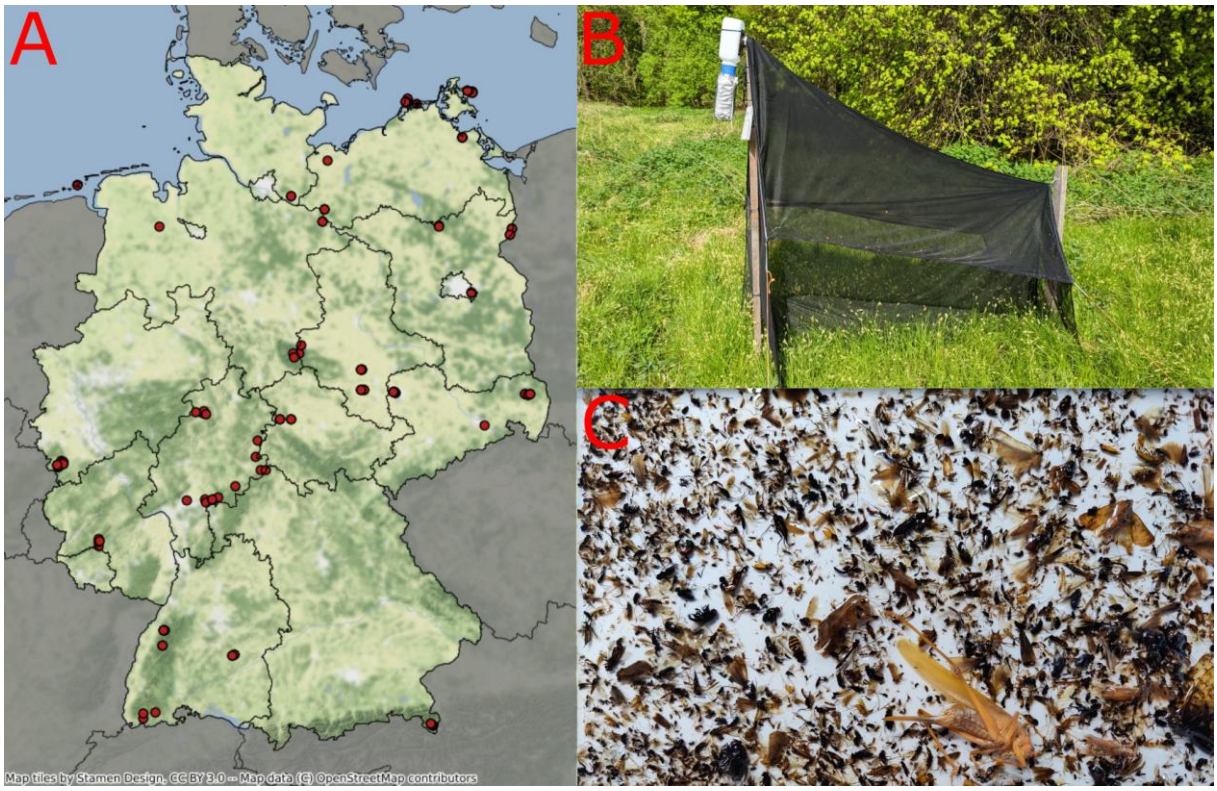

Figure 1: A: Location of the in total 75 Malaise traps across Germany**, B: Lateral view** of a Malaise trap with collection bottle protected from sunlight at the upper left end, and **C: top view** of a preserved Malaise trap sample spread in a white tray.

Sample processing

All samples were divided into two size classes (small ≤ 4 mm; large > 4 mm) to increase taxon recovery rates of small taxa (Elbrecht et al. 2021). Samples spread on a perforated plate sieve (4 mm hole diameter) were stirred using a magnetic stirrer (750 rpm) in ethanol (Figure S1), so that small individuals passed through the holes, whereas the large ones were retained on top of the sieve. The size fractions were homogenized following the protocol described in Buchner et al. (Buchner et al. 2021a), except that the homogenization time was reduced to 30 s. The two size fractions were subsequently pooled at a ratio of 1:4 (large 200 µL: small 800 µL) as recommended by Elbrecht et al. (2021).

DNA extraction

Generally, the laboratory steps followed the workflow described in Buchner et al. (Buchner et al. 2021c). All procedures are available as step-by-step protocols in the protocols.io repository (Buchner 2022b). All subsequent steps after size-sorting and homogenization were completed on a Biomek FX^P liquid handling workstation (Beckman Coulter, Brea, CA, USA). After sample lysis (Buchner 2022e), samples were processed in duplicate during the entire library preparation to control for possible cross-contamination. Additionally, in each 96-well plate twelve negative controls were included. DNA was extracted using a magnetic bead protocol (Buchner 2022a). Extraction success was verified on a 1% agarose gel. For all samples that did not amplify, the extraction was repeated with silica spin columns (Buchner 2022a), which were always successful.

Sequencing library preparation

The PCR for the metabarcoding library followed a two-step PCR protocol (Zizka et al. 2019) targeting a 205 bp fragment (Vamos et al. 2017) of the cytochrome oxidase c subunit I (COI) gene. DNA was amplified in a first PCR using the Qiagen Multiplex Plus Kit (Qiagen, Hilden, Germany) with a final concentration of 1x Multiplex Mastermix, 200 nM of each primer (fwh2F, fwhR2n (Vamos et al. 2017)) and 1 µL of DNA, filled up with PCR-grade water to a final volume of 10 µL. The amplification protocol was 5 min of initial denaturation at 95 °C, 20 cycles of 30 s denaturation at 95 °C, 90 s of annealing at 58 °C and 30 s of extension at 72 °C, followed by a final elongation step of 10 min at 68 °C. Each of the PCR plates used in the first step was tagged with a unique combination of inline tags. Additionally, the primers contained a universal binding site for the primer used in the second PCR step to anneal (Table S1). The PCR products were purified using a bead-based protocol and a ratio of 0.8x and an elution volume of 40 µL to remove remaining primers and potential primer dimers (Buchner 2022c).

In the second PCR, DNA was amplified at a final concentration of 1x Multiplex Mastermix, 100 nM of each primer (Table S1), 1x Corralload Loading Dye, 2 µL of the cleaned-up product of the first PCR in a final volume of 10 µL. The amplification protocol was 5 min of initial denaturation at 95°C, 25 cycles of 30 s denaturation at 95°C, 90 s of annealing at 61°C and 30 s of extension at 72°C, followed by a final elongation step of 10 min at 68 °C. PCR success was visualized on a 1% agarose gel.

To achieve a similar sequencing depth, the PCR products were normalized to equal concentrations. Normalization was achieved with a bead-based protocol and a ratio of 0.7x (Buchner 2022d) and an elution volume of 40 µL. The whole volume of the normalized samples was then pooled into the final libraries. Libraries were then concentrated using a silica spin-column protocol (Buchner 2022a). Library concentrations were quantified on a Fragment Analyzer (High Sensitivity NGS Fragment Analysis Kit; Advanced Analytical, Ankeny, IA, USA). The libraries were sequenced at Macrogen Europe using the HiSeq X platform with a paired-end (2x150 bp, 15 lanes) kit or at CeGaT (Tübingen, Germany) using the MiSeq V2 platform (2x150 bp, 1 lane).

Bioinformatics

Raw data of the sequencing runs were delivered demultiplexed by index reads. Since no differences were detected between sequencing runs, they were all pooled before subsequent analyses. Additional demultiplexing of the inline tags was achieved with the Python package “demultiplexer” (v1.1.0, https://github.com/DominikBuchner/demultiplexer). Reads were further processed with the APSCALE pipeline (Buchner et al. 2022) (v1.4.0, https://github.com/DominikBuchner/apscale using default settings. Briefly, paired-end reads were first merged using vsearch (Rognes et al. 2016) (v2.21.1) before the primer sequences were trimmed using cutadapt (Martin 2011) (v3.5). Only reads with a length of 205 bp (±10) and with a maximum expected error of 1 were retained. Identical reads less abundant than 4 were discarded and all others dereplicated before OTUs were clustered based on a similarity threshold of 97% and mapped to OTUs including

singletons. The resulting OTU table was filtered for erroneous OTUs with the LULU algorithm (Frøslev et al. 2017) as implemented in APSCALE. Taxonomic assignment was performed using BOLDigger (Buchner and Leese 2020) (v1.5.4, https://github.com/DominikBuchner/BOLDigger). The best hit was determined with the BOLDigger method and the API verification method. This resulted in a raw OTU table (Table S2) that was used in subsequent analysis.

Data filtering

To control for possible contamination during the laboratory workflow, the technical replicates of each sample, as well as the negative controls, were merged by summing up the reads, provided that the reads were present in both replicates. Subsequently, the maximum number of reads per OTU present in all of the negative controls was subtracted from the respective OTU (Table S3). All OTUs were analyzed for stop-codons, and any OTUs containing stop-codons were removed. We analyzed two datasets: 1) OTUs assigned at the species level and 2) OTUs assigned to insects. To further clean dataset 1, all OTUs sharing species assignments were merged by summing up their reads. Retrieved species names containing numbers or punctuation marks were also removed (e.g., incomplete database records). The resulting final species list (Table S4) for all samples was used for the validation procedure.

Validation of taxonomic assignment

To validate the resulting species-level list, three different approaches were used. First, the occurrences of all species were checked by taxonomic experts at the Entomological Society Krefeld, Germany. As a basis, the experts used the digitally available insect species catalogue from the Entomofauna Germanica (Klausnitzer 2005), which is best resembled through the list from the German Barcode of Life portal (https://gbol.bolgermany.de). All records of species found were then validated against known changes in synonymies as well as new records of insect species from Germany reported in scientific primarily literature available to them. Second, all detected species were checked for GBIF records within a 200 km radius around the given trap. This value was selected to be sufficiently high to accommodate also for the often patchy and rare species records in GBIF, the large size of the study area and the fact that many flying insect species are highly mobile. To do so, a polygon was drawn around all trap locations with occurrences of the respective species, and the border of this polygon was then extended by 200 km (Figure S2). Records were extracted from the GBIF database using the “rgbif” package (Chamberlain et al. 2022). Lastly, all species were checked against the German Barcode of Life database, which was downloaded with a custom Python script (Script S1). A record was accepted as valid when two of the three validation criteria were met, which also helps to control for false positives (Table S5).

Statistical analysis

To assess if sequencing depth was sufficient across all samples, we conducted a read-based rarefaction using a custom python script. This approach involved randomly sampling reads without replacement from each sample in increments of 0.1% (with 50 iterations per step) of the total read count. Subsequently, we fitted a Michaelis-Menten-type equation to the resulting rarefaction curve. Using this function, we computed OTU

richness when doubling the sequencing depth. If OTU richness increased by less than 5% when doubling the sequencing depth, the sequencing depth was considered sufficient (Figure S3, Table S6).

In addition to examining sequencing rarefaction curves, we also examined rarefaction curves for both validated and plausible species using the ¡Chao2 (Chiu et al. 2014) estimator. We did so to determine the potential influence of collecting more samples from each site, such as by sampling for a longer time period or more frequently, on the number of detected insect species. This was done by randomly drawing subsets of samples (50 iterations each) from the dataset without replacement in increments of 5 from 5 to 1815.

Plausible species and dark taxa estimation

To address the potential over-estimation of species diversity based on OTU numbers we computed the mean number of OTUs per validated species for each of the 20 insect orders within the dataset. This mean was then used to normalize the OTU count for each order. Specifically, we divided the number of OTUs per order by the calculated mean, providing an estimate of additional species present in the dataset that have not yet been assigned a species name. We refer to these additional species as plausible species. Furthermore, we obtained data from the Barcode of Life datasystems (accession date: 14th of April, 2023), including the number of species barcoded with a voucher specimen collected in Germany, along with the total number insect species known from Germany (Klausnitzer 2005). This additional information allowed us to distinguish plausible species lacking a reference sequence from potential dark data within the dataset. The described correction was performed on order as well as on family level, but we simplicity we focused the subsequent analysis on the order level.

Time and cost estimations

To estimate the time and costs needed for our nationwide insect diversity assessment via metabarcoding, we used vendor list prices for materials, as well as runtimes on the liquid handler plus estimates for laboratory set-up before and after each step of the workflow. These costs include all labware and chemicals needed to complete the respective step of the protocol. Incubation times are not included in the time estimates, because the time can be used to process other samples. The calculated price per sample is based on 1,815 samples including replicates and negative controls, resulting in a total number of 4,149 individual reactions spread across 44 96-well PCR plates. A total of 10 blenders were used for homogenization, which are included in the cost estimate. For sequencing, cost estimates are based on the current costs of 1,100 € for 110 Gb output at a commercial service provider (Macrogen Europe) using the NovaSeq 6000 S4, which has replaced the HiSeq platform, and a sequencing depth of at least 1.5 million reads per sample (~ 817 Gb). Labor costs were estimated at 60 € per hour, corresponding to an experienced scientist in Germany. The costs for the use of the liquid handler were estimated based on a linear depreciation over a total expected lifetime of 10 years. Annual gross maintenance costs of 40,000 € were assumed for instruments, resulting in total depreciation and maintenance costs of approximately 4,000 € for the present study. No rental and auxiliary costs needed to be applied in this study.

Results

Sequencing results and species validation

We performed high-throughput DNA metabarcoding including replicates and negative controls on 1,815 Malaise trap insect samples (775 for 2019 and 1,040 for 2020; Table 1) using a 205-bp fragment of the mitochondrial cytochrome *c* oxidase I (COI) gene as a marker. All samples and sequencing runs combined yielded 3,999,082,169 demultiplexed read pairs. The average read number per sample in the final read table was 1,401,469 (±SD of 618,631). Sequencing depth was sufficient for all samples, with a mean increase in richness of 0.17% by doubling the sequencing depth (0% – 1.46%; Table S6).

Sequencing yielded a total of 52,981 raw OTUs, 50,087 of which were assigned to insects. We were able to assign 11,776 species names to a total of 15,042 OTUs. Two thirds of these species (10,803) were validated via three different criteria involving i) expert judgment from entomologists with particular knowledge of long-term Malaise trap community data from German Malaise traps, ii) a comparison with an online database that includes the known German species (Klausnitzer 2005) (GBOL), and iii) a GBIF record check within a 200 km radius. Species were regarded as 'validated' when two of the three validation criteria supported them. The three different validation criteria showed a pairwise agreement exceeding 80%. Taxonomic experts validated an additional 355 species not yet not in the list of species reported from Germany (Table 1), leading to an overlap of 97% with the GBOL database. GBIF and taxonomic expert validation had an overlap of 84%, and GBIF and GBOL of 82% (Table S7). Consequently, a total of 35,045 insect OTUs remained unassigned, either because species-level reference data are lacking or the species are truly unknown (i.e., dark taxa; Table 1), resulting in a large discrepancy between the number of named species and recorded insect OTUs. Although sometimes several OTUs were assigned to the same species, 9,061 (83.9%) of our validated species were represented by a single OTU (Table S8).

Table 1: Summary statistics of OTUs, assigned insect species and OTUs assigned to validated insect species according to the two-out-of-three validation criterion. Plausible species calculated from the average number of OTUs per validated species and family. n = number of Malaise trap samples.

	2019 (n=775)	**2020 (n=1,040)**	**Σ (n=1,815)**
Raw OTUs	41,189	50,166	52,981
Insect OTUs	39,367	47,652	50,087
Insect species	9,728	11,183	11,776
Validated species according to:			
Expert validation	9,132	10,475	11,030
GBIF validation	7,969	8,867	9,254
GBOL validation	8,777	10,056	10,574
Validated species	**8,983**	**10,276**	**10,803**
Plausible species			**21,403**
Total species			**31,846**

Species richness estimation

Our sampling effort captured the majority of OTUs. Based on rarefaction curves, a greatly increased sampling effort in each site, for instance by sampling more frequently or for a longer period at all sites, may have only resulted in detecting an additional 4,725 OTUs (+9.4%) and 934 validated species (+8.6%; Figure 2). Additionally, most of the OTUs were found in both sampling years (36,932 = 73.7%) with more insect OTUs occurring exclusively in 2020 (10,720) than in 2019 (2,435). The same is true for the validated insect species. Most were found in both years (8,456 = 78.3%) but more than three times as many occurred exclusively in 2020 (1,820) compared to 2019 (527).

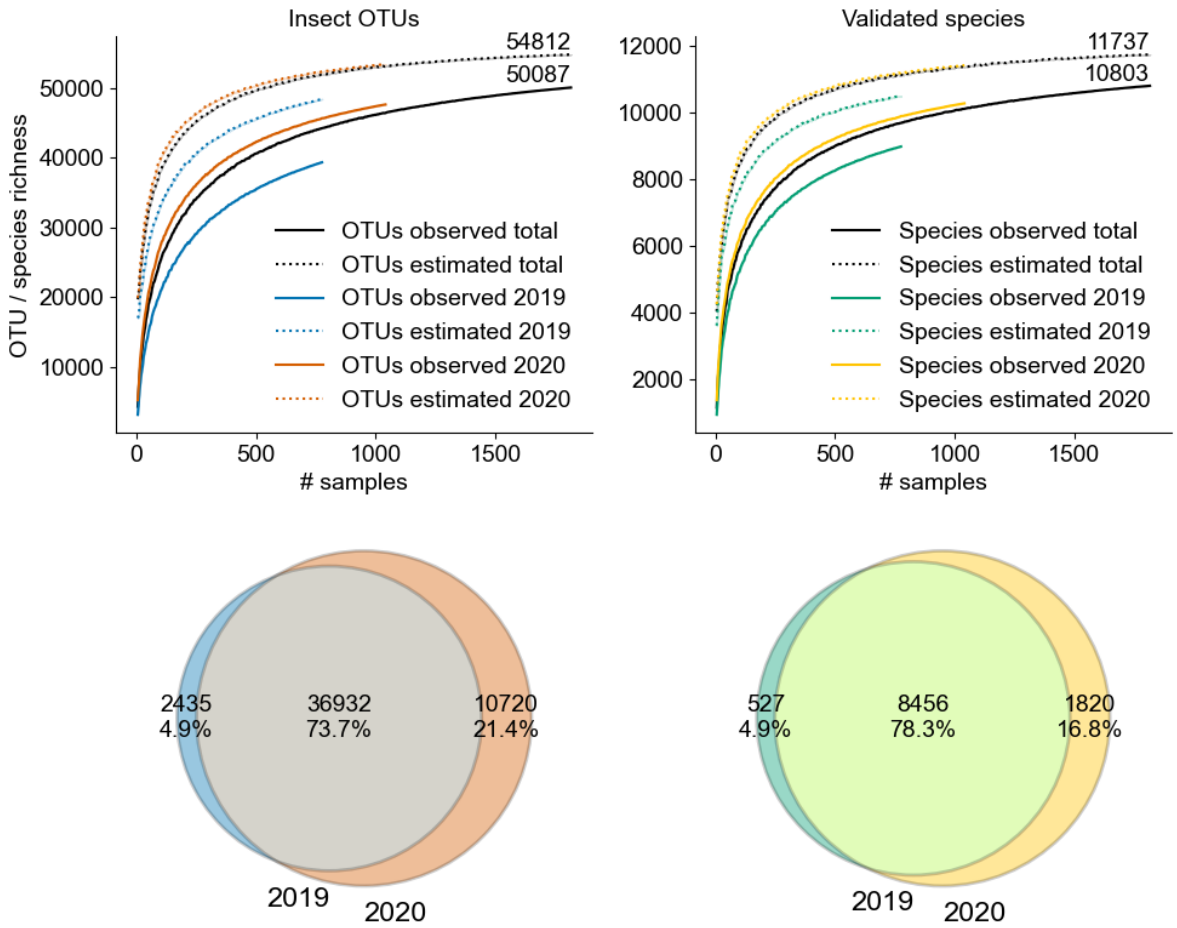

Figure 2: Top row: Rarefaction curves showing richness of insect OTUs and validated species richness that were either observed (solid lines) or estimated (dotted lines). **Bottom row:** Shared and unique absolute numbers and proportions of the insect OTUs (left circles) and validated insect species (right circles) collected in 2019 and 2020.

OTU distribution across insect groups and unknown taxa

The four most diverse insect orders in Germany (Diptera, Hymenoptera, Lepidoptera, and Coleoptera) were well represented in the dataset (Figure 3, top row). Diptera and Hymenoptera were the most common orders both at the OTU (22,732 = 45.4% and 12,823 = 25.6%) and species level (3,851 = 35.6% and 2,370 = 21.9%). The 10,803 validated insect species represent 33.5% of the 35,500 insect species known in

Germany (Klausnitzer 2005) and 82.6% of the 13,076 barcoded insect species recorded in the country (Table S9).

The percentage of OTUs assigned to species differed considerably among insect orders, being highest for Lepidoptera (75%), followed by Coleoptera (59%). Less than a fifth of the OTUs identified as Diptera and Hymenoptera could be assigned at the species-level (17% and 18%, respectively). The lowest percentage of OTUs assigned to the species level was for Orthoptera (<1%) (Figure 3, bottom left panel). The mean number of OTUs per validated insect species varied slightly among insect orders and families, typically between 1 and 1.5, but Orthoptera (specifically the Acrididae) were represented by >4 and Zygentoma (specifically Lepismatidae) by ~2 OTUs per species (Figure 3, bottom right panel).

Based on the mean number of OTUs per validated insect species calculated at either the order or family level, we estimated that our data set respectively included either an additional 22,496 or 21,403 plausible insect species (Table S10). All OTUs belonging to Orthoptera were removed for these estimates to avoid artificially inflating the total number of species (Information S1). The lower of the two numbers is a more conservative estimate, but since we identified 435 different families (see Table S11 for further family level information), our analysis focuses, for simplicity reasons, on the order level. The additional plausible species could be those that either a) could not be identified to species level due to a lack of a reference in the present reference database, b) had not previously been recorded from Germany, or c) represent new species to science. Examples of species not reported from Germany or new to science ('potential dark taxa' in Figure 3, Table S10) were particularly evident in the Diptera and Hymenoptera, for which we respectively found ~6,600 and ~1,200 species while the missing reference data aspect affected the assignment in all orders except the Mantodea (with 1 species only).

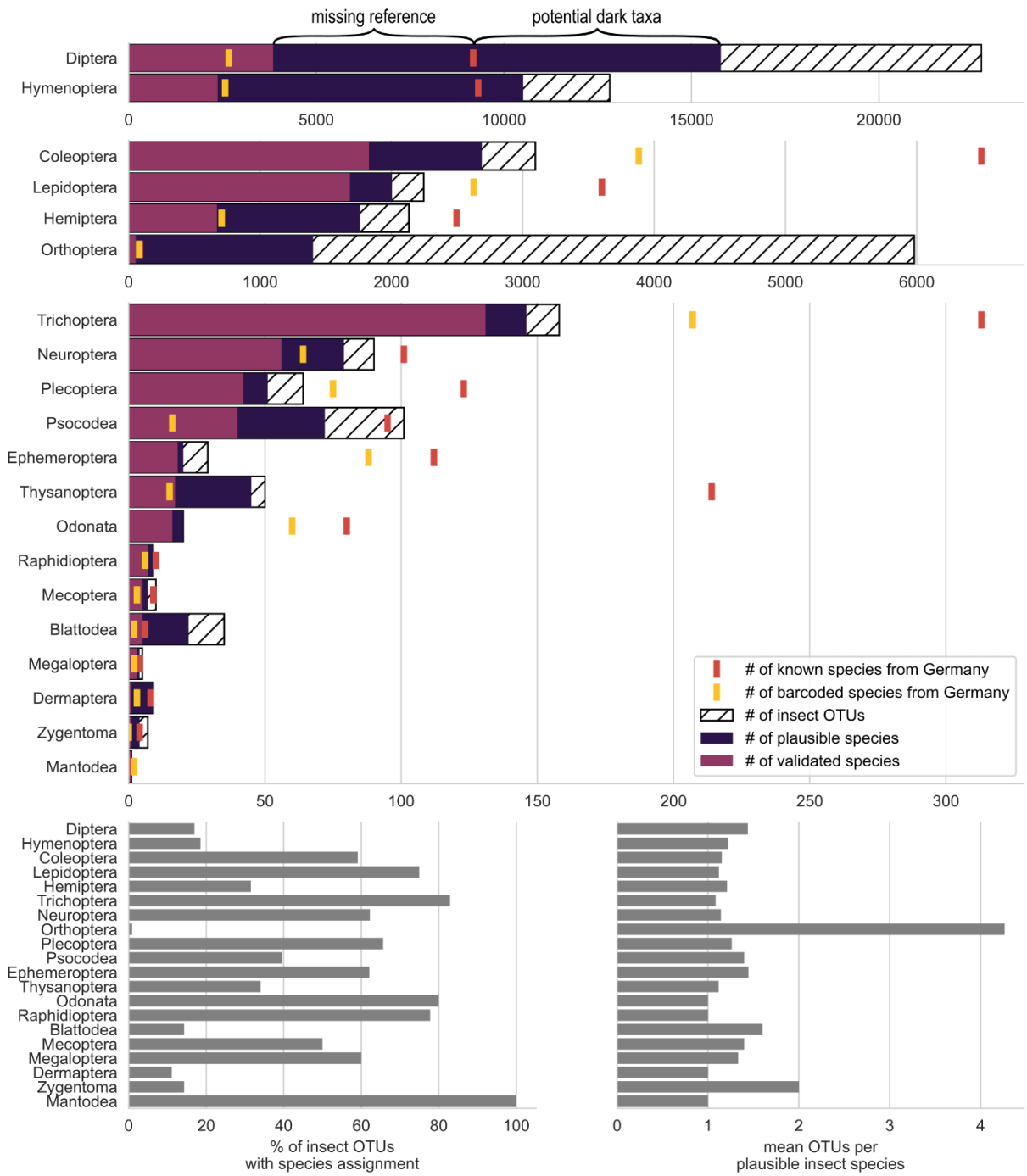

Figure 3: Top panels: Number of total insect OTUs, plausible and validated insect species per order identified in the present study compared to the total number of species reported from Germany or sampled in Germany and included in the BOLD database (data accessed on 14th of April, 2023). **Bottom left panel**: Percentage of OTUs assigned to species. **Bottom right panel:** Mean number of distinct OTUs assigned to a given plausible species.

Time and cost estimation

The total costs of the laboratory workflow for duplicate sample processing and negative controls - including all needed labware, chemicals, sequencing and salaries - were estimated at 88,000 €, equivalent to about 46 € per sample (Tables 2, S12). Costs for laboratory materials accounted for 12 € (26%) of the total costs per sample, and salaries for 34 € (74%). Sequencing was the most expensive step (contributing 4.85 € per sample), followed by enzymatic steps such as PCR (1.65 €) and sample lysis (1.54 €). The total processing time for all 1,815 samples was 1,030 working hours or 27.5 weeks, equivalent to 141 samples within 2 weeks for one person working full time and supported by one liquid handling robot. Most of the processing time was needed for sample size-sorting and homogenization. All subsequent steps were completed within 6 weeks.

Table 2: Cost estimates for the proposed workflow: Cost estimates are based on vendor list prices in 2022 where available. Time estimates are either hands-on time in the laboratory or mean runtimes of the robotic protocols. Costs and time required per sample are based on 1,815 samples. Total time and total costs are based on the respective number of replicates processed in each step.

Processing step	Material costs per sample [€]	Total material costs [€]	Labor costs per sample [€]	Time per sample [h:min:s]	Total time [h:min:s]	Total costs [€]
Size-sorting	-	-	15.09	0:15:00	453:45:00	27,383.81
Homogenization	0.21	379.90	11.50	0:11:26	345:45:00	21,252.45
Sample lysis	1.54	2,795.64	1.91	0:01:54	57:37:30	6,264.26
DNA extraction	1.02	1,845.55	1.09	0:01:05	33:00:00	3,823.27
QA DNA extraction	0.04	78.48	0.74	0:00:44	22:00:00	1,417.24
1st PCR	1.65	2,994.43	0.25	0:00:15	07:20:00	3,450.83
PCR cleanup	0.50	898.46	0.74	0:00:44	22:00:00	2,237.22
2nd PCR	1.65	2,994.43	0.37	0:00:22	11:00:00	3,663.81
Normalization	0.47	844.47	0.74	0:00:44	22:00:00	2,183.23
Pooling	<0.01	0.45	0.13	0:00:08	04:00:00	243.86
QC libraries	<0.01	10.86	0.13	0:00:08	04:00:00	254.27
Sequencing	4.85	8,800.00	-	-	-	8,800.00
Bioinformatic analysis	-	-	1.59	0:01:35	48:00:00	2,890.51
Liquid handler depreciation and maintenance			-	-	-	4,000.00
Total	11.92	21,642.67	34.28	34:04	1030:27:00	**87,864.76**

Discussion

A metabarcoding workflow for large-scale biodiversity monitoring

We here present a comprehensive, scalable, and cost-efficient DNA metabarcoding workflow which allows analyzing thousands of specimen- and species-rich samples within several weeks to months depending on the available workforce. The proposed DNA metabarcoding workflow differs from others (e.g. Braukmann et al. 2019, Hardulak et al. 2020, Hausmann et al. 2020) in several important aspects that are key to high sample throughput at reduced costs and processing time while ensuring high-quality data. Key differences include: i) homogenization of samples in preservative liquid to avoid a time-consuming drying step (Buchner et al. 2021a); ii) processing of

all samples in duplicate before DNA extraction as an essential quality assurance measure to reduce the probability of false positive signals; iii) completion of all laboratory procedures by an automated liquid handling robot to minimize processing time (Buchner et al. 2021c) (except for the gel electrophoresis) and maximize consistency; iv) a mean sequencing depth increased to ~1.4 M reads per sample to boost species detectability; and v) publication of all laboratory procedures as open-source protocols (Buchner 2022b) or programs (Buchner and Leese 2020, Buchner et al. 2022) to ensure full transparency and reproducibility following the FAIR principles (Wilkinson et al. 2016).

Labor and material costs of metabarcoding protocols are often considered prohibitively high for large-scale monitoring programs (Borrell et al. 2017, Montgomery et al. 2021), frequently exceeding 200 € per sample and up to 400 € (Ji et al. 2013, Elbrecht et al. 2017, Aylagas et al. 2018). The workflow presented here considerably reduces these costs, down to <50 €. This is primarily achieved by automating crucial laboratory steps and by preparing all required solutions instead of purchasing expensive commercial kits. Costs could be further cut in future large-scale programs by ordering chemicals and consumables in bulk and by further reducing reaction volumes for all laboratory steps involving enzymes, wherever possible (Buchner et al. 2021b). An important consideration to lower labor costs is to reduce the processing time per sample, notably for size-sorting and homogenization, ideally by automating these steps as well. Furthermore, expenses related to sampling in the field are not included in our analysis, although they can easily match laboratory costs. For example, in our monitoring program, costs for Malaise trap installation, maintenance and biweekly sampling from April to October (15 samples per site and year in total) amounted to about 12 working hours or 240 € for a student aid. Adding travel time (highly variable but assumed here to average 15 hours or 300 €) and material (approx. 450 €) would result in total costs for field sampling of 66 € per sample, unless much of the work is accomplished by volunteers. This sum ignores expenses for rent and additional costs (heating, electricity etc.), which need to be added when commercial providers are solicited.

Another crucial but often neglected step in metabarcoding workflows is species validation. Validating species records poses a significant challenge for many taxa due to the dearth of reference sequences, few checklists and experts, and diverse algorithmic approaches (Hleap et al. 2021). However, these challenges could be partly addressed by using publicly available species databases, as suggested by the remarkably close agreement we found among our three validation procedures of insect experts, GBIF, and GBOL matching. The taxonomic experts added further species to the list of known species from Germany based on recent scientific evidence. This agreement suggests that relying solely on public databases, like GBIF (Telenius 2011) and the GBOL database, which most closely resembles the reports of the Entomofauna Germanica (Klausnitzer 2005), is sufficiently accurate for large-scale monitoring where expert validation tends to be prohibitive. This is not to advocate disregarding expert knowledge, which is critical to minimize the well-known errors inherent in automated validation methods (e.g. Meiklejohn et al. 2019). Instead, we would like to raise awareness that acceptable alternatives exist to overcome time constraints, as they typically arise in large biodiversity monitoring programs. FAIR and curated reference databases with up-to-date taxonomic names and integration of synonyms is essential

for harmonized biodiversity monitoring and establishing such systems a key challenge (Keck and Altermatt 2023).

A reliable account of biodiversity

Rarefaction indicates that the 10,803 insect species identified (without considering unassigned OTUs) with our protocol would increase by only 934 species (8.6%) if sampling efforts were ramped up at the established study sites. However, we acknowledge that higher species numbers would be obtained if more sites were sampled, in particular including new habitat types. This expectation is consistent with the larger number of validated insect species in 2020 (+14.4%) when 19 more sites were sampled compared to 2019. The 10,803 detected species represent one third of all insect species reported from Germany and ~83.1% of all barcoded insects from this country. This includes nearly 100% of the up-to-date barcoded Hymenoptera and Hemiptera, indicating that with just 75 sampling sites our approach captures a large portion of the known insect biodiversity of Germany. Other metabarcoding studies from regions within Germany have identified ~5,900 species (Uhler et al. 2021) and ~ 11,984 insect OTUs (Habel et al. 2023). Similarly, a DNA barcoding study investigating 1% of the samples from the Swedish Malaise trap program found >4,000 species (Karlsson et al. 2020), and another recent DNA barcoding study based on Malaise traps across eight countries from four continents found >25,000 species (Srivathsan et al. 2023). While we acknowledge that comparisons of insect species and OTUs numbers among studies are difficult due to differences in protocols, these examples indicate, along with the high proportion of known or barcoded insects from Germany, that rapid and cost effective metabarcoding can provide a robust inventory of biodiversity.

At the scale of our nationwide sampling network, we consistently found the same species in two consecutive years (~75% species and ~80% OTU-similarity among years), highlighting the spatiotemporal reliability of taxa lists derived from metabarcoding when used in large-scale monitoring. One of the principal concerns in large-scale sampling is that high community variability in space and time, and sampling idiosyncrasies, make it difficult to detect the same species even in spatially or temporally proximate samples. This issue is particularly relevant when using methods, like metabarcoding, that can identify thousands of species, most of which will be rare and so may only occur sporadically in a single site (Jeliazkov et al. 2022). For example, traps just tens of meters apart (e.g., Steinke et al. 2021) or in successive sampling periods (e.g., Sinclair et al., in prep, Information S1) can capture very different insect communities. However, the between-year consistency of our data suggests this issue may diminish as the spatial scale of sampling for metabarcoding increases, such as in larger-scale biodiversity monitoring conducted across whole countries.

Assessing dark taxa biodiversity

A key finding of our study is the discovery of approximately 5 times more insect OTUs than validated insect species, demonstrating the potential value of cost-effective metabarcoding for uncovering as yet unknown biodiversity in large-scale monitoring. One explanation for the higher number of OTUs we found is that many unassigned OTUs represent described species that cannot be assigned to species level because reference sequences are lacking. Alternatively, many of these OTUs could represent

dark taxa, i.e. species new to science, as highlighted by Karlsson et al. (2020) and Srivathsan et al. (2023). While it is difficult to estimate the number of OTUs attributed to either known species without a reference barcode or to new species, we can roughly quantify this by comparing OTU-to-species ratios. Our mean number of OTUs per validated insect species of 1.0-1.5 suggests that our chosen cutoff of 3% sequence identity generally delineates different species accurately. The only exception is the Orthoptera, which are known to exhibit many pseudogenes (see Information S1) and that would inflate the OUT-to-species ratio. Specific correction factors for different taxonomic groups can also be estimated to infer the portion of undescribed species in the OTU datasets, which our results suggest is a large proportion. This is particularly true for the Hymenoptera and Diptera, where, after correcting the OTU numbers, we estimate about 6,600 and 1,200 potential new species to science for the two groups. This result agrees with reports from the Swedish Malaise Trap Program where most of the observed 700 species new to science (Karlsson et al. 2020) belonged to Diptera and Hymenoptera (Ronquist et al. 2020). However, our estimate of putatively undiscovered species in Germany is obviously a significant underrepresentation for two main reasons: First, while Malaise traps capture many different insect species, they likely miss many non-flying insects, insects flying high above ground, e.g. in tree canopies, or avoiding or escaping Malaise traps (e.g. Habel et al. 2023). Second, while our 75 Malaise traps reflect almost entirely the north-south and east-west gradients of Germany, they do not cover all regions in the country. Consequently, there is likely a trove of undiscovered species, even in countries like Germany and Sweden with a relatively species-poor fauna and a long taxonomic tradition. DNA metabarcoding can help discover the extent of this unknown biodiversity.

Technical challenges and ways forward

Metabarcoding comes with some drawbacks and technical challenges that must be considered. Firstly, our approach is a compromise between improving feasibility for large-scale monitoring and the chances of detecting all taxa. Additional sample processing, such as complementary mild lysis (Marquina et al. 2022, Iwaszkiewicz-Eggebrecht et al. 2023), greater replication (Zizka et al. 2022b) and using multiple gene markers or primers (Elbrecht et al. 2019, Hajibabaei et al. 2019), as well as increasing sequencing depth (this study), would undoubtedly recover more species. However, the additional effort required must be weighed against the extra information gained, which may not be substantial (Buchner et al. 2021a) Secondly, the incompleteness of regional and interregional reference sequence databases, and the generation of reference sequences of unknown species, remains a persistent challenge (e.g. Karlsson et al. 2020, Chua et al. 2023). Poorly studied groups present the largest hurdle, such as Diptera, Hymenoptera and various beetle families present the largest impediment. Improved collaboration between taxonomists, molecular ecologists, and experts from other fields, such as computer vision and deep learning (Høye et al. 2021, Wührl et al. 2022), will be needed to fill these gaps and to improve biodiversity assessments in general. Specifically, sequence data need to be linked with valid taxonomic names and undescribed species need to be described. This could be done by re-sampling those sites, where an unknown or undescribed species has been identified by metabarcoding. Alternatively, techniques like mild lysis need to be further developed, to allow a similar high coverage of species while also allowing

morphological identification. Vouchers would also help obtain the ecological trait information needed link the biodiversity changes with ecosystem function changes. Lastly, a key advantage of molecular methods is that tissue and DNA samples can easily be stored in miniaturized formats and be reanalyzed in the future (Zizka et al. 2022a). Emerging methods such as metagenomics can thus be applied in the future to stored samples to gain additional insights through reanalysis. This also makes direct intercalibration of different methods possible enabling harmonized biodiversity monitoring despite methodological advancements.

Implications beyond insect monitoring

We used insects collected with Malaise traps to demonstrate the value of a new DNA metabarcoding workflow that is reliable, scalable, fast, cost-effective, and particularly well-suited for large-scale monitoring of highly diverse taxonomic groups. The approach is, however, by no means limited to insects from Malaise traps, given that many of the key advances we highlight (e.g., automated workflow, robust species validation) are applicable to a variety of sampling methods, other invertebrate and vertebrate taxa, and even environmental DNA from various sources including soil and sediment (Pawlowski et al. 2022), by aligning all post-sampling processing steps with the requirements for robotic high sample throughput (Buchner et al. 2021c). The presented workflow thus moves us closer to realizing the overall vision for metabarcoding, i.e., to generate and link high-throughput biodiversity analyses with large-scale monitoring (Bush et al. 2017). Such integration would greatly enhance assessments of the massive ongoing changes in global biodiversity experienced at the present (e.g., Sinclair et al., in prep, Information S2) and biodiversity protection (e.g., the Kunming-Montreal Global Biodiversity Framework of the CBD), including Red List and invasive species assessments as part of policy frameworks on biodiversity conservation (e.g., Wetzel et al. 2015). As demonstrated here, integrating metabarcoding into large-scale monitoring networks is a powerful means to improving our understanding of biodiversity change and supporting conservation actions.

Acknowledgements

This project received funding from the Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz of the German federal State of Hesse (Center for Translational Biodiversity Genomics; LOEWE-TBG), the Hessische Landesamt für Naturschutz, Umwelt und Geologie (HLNUG) and the EU Horizon 2020 project eLTER PLUS (grant agreement no. 871128). We are particular grateful to the participating LTER-D and NNL sites. We furthermore want to thank Michael Sachtleben, Monika Degebrodt and Katrin Preuß (IGB), Thomas Hahn and Petra Möhl (Ranger Nature Park Stechling-Ruppiner Land), Dr. Mario Schrumpf (Nature Park Stechlin-Ruppiner Land), Sebastian Flinkerbusch, Michael Hinz, Enno Klipp and Roland Wollgarten (Eifel National Park) and Günter Hoenselaar (Kellerwald-Edersee National Park) and all members involved in the operation of traps and the sample collection for their assistance and support of the project.

References

Aylagas E, Borja Á, Muxika I, Rodríguez-Ezpeleta N (2018) Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. Ecological Indicators 95: 194–202. https://doi.org/10.1016/j.ecolind.2018.07.044

Beermann AJ, Zizka VMA, Elbrecht V, Baranov V, Leese F (2018) DNA metabarcoding reveals the complex and hidden responses of chironomids to multiple stressors. Environmental Sciences Europe 30: 26. https://doi.org/10.1186/s12302-018-0157-x

Borrell YJ, Miralles L, Huu HD, Mohammed-Geba K, Garcia-Vazquez E (2017) DNA in a bottle—Rapid metabarcoding survey for early alerts of invasive species in ports. PLOS ONE 12: e0183347. https://doi.org/10.1371/journal.pone.0183347

Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, de Waard JR, Sones JE, Zakharov EV, Hebert PDN (2019) Metabarcoding a diverse arthropod mock community. Molecular Ecology Resources 19: 711–727. https://doi.org/10.1111/1755-0998.13008

Buchner D (2022a) Guanidine-based DNA extraction with silica-coated beads or silica spin columns. https://doi.org/10.17504/protocols.io.eq2ly73mmlx9/v2

Buchner D (2022b) Invertebrate bulk sample metabarcoding protocol collection. https://doi.org/10.17504/protocols.io.j8nlkw4n6l5r/v4

Buchner D (2022c) PCR cleanup and size selection with magnetic beads. https://doi.org/10.17504/protocols.io.36wgqj45xvk5/v3

Buchner D (2022d) PCR normalization and size selection with magnetic beads. https://doi.org/10.17504/protocols.io.q26g7y859gwz/v3

Buchner D (2022e) Sample preparation and lysis of homogenized malaise trap samples. https://doi.org/10.17504/protocols.io.dm6gpjrmjgzp/v1

Buchner D, Leese F (2020) BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. Metabarcoding and Metagenomics 4: e53535. https://doi.org/10.3897/mbmg.4.53535

Buchner D, Haase P, Leese F (2021a) Wet grinding of invertebrate bulk samples – a scalable and cost-efficient protocol for metabarcoding and metagenomics. Metabarcoding and Metagenomics 5: e67533. https://doi.org/10.3897/mbmg.5.67533

Buchner D, Macher T-H, Leese F (2022) APSCALE: advanced pipeline for simple yet comprehensive analyses of DNA metabarcoding data. Bioinformatics 38: 4817–4819. https://doi.org/10.1093/bioinformatics/btac588

Buchner D, Beermann AJ, Leese F, Weiss M (2021b) Cooking small and large portions of “biodiversity-soup”: Miniaturized DNA metabarcoding PCRs perform as good as large-volume PCRs. Ecology and Evolution. https://doi.org/10.1002/ece3.7753

Buchner D, Macher T-H, Beermann AJ, Werner M-T, Leese F (2021c) Standardized high-throughput biomonitoring using DNA metabarcoding: Strategies for the adoption of automated liquid handlers. Environmental Science and Ecotechnology 8: 100122. https://doi.org/10.1016/j.ese.2021.100122

Bush A, Sollmann R, Wilting A, Bohmann K, Cole B, Balzter H, Martius C, Zlinszky A, Calvignac-Spencer S, Cobbold CA, Dawson TP, Emerson BC, Ferrier S, Gilbert MTP, Herold M, Jones L, Leendertz FH, Matthews L, Millington JDA, Olson JR, Ovaskainen O, Raffaelli D, Reeve R, Rödel M-O, Rodgers TW, Snape S, Visseren-Hamakers I, Vogler AP, White PCL, Wooster MJ, Yu DW (2017) Connecting Earth observation to high-throughput biodiversity data. Nature Ecology & Evolution 1: 1–9. https://doi.org/10.1038/s41559-017-0176

Cardoso P, Barton PS, Birkhofer K, Chichorro F, Deacon C, Fartmann T, Fukushima CS, Gaigher R, Habel JC, Hallmann CA, Hill MJ, Hochkirch A, Kwak ML, Mammola S, Ari Noriega J, Orfinger AB, Pedraza F, Pryke JS, Roque FO, Settele J, Simaika JP, Stork NE, Suhling F, Vorster C, Samways MJ (2020) Scientists’ warning to humanity on insect extinctions. Biological Conservation 242: 108426. https://doi.org/10.1016/j.biocon.2020.108426

Chamberlain S, Oldoni D, Waller J (2022) rgbif: Interface to the Global Biodiversity Information Facility API. https://doi.org/10.5281/zenodo.6023735

Chimeno C, Hausmann A, Schmidt S, Raupach MJ, Doczkal D, Baranov V, Hübner J, Höcherl A, Albrecht R, Jaschhof M, Haszprunar G, Hebert PDN (2022) Peering into the darkness: DNA barcoding reveals surprisingly high diversity of unknown species of Diptera (Insecta) in Germany. Insects 13: 82. https://doi.org/10.3390/insects13010082

Chiu C-H, Wang Y-T, Walther BA, Chao A (2014) An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. Biometrics 70: 671–682. https://doi.org/10.1111/biom.12200

Chua PYS, Bourlat SJ, Ferguson C, Korlevic P, Zhao L, Ekrem T, Meier R, Lawniczak MKN (2023) Future of DNA-based insect monitoring. Trends in Genetics 0. https://doi.org/10.1016/j.tig.2023.02.012

Directorate-General for Environment (European Commission), Hochkirch A, Casino A, Penev L, Allen D, Tilley L, Georgiev T, Gospodinov K, Barov B (2022) European Red List of insect taxonomists. Publications Office of the European Union, LU. Available from: https://data.europa.eu/doi/10.2779/364246 (December 15, 2023).

Elbrecht V, Vamos EE, 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. https://doi.org/10.1111/2041-210X.12789

Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D (2019) Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ 7: e7745. https://doi.org/10.7717/peerj.7745

Elbrecht V, Bourlat SJ, Hörren T, Lindner A, Mordente A, Noll NW, Schäffler L, Sorg M, Zizka VMA (2021) Pooling size sorted Malaise trap fractions to maximize taxon recovery with metabarcoding. PeerJ 9: e12177. https://doi.org/10.7717/peerj.12177

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8: 1188. https://doi.org/10.1038/s41467-017-01312-x

Geiger MF, Moriniere J, Hausmann A, Haszprunar G, Wägele W, Hebert PDN, Rulik B (2016) Testing the Global Malaise Trap Program – How well does the current barcode reference library identify flying insects in Germany? Biodiversity Data Journal: e10671. https://doi.org/10.3897/BDJ.4.e10671

Haase P, Frenzel M, Klotz S, Musche M, Stoll S (2016) The long-term ecological research (LTER) network: Relevance, current status, future perspective and examples from marine, freshwater and terrestrial long-term observation. Ecological Indicators 65: 1–3. https://doi.org/10.1016/j.ecolind.2016.01.040

Habel JC, Ulrich W, Segerer AH, Greifenstein T, Knubben J, Morinière J, Bozicevic V, Günter A, Hausmann A (2023) Insect diversity in heterogeneous agro-environments of Central Europe. Biodiversity and Conservation 32: 4665–4678. https://doi.org/10.1007/s10531-023-02717-5

Hajibabaei M, Porter TM, Wright M, Rudar J (2019) COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. PLOS ONE 14: e0220953. https://doi.org/10.1371/journal.pone.0220953

Hallmann CA, Sorg M, Jongejans E, Siepel H, Hofland N, Schwan H, Stenmans W, Müller A, Sumser H, Hörren T, Goulson D, Kroon H de (2017) More than 75 percent decline over 27 years in total flying insect biomass in protected areas. PLOS ONE 12: e0185809. https://doi.org/10.1371/journal.pone.0185809

Hardulak LA, Morinière J, Hausmann A, Hendrich L, Schmidt S, Doczkal D, Müller J, Hebert PDN, Haszprunar G (2020) DNA metabarcoding for biodiversity monitoring in a national park: Screening for invasive and pest species. Molecular Ecology Resources 20: 1542–1557. https://doi.org/10.1111/1755-0998.13212

Hartop E (2021) A multi-faceted approach to a “dark taxon”: The hyperdiverse and poorly known scuttle flies (Diptera: Phoridae). Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-192276 (April 12, 2023).

Hausmann A, Segerer AH, Greifenstein T, Knubben J, Morinière J, Bozicevic V, Doczkal D, Günter A, Ulrich W, Habel JC (2020) Toward a standardized

quantitative and qualitative insect monitoring scheme. Ecology and Evolution 10: 4009–4020. https://doi.org/10.1002/ece3.6166

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources 21: 2190–2203. https://doi.org/10.1111/1755-0998.13407

Hobern D (2021) BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome 64: 161–164. https://doi.org/10.1139/gen-2020-0009

Hoppeler F, Tachamo Shah RD, Shah DN, Jähnig SC, Tonkin JD, Sharma S, Pauls SU (2016) Environmental and spatial characterisation of an unknown fauna using DNA sequencing – an example with Himalayan Hydropsychidae (Insecta: Trichoptera). Freshwater Biology 61: 1905–1920. https://doi.org/10.1111/fwb.12824

Høye TT, Ärje J, Bjerge K, Hansen OLP, Iosifidis A, Leese F, Mann HMR, Meissner K, Melvad C, Raitoharju J (2021) Deep learning and computer vision will transform entomology. Proceedings of the National Academy of Sciences 118: e2002545117. https://doi.org/10.1073/pnas.2002545117

Huang J, Miao X, Wang Q, Menzel F, Tang P, Yang D, Wu H, Vogler AP (2022) Metabarcoding reveals massive species diversity of Diptera in a subtropical ecosystem. Ecology and Evolution 12: e8535. https://doi.org/10.1002/ece3.8535

IPBES (2019) Summary for policymakers of the global assessment report on biodiversity and ecosystem services. Zenodo https://doi.org/10.5281/zenodo.3553579

Iwaszkiewicz-Eggebrecht E, Granqvist E, Buczek M, Prus M, Kudlicka J, Roslin T, Tack AJM, Andersson AF, Miraldo A, Ronquist F, Łukasik P (2023) Optimizing insect metabarcoding using replicated mock communities. Methods in Ecology and Evolution 14: 1130–1146. https://doi.org/10.1111/2041-210X.14073

Jeliazkov A, Gavish Y, Marsh CJ, Geschke J, Brummitt N, Rocchini D, Haase P, Kunin WE, Henle K (2022) Sampling and modelling rare species: Conceptual guidelines for the neglected majority. Global Change Biology 28: 3754–3777. https://doi.org/10.1111/gcb.16114

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16: 1245–1257. https://doi.org/10.1111/ele.12162

Karlsson D, Hartop E, Forshage M, Jaschhof M, Ronquist F (2020) The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. Biodiversity Data Journal 8: e47255. https://doi.org/10.3897/BDJ.8.e47255

Keck F, Altermatt F (2023) Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. Molecular Ecology Resources 23: 511–518. https://doi.org/10.1111/1755-0998.13723

Klausnitzer B (2005) Die Insektenfauna Deutschlands (“Entomofauna Germanica”) - ein Gesamtüberblick. 37: 87–97.

van Klink R, Bowler DE, Gongalsky KB, Swengel AB, Gentile A, Chase JM (2020) Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. Science 368: 417–420. https://doi.org/10.1126/science.aax9931

Li M, Lei T, Wang G, Zhang D, Liu H, Zhang Z (2023) Monitoring insect biodiversity and comparison of sampling strategies using metabarcoding: A case study in the Yanshan Mountains, China. Ecology and Evolution 13: e10031. https://doi.org/10.1002/ece3.10031

Marquina D, Roslin T, Łukasik P, Ronquist F (2022) Evaluation of non-destructive DNA extraction protocols for insect metabarcoding: gentler and shorter is better. Metabarcoding and Metagenomics 6: e78871. https://doi.org/10.3897/mbmg.6.78871

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12. https://doi.org/10.14806/ej.17.1.200

McGee KM, Robinson CV, Hajibabaei M (2019) Gaps in DNA-based biomonitoring across the globe. Frontiers in Ecology and Evolution 7. Available from: https://www.frontiersin.org/articles/10.3389/fevo.2019.00337 (December 15, 2023).

Meiklejohn KA, Damaso N, Robertson JM (2019) Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. PLOS ONE 14: e0217084. https://doi.org/10.1371/journal.pone.0217084

Mirtl M, T. Borer E, Djukic I, Forsius M, Haubold H, Hugo W, Jourdan J, Lindenmayer D, McDowell WH, Muraoka H, Orenstein DE, Pauw JC, Peterseil J, Shibata H, Wohner C, Yu X, Haase P (2018) Genesis, goals and achievements of Long-Term Ecological Research at the global scale: A critical review of ILTER and future directions. Science of The Total Environment 626: 1439–1462. https://doi.org/10.1016/j.scitotenv.2017.12.001

Montgomery GA, Belitz MW, Guralnick RP, Tingley MW (2021) Standards and best practices for monitoring and benchmarking insects. Frontiers in Ecology and Evolution 8. Available from: https://www.frontiersin.org/articles/10.3389/fevo.2020.579193 (March 17, 2023).

Pawlowski J, Bruce K, Panksep K, Aguirre FI, Amalfitano S, Apothéloz-Perret-Gentil L, Baussant T, Bouchez A, Carugati L, Cermakova K, Cordier T, Corinaldesi C, Costa FO, Danovaro R, Dell’Anno A, Duarte S, Eisendle U, Ferrari BJD, Frontalini F, Frühe L, Haegerbaeumer A, Kisand V, Krolicka A, Lanzén A, Leese F, Lejzerowicz F, Lyautey E, Maček I, Sagova-Marečková M, Pearman JK,

Pochon X, Stoeck T, Vivien R, Weigand A, Fazi S (2022) Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. Science of The Total Environment 818: 151783. https://doi.org/10.1016/j.scitotenv.2021.151783

Ratnasingham S, Hebert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham S, Hebert PDN (2013) A DNA-Based registry for all animal species: The Barcode Index Number (BIN) system. PLOS ONE 8: e66213. https://doi.org/10.1371/journal.pone.0066213

Resh V, Jackson J (1993) Rapid assessment approaches to biomonitoring using benthic macroinvertebrates. In: , 195–233.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584

Ronquist F, Forshage M, Häggqvist S, Karlsson D, Hovmöller R, Bergsten J, Holston K, Britton T, Abenius J, Andersson B, Buhl PN, Coulianos C-C, Fjellberg A, Gertsson C-A, Hellqvist S, Jaschhof M, Kjærandsen J, Klopfstein S, Kobro S, Liston A, Meier R, Pollet M, Riedel M, Roháček J, Schuppenhauer M, Stigenberg J, Struwe I, Taeger A, Ulefors S-O, Varga O, Withers P, Gärdenfors U (2020) Completing Linnaeus's inventory of the Swedish insect fauna: Only 5,000 species left? PLOS ONE 15: e0228561. https://doi.org/10.1371/journal.pone.0228561

Sickel W, Zizka V, Scherges A, Bourlat SJ, Dieker P (2023) Abundance estimation with DNA metabarcoding – recent advancements for terrestrial arthropods. Metabarcoding and Metagenomics 7: e112290. https://doi.org/10.3897/mbmg.7.112290

Srivathsan A, Ang Y, Heraty JM, Hwang WS, Jusoh WFA, Kutty SN, Puniamoorthy J, Yeo D, Roslin T, Meier R (2023) Convergence of dominance and neglect in flying insect diversity. Nature Ecology & Evolution 7: 1012–1021. https://doi.org/10.1038/s41559-023-02066-0

Steinke D, Braukmann TW, Manerus L, Woodhouse A, Elbrecht V (2021) Effects of Malaise trap spacing on species richness and composition of terrestrial arthropod bulk samples. Metabarcoding and Metagenomics 5: e59201. https://doi.org/10.3897/mbmg.5.59201

Stork NE (2018) How many species of insects and other terrestrial arthropods are there on Earth? Annual Review of Entomology 63: 31–45. https://doi.org/10.1146/annurev-ento-020117-043348

Sturmbauer C, Opadiya GB, Niederstätter H, Riedmann A, Dallinger R (1999) Mitochondrial DNA reveals cryptic oligochaete species differing in cadmium resistance. Molecular Biology and Evolution 16: 967–974. https://doi.org/10.1093/oxfordjournals.molbev.a026186

Telenius A (2011) Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany 29: 378–381. https://doi.org/10.1111/j.1756-1051.2011.01167.x

Uhler J, Redlich S, Zhang J, Hothorn T, Tobisch C, Ewald J, Thorn S, Seibold S, Mitesser O, Morinière J, Bozicevic V, Benjamin CS, Englmeier J, Fricke U, Ganuza C, Haensel M, Riebl R, Rojas-Botero S, Rummler T, Uphus L, Schmidt S, Steffan-Dewenter I, Müller J (2021) Relationship of insect biomass and richness with land use along a climate gradient. Nature Communications 12: 5946. https://doi.org/10.1038/s41467-021-26181-3

Vamos E, Elbrecht V, Leese F (2017) Short COI markers for freshwater macroinvertebrate metabarcoding. Metabarcoding and Metagenomics 1: e14625. https://doi.org/10.3897/mbmg.1.14625

Wagner DL (2020) Insect declines in the Anthropocene. Annual Review of Entomology 65: 457–480. https://doi.org/10.1146/annurev-ento-011019-025151

Welti EAR, Zajicek P, Frenzel M, Ayasse M, Bornholdt T, Buse J, Classen A, Dziock F, Engelmann RA, Englmeier J, Fellendorf M, Förschler MI, Fricke U, Ganuza C, Hippke M, Hoenselaar G, Kaus-Thiel A, Kerner J, Kilian D, Mandery K, Marten A, Monaghan MT, Morkel C, Müller J, Puffpaff S, Redlich S, Richter R, Rojas-Botero S, Scharnweber T, Scheiffarth G, Yáñez PS, Schumann R, Seibold S, Steffan-Dewenter I, Stoll S, Tobisch C, Twietmeyer S, Uhler J, Vogt J, Weis D, Weisser WW, Wilmking M, Haase P (2022) Temperature drives variation in flying insect biomass across a German malaise trap network. Insect Conservation and Diversity 15: 168–180. https://doi.org/10.1111/icad.12555

Wetzel FT, Saarenmaa H, Regan E, Martin CS, Mergen P, Smirnova L, Tuama ÉÓ, García Camacho FA, Hoffmann A, Vohland K, Häuser CL (2015) The roles and contributions of Biodiversity Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case study. Biodiversity 16: 137–149. https://doi.org/10.1080/14888386.2015.1075902

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Wührl L, Pylatiuk C, Giersch M, Lapp F, von Rintelen T, Balke M, Schmidt S, Cerretti P, Meier R (2022) DiversityScanner: Robotic handling of small invertebrates with machine learning methods. Molecular Ecology Resources 22: 1626–1638. https://doi.org/10.1111/1755-0998.13567

Zizka VMA, Elbrecht V, Macher J-N, Leese F (2019) Assessing the influence of sample tagging and library preparation on DNA metabarcoding. Molecular Ecology Resources 19: 893–899. https://doi.org/10.1111/1755-0998.13018

Zizka VMA, Koschorreck J, Khan CC, Astrin JJ (2022a) Long-term archival of environmental samples empowers biodiversity monitoring and ecological research. Environmental Sciences Europe 34: 40. https://doi.org/10.1186/s12302-022-00618-y

Zizka VMA, Geiger MF, Hörren T, Kirse A, Noll NW, Schäffler L, Scherges AM, Sorg M (2022b) Repeated subsamples during DNA extraction reveal increased diversity estimates in DNA metabarcoding of Malaise traps. Ecology and Evolution 12: e9502. https://doi.org/10.1002/ece3.9502

Data accessibility statement

Demultiplexed raw read data for this study has been uploaded to the European Nucleotide Archive under the accession number PRJEB71324.

Author contributions

DB, FL and PH conceived the study. DB performed the analyses and wrote the manuscript with contributions from all authors. DB performed the laboratory processing and the sequence analysis with help from FL and YL. TH and MS checked the taxalist for plausibility. MOG, JM, SUP, SSt, JB, JE, TH, CM, RR, TS, ST, WW, BW, MW and GZ set up Malaise traps in their respective regions, maintained traps and collected samples.

Competing interest

The authors declare no competing interests.