PLOS ONE

LAB PROTOCOL

FAVIS: Fast and versatile protocol for non-destructive metabarcoding of bulk insect samples

Elzbieta Iwaszkiewicz-Eggebrecht[1]*, Piotr Łukasik[2], Mateusz Buczek[2], Junchen Deng[2,3], Emily A. Hartop[4,5], Harald Havnås[4], Monika Prus-Frankowska[2], Carina R. Ugarph[4], Paulina Viteri[1], Anders F. Andersson[6], Tomas Roslin[7], Ayco J. M. Tack[8], Fredrik Ronquist[1‡], Andreia Miraldo[1‡]*

1 Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden, **2** Institute of Environmental Sciences, Faculty of Biology, Jagiellonian University, Kraków, Poland, **3** Doctoral School of Exact and Natural Sciences, Jagiellonian University, Kraków, Poland, **4** Station Linné, Färjestaden, Sweden, **5** Center for Integrative Biodiversity Discovery, Museum für Naturkunde—Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany, **6** Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden, **7** Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden, **8** Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden

‡ FR and AM are equal senior authorship to this work.
* andreia.miraldo@nrm.se (AM); ela.iwaszkiewicz@nrm.se (EIE)

OPEN ACCESS

Citation: Iwaszkiewicz-Eggebrecht E, Łukasik P, Buczek M, Deng J, Hartop EA, Havnås H, et al. (2023) FAVIS: Fast and versatile protocol for non-destructive metabarcoding of bulk insect samples. PLoS ONE 18(7): e0286272. https://doi.org/10.1371/journal.pone.0286272

Editor: Ruslan Kalendar, University of Helsinki: Helsingin Yliopisto, FINLAND

Received: March 21, 2023

Accepted: May 11, 2023

Published: July 19, 2023

Data Availability Statement: Raw data is made available on The Sequence Read Archive (SRA) with accession number PRJNA946790.

Funding: This project was supported by the Knut and Alice Wallenberg Foundation, URL: https://url11.mailanyone.net/scanner?m=1pnO6j-0001w3-3a&d=4%7Cmail%2F90%2F1681496400%2F1pnO6j-0001w3-3a%7Cin11d%7C57e1b682%7C12918722%7C9499237%7C64399A89A36B1B445C9A64A73745266C&o=%

Abstract

Insects are diverse and sustain essential ecosystem functions, yet remain understudied. Recent reports about declines in insect abundance and diversity have highlighted a pressing need for comprehensive large-scale monitoring. Metabarcoding (high-throughput bulk sequencing of marker gene amplicons) offers a cost-effective and relatively fast method for characterizing insect community samples. However, the methodology applied varies greatly among studies, thus complicating the design of large-scale and repeatable monitoring schemes. Here we describe a non-destructive metabarcoding protocol that is optimized for high-throughput processing of Malaise trap samples and other bulk insect samples. The protocol details the process from obtaining bulk samples up to submitting libraries for sequencing. It is divided into four sections: 1) Laboratory workspace preparation; 2) Sample processing—decanting ethanol, measuring the wet-weight biomass and the concentration of the preservative ethanol, performing non-destructive lysis and preserving the insect material for future work; 3) DNA extraction and purification; and 4) Library preparation and sequencing. The protocol relies on readily available reagents and materials. For steps that require expensive infrastructure, such as the DNA purification robots, we suggest alternative low-cost solutions. The use of this protocol yields a comprehensive assessment of the number of species present in a given sample, their relative read abundances and the overall insect biomass. To date, we have successfully applied the protocol to more than 7000 Malaise trap samples obtained from Sweden and Madagascar. We demonstrate the data yield from the protocol using a small subset of these samples.

2Fphta%3A%2Fktsnaw.elebwlne%2Fgro.gr&s=ocsMSCxtEbWPKspOLHYioJenJH8 (grant KAW 2017.088 to FR), Swedish Research Council, URL: https://www.vr.se/english.html (grant 2018-04620 to FR, 2019-04493 to AJMT and 2018-05973 to The Swedish National Infrastructure for Computing (SNIC)), Polish National Agency for Academic Exchange, URL: https://url11.mailanyone.net/scanner?m=1pnO6j-0001w3-3a&d=4%7Cmail%2F90%2F1681496400%2F1pnO6j-0001w3-3a%7Cin11d%7C57e1b682%7C12918722%7C9499237%7C64399A89A36B1B445C9A64A73745266C&o=%2Fphta%3A%2Fntspgwa%2Fv.l.o%26nenbwan%2Faps&s=MHyRVvm2JttgRlGPNhtBu1kbxIU; (grant PPN/PPO/2018/1/00015 to PL) and Polish National Science Centre, URL: https://url11.mailanyone.net/scanner?m=1pnO6j-0001w3-3a&d=4%7Cmail%2F90%2F1681496400%2F1pnO6j-0001w3-3a%7Cin11d%7C57e1b682%7C12918722%7C9499237%7C64399A89A36B1B445C9A64A73745266C&o=%2Fphtw%3A%2Fwtsocw. . .gvnnne%2Flp&s=xp6RyBR_KENx9dKoeP-UipbjmGM (grant 2018/31/B/NZ8/01158 to PL). TR was funded by the European Research Council Synergy, https://url11.mailanyone.net/scanner?m=1pnO6j-0001w3-3a&d=4%7Cmail%2F90%2F1681496400%2F1pnO6j-0001w3-3a%7Cin11d%7C57e1b682%7C12918722%7C9499237%7C64399A89A36B1B445C9A64A73745266C&o=%2Fphtr%3A%2Fetsauc.eop.eraou%2Feepghm&s=d7238XBRw7A7kzQXg7KkyKOMFcA, Grant 856506 (LIFEPLAN) and a Career Support grant from the Swedish University of Agricultural Sciences, https://url11.mailanyone.net/scanner?m=1pnO6j-0001w3-3a&d=4%7Cmail%2F90%2F1681496400%2F1pnO6j-0001w3-3a%7Cin11d%7C57e1b682%7C12918722%7C9499237%7C64399A89A36B1B445C9A64A73745266C&o=%2Fphtw%3A%2Fwtselw.e.s%2Fsu%2Fn&s=B6saXQFQvI-SFi3HzbQiNgfzdfo. The funders did not and will not have a role in study design, data collection and analysis, decision to publish, or preparation of the manuscript."

Competing interests: The authors have declared that no competing interests exist.

Introduction

Insects are key players in ecosystems—they are a crucial part of food webs and provide a wealth of ecosystem functions and services. They are therefore indispensable for the maintenance of natural systems as well as for food production [1]. Insects are also highly diverse with estimates ranging from 4 to 7 million species, which makes them one of the most species-rich groups of animals on Earth [2–4]. However, despite insects' tremendous diversity and ecological importance, our knowledge about them is still fragmentary, and an estimated 75 to 85% of insect species still remain undescribed [5]. Worryingly, recent studies of trends in insect abundance and diversity [6–8] have raised alarm about worldwide insect declines and subsequent threats to the stability of terrestrial ecosystems [9]. Thus, there is a pressing need to speed up efforts in insect diversity discovery and monitoring.

Traditional methods used to study and describe insects involve collecting specimens with a range of different traps, followed by sorting and classification of samples into taxonomic fractions. Whilst the first part of this process—sampling—is relatively straightforward and can be performed by volunteers [10], the latter—taxonomic identification—is complex, demanding specialized knowledge which is in short supply, and can be incredibly time-consuming. For instance, *the Swedish Malaise Trap Project*—an ambitious project aiming to characterize the entire insect fauna of Sweden—collected 1919 bulk insect community samples, containing an estimated 20 million individuals, over three years. Sorting those samples into some 350 taxonomic fractions suitable for processing by specialists took 15 years, despite a considerable investment in manpower [11]. Furthermore, the material identified to species level accounted for only 1% of the total specimen number [12]. In another study, the mapping of insect diversity in a single tropical forest took a decade, and involved the work of 110 taxonomists [13]. Such delays in obtaining results hamper the development of meaningful conservation or protection measures in a timely fashion. For adequate insect diversity discovery, insect community monitoring and real-time study of the spatio-temporal dynamics of insect communities, it is imperative that we develop high-throughput methods for taxonomic processing of insect samples.

DNA-based methods appear particularly well suited to address these high-throughput needs [14]. Reference databases are constantly growing and the cost of sequencing is decreasing, adding to their appeal. Methods such as metagenomics or genome skimming–i.e., filtering of high-copy loci, such as mitochondria, chloroplasts or rRNA, mitochondrial sequences after sequencing–can provide high taxonomic resolution and even promise to provide accurate abundance estimates from bulk samples [15, 16]. However, these techniques still remain prohibitively expensive for most large-scale insect monitoring projects. Metabarcoding—i.e., the amplification of large numbers of barcode sequences from bulk samples—is a cost-effective alternative that has gained popularity in recent years [17–19] and has been successfully applied in arthropod community surveys [14, 20, 21]. Metabarcoding relies on the use of the DNA barcoding technique, developed by Hebert and colleagues [22, 23], in which a short DNA fragment of an individual (i.e., a barcode) can provide us with species-level identification. The standard barcode used in eukaryotic diversity studies is the Folmer region [24] of the mitochondrial cytochrome *c* oxidase subunit 1 (COI) gene, for which vast reference databases exist [25]. In metabarcoding studies, DNA is extracted from bulk, multi-species samples (as derived from e.g. a Malaise trap, or a water or soil sample). Then barcodes are amplified via PCR, sequenced and compared to the reference database for taxonomic identification. The species can be named by matching the barcode to a reference database, providing that the species is represented in the database. Due to high insect diversity and large knowledge gaps, certain taxonomic groups are poorly represented among the references—both because of the lack of

voucher material for described species, and because of a high proportion of undescribed species. Both aspects will contribute to lowering the success of species-level identification. Nevertheless, even for those poorly represented groups, it is still possible to group sequences into clusters based on their genetic similarity, obtain taxonomic assignment for these clusters at higher levels (i.e., order, family or genus), and compare their presence among samples–thereby allowing the efficient characterization of the community composition of the original sample collection.

Despite the great potential of metabarcoding, many methodological questions concerning early stages of sample processing remain open. Perhaps most importantly, the operating procedures for large-scale insect monitoring projects remain motley and poorly documented. In recent years, many different protocols have emerged. Some advocate destructive DNA extraction methods like homogenizing specimens into an "insect soup" [18, 26–28]. Others propose non-destructive mild lysis treatments, in which insects soak in a buffer, gradually releasing their DNA, with minimal damage to specimens [19, 29–31]. The mild lysis treatment yields smaller DNA amounts [32] but is less laborious and preserves specimens for future molecular and taxonomic work [33–36]. Furthermore, it was recently shown that mild lysis also decreases the rate of false negatives during metabarcoding, as the capability to detect small specimens is improved [32].

Each laboratory and institution has to design a workflow best fitting their aims and needs. To aid those searching for a versatile and scalable solution for their purposes, here we present a complete metabarcoding protocol, from insect bulk samples to sequencing data, initially designed for a large-scale insect monitoring project—the *Insect Biome Atlas* (www.insectbiomeatlas.org). The project's field campaign took place in Sweden and Madagascar over 12 months during 2019–2020, and yielded 7398 insect community samples collected with Malaise traps, each sample typically representing one week. All samples were processed using this protocol within 12 months. When adapted and optimized, the wet-lab protocol allows one lab technician to process 180 insect community samples from bulk samples to submission for sequencing in one week, allowing the timely delivery of results.

The use of the protocol and further bioinformatic processing result in a dataset that can be used to produce comprehensive lists of species present in a sample, their relative read abundances, and the overall insect biomass. In defining the protocol, we made efforts to reduce costs and adopt universal reagents and materials that can be easily obtained worldwide. For steps that remain costly or inaccessible, such as DNA purification robots, we suggest alternative low-cost solutions when possible. We opted for a non-destructive lysis protocol with a short incubation time (2h 45 min) in a mild lysis buffer [37] as this allows the efficient processing of a large number of samples per day whilst maximizing the power to recover the original species composition of each sample [32]. In order to introduce a correction factor and allow more accurate estimates of species' abundances, we added to each sample a pre-defined number of biological spike-ins—size-standardized insect species that do not occur in our sampled area (e.g., in the processing of Swedish Malaise trap collection we selected six tropical species that have never been detected in Sweden or neighboring countries). Furthermore, we minimize the damage to specimens and preserve the insect material for further taxonomic or molecular work by returning them to ethanol immediately after the lysis step. Another important aspect of the protocol is the fact that insects never leave the collection bottle, minimizing the risk of cross-sample contamination during sample processing and DNA extraction. The two-step-PCR strategy for COI amplicon library preparation results in double-uniquely indexed libraries obtained using broad-spectrum BF3-BR2 primers [38] with variable-length inserts (phased), reducing cross-contamination through index hopping and increasing signal complexity within the sequencing lane, thus translating to higher quality of results [39].

Materials and methods

The protocol described in this article is published on protocols.io https://www.protocols.io/private/C609E2107CD8B7CFF46EFF1461DBE4C3 and is included for printing as S1 File with this article.

The protocol is divided into four sections. Section 1 (*Preparation*) describes how to prepare workspace and equipment before starting to process samples. The remaining three sections—sections 2 (*Sample Processing*), 3 (*DNA Purification*) and 4 (*Library Preparation and Sequencing*)—cover the main parts of the protocol (**Fig 1**).

To demonstrate the utility of the protocol we summarize the sequence data obtained by processing fifteen Malaise trap samples representing three different habitats in Southern Sweden: a forest, a wetland, and a grassland. From each of these habitats, we present data from samples collected during five consecutive weeks between April and May 2019. The 15 samples presented here were processed as part of our high-throughput sample processing, which involved processing 180 bulk insect samples per week. After completing all steps of the protocol and sequencing on an Illumina NovaSeq 6000 SPrime flow cell, sequencing data was processed bioinformatically following pipelines that can be accessed via links: https://github.com/biodiversitydata-se/amplicon-multi-cutadapt (read trimming and filtering); https://nf-co.re/ampliseq (ASV reconstruction and taxonomic annotation). In short, we use *cutadapt v.3.2* [40] for primer trimming and *R* package *DADA2 v.4.2.1* for denoising [41]. Then we use *SINTAX* [42] in order to get the taxonomic assignment for all ASVs using a custom-made reference COI database (https://doi.org/10.17044/scilifelab.20514192.v4). Krona plots were prepared with the *q2-krona* plug-in from the *qiimeII v.2022.2* library [43, 44]. Visualizations of the results were done with *ggplot2 v.3.4.1* [45] and *ggvenn v.0.1.9* [46] packages in the *R* environment [47]. Non-metric multidimensional scaling (nMDS) was calculated using the metaMDS function from the *vegan v.2.6–4* package [48]. Code used for data manipulation and plotting of the results as well as interactive Krona plots are available on GitHub under https://github.com/ela-iwaszkiewicz/Lab_protocols.git.

Ethics statement

Samples used in this study were covered by Sweden´s right of access to private land (Allemansrätten) and did not necessitate a collection permit. More information about utilizing Swedish genetic resources can be found at the Swedish Environmental Protection Agency website: https://www.naturvardsverket.se/en/guidance/species-protection/utilizing-genetic-resources

Results

The wet weight of the collected insect material ranged from 0.2 to 5.3 g (**Fig 3A**). Concentration of the DNA purified from lysates ranged from 0.7 to 3.8 ng/μl (**Fig 2C**). PCR reactions were successful for all samples included in this study as well as for the positive DNA control (qualitative control of the DNA purification step). Negative controls of PCR I, PCR II, DNA extraction and sample processing (buffer blank) did not show detectable contamination in gel electrophoresis (**Fig 2A and 2B**). In order to even out the amount of each library in the sequencing pool, samples were divided into four categories based on the PCR II product band brightness (**Fig 2C**).

After filtering out amplicon sequence variants (ASVs) that did not have the correct primer sequences, had stop codon(s) in the expected reading frame (possible nuclear mitochondrial DNA segments, i.e. NUMTs), or were of bacterial origin, we ended up with a total of 10,697,001 read-pairs (2 x 250bp). This amounted to an average of 713,133 verified read-pairs per sample (range: 368,060–1,299,458) (**Fig 3A**). These reads represented 1,038 ASVs. For

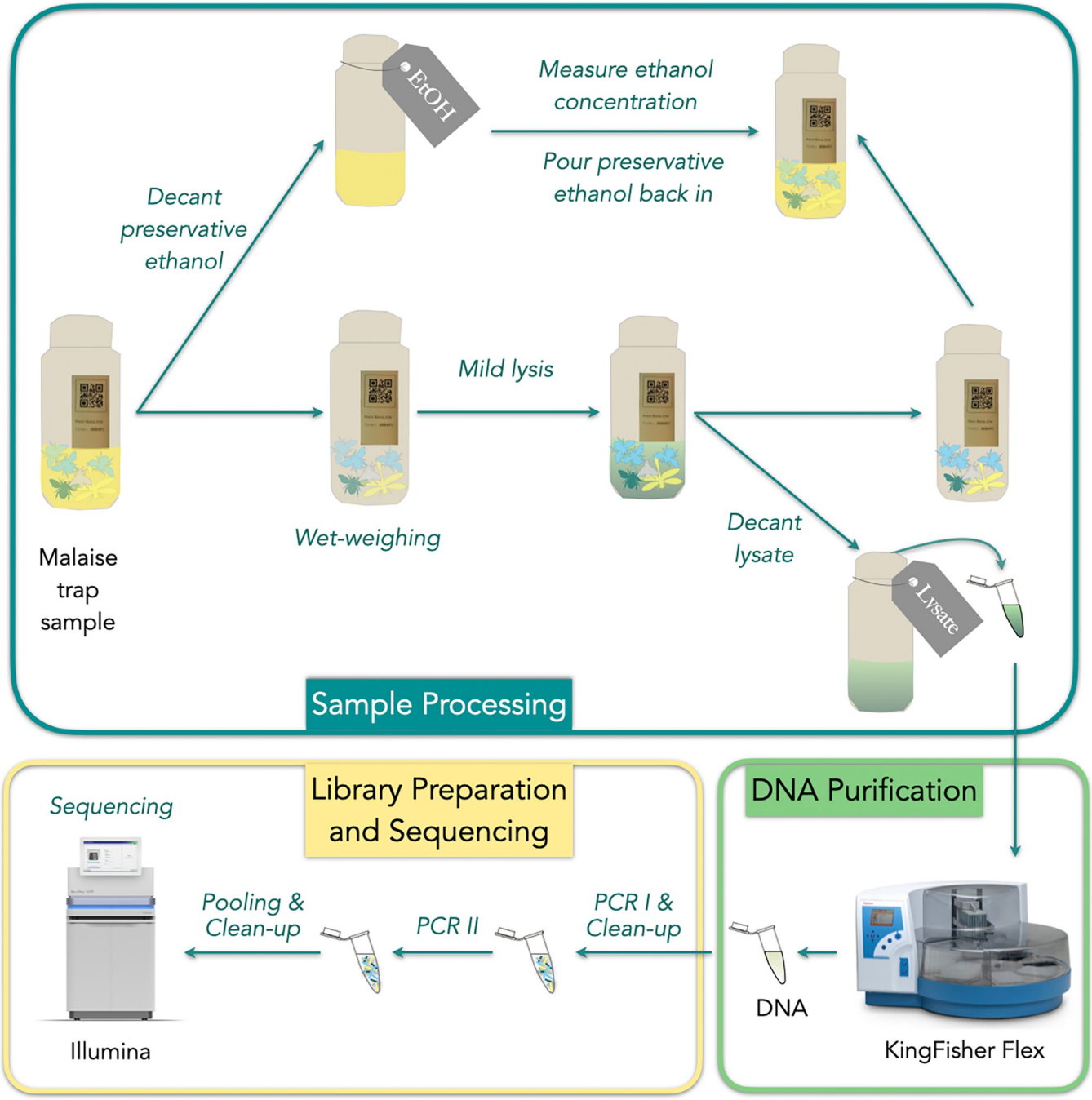


Fig 1. Schematic workflow of FAVIS metabarcoding protocol, comprising three main sections. Sample processing consists of decanting ethanol, measuring ethanol concentration, wet-weighing the sample, adding lysis buffer, incubating the sample, decanting lysates, taking lysate aliquot, and refilling the insect community sample with the previously decanted ethanol. DNA is then extracted and purified from the lysate aliquot with magnetic beads and subsequently used as a template for the amplification of the target COI fragment via PCR (PCR I). After amplification, PCR products are cleaned using magnetic beads and used as template in the second round of PCR (PCR II), where sample-specific tags are added and Illumina adapters completed. The concentration of PCR II products is assessed based on the band brightness on an agarose gel, and all samples are pooled approximately equimolarly to form the sequencing pool. Finally, the pool is purified with magnetic beads and then sequenced on an Illumina NovaSeq platform.

https://doi.org/10.1371/journal.pone.0286272.g001

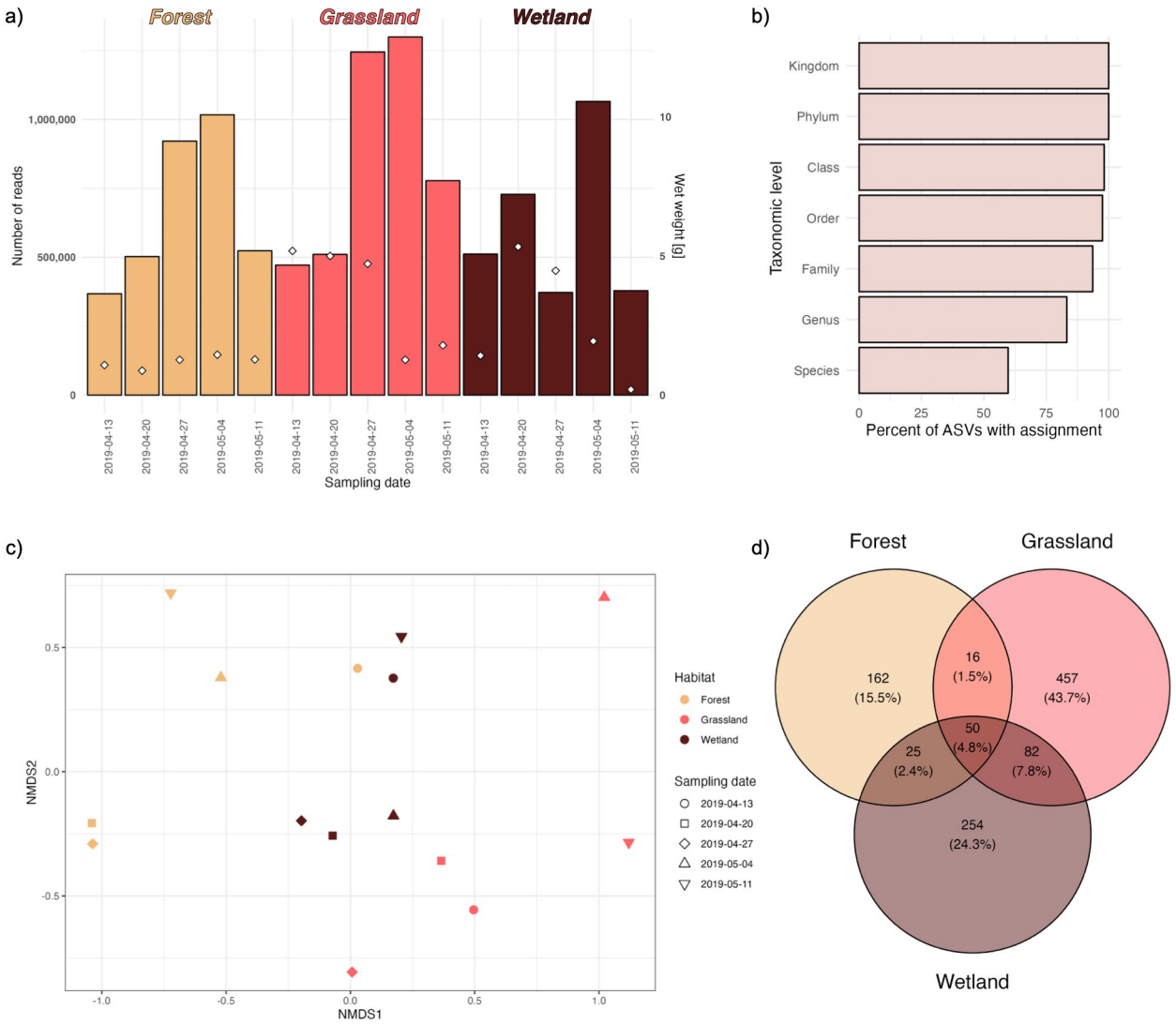
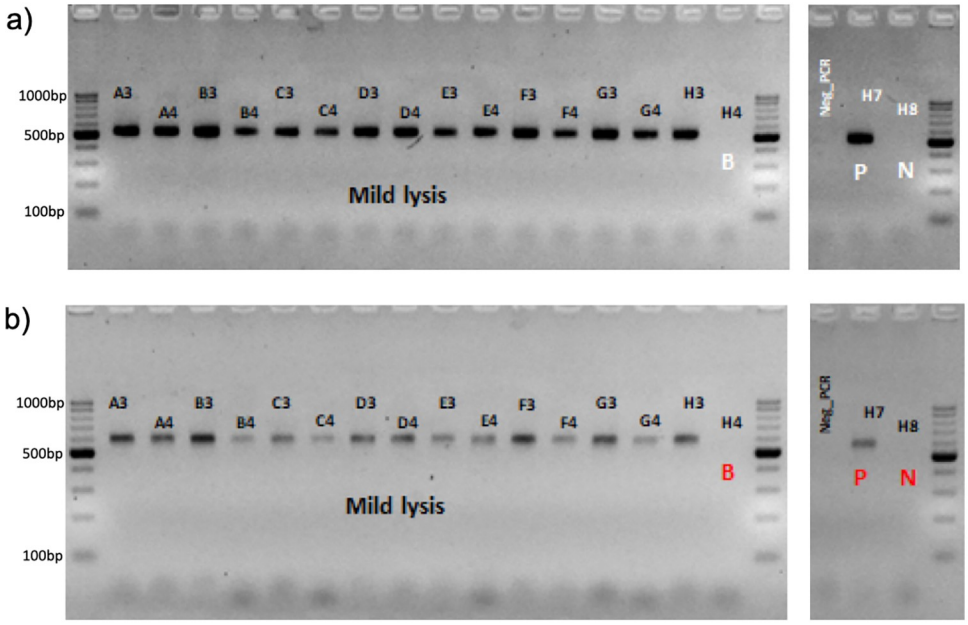


Fig 3. Visual summary of sequencing results. a) Wet weight and sequencing depth for 15 samples. The bars show the number of COI reads that passed our quality filters and were used for further analyses. Colors indicate habitat and white diamonds represent sample wet weight (scale shown on the right y-axis). **b)** Percentages of ASVs with taxonomic assignment at a given rank. **c)** Non-metric multidimensional scaling (nMDS) plot based on pairwise Bray-Curtis distances calculated from relative abundances of ASV sequences.Stress factor associated with the ordination is 0.08. **d)** Venn-diagram displaying numbers and percentages of ASVs unique to the different habitats as well as shared between habitats. Note that the data includes seven ASVs representing the biological spike-in species which are shared between all samples. Therefore, the real number of ASVs shared between habitats is lower, i.e. 43 ASVs.

https://doi.org/10.1371/journal.pone.0286272.g003

93.6% of these we obtained taxonomic assignments at least at the family level, for 83.2% at the genus level and for 59.7% at the species level (Fig 3B). PCR I and PCR II negative controls did not yield any sequencing results. ASVs detected in the sample processing negative control (buffer blank) were inspected but not excluded from the results. In large-scale processing, we recommend using multiple negative controls—i.e. one buffer blank per each day of sample processing, one DNA negative control per extraction batch and one negative control for each PCR reaction plate—and then excluding from the entire dataset all ASVs that are found in more than 5% of negative controls.

Most of the ASVs were found in only one habitat, with Grassland having the highest barcode richness (Fig 3D). Multivariate ordination methods, such as nMDS, allow visual



c)

DNA conc. [ng/µl]		Sample
1.720	A3	Wetland_04/20
1.570	A4	Forest_04/27
3.810	B3	Wetland_04/27
0.704	B4	Forest_04/13
0.928	C3	Grassland_04/20
0.779	C4	Forest_05/11
1.910	D3	Grassland_04/27
1.070	D4	Forest_05/04
0.739	E3	Grassland_04/13
1.440	E4	Forest_04/20
1.330	F3	Grassland_05/11
0.751	F4	Wetland_04/13
2.650	G3	Wetland_05/04
0.847	G4	Wetland_05/11
1.060	H3	Grassland_05/04
0.300	H4	Buffer Blank
3.610	H7	Positive DNA control
	Neg_PCR	Negative PCR I / II control
	H8	Negative DNA control

Fig 2. Electrophoresis gel pictures of PCR products. Panel a) shows products after PCR I and b) after PCR II. Panel c) provides sample codes; cell color represent manually assigned PCR II band brightness scores, which determined the amount of product included in the library pool for sequencing (coral = strong = 1µl, brown = medium = 2µl, beige = weak = 4µl, gray = empty = 8µl). The length of the PCR I product was ca. 490bp and of PCR II product, ca. 563bp. Gel was 2.5% agarose and the DNA ladder used was Perfect Ladder 100-1000bp (EurX, Poland). Original, uncropped gel pictures are provided in S1 and S2 Figs.

https://doi.org/10.1371/journal.pone.0286272.g002

inspection of the data [49]. Here, the distance between different points—samples—reflect their relative similarities and differences in terms of the ASV read counts. The further the samples are situated from one another the more dissimilar they are. In our dataset, the multidimensional ordination of the samples based on beta-diversity (Bray-Curtis distances) revealed moderate grouping of the samples coming from different habitats. Forest and grassland samples were separated along the first nMDS axis, with wetland being intermediate (**Fig 3C**).

Modern data visualization tools for metabarcoding data, like Krona plots [44], allow users to visualize and compare the sample composition at different taxonomic levels (**Fig 4**). Note that the spike-in species—insects added to each sample before processing—are present in all habitats (blue arrows in **Fig 4**) and all samples. Their read abundances are substantial but do not take over, serving as a positive control of the protocol.

The traditional way of comparing community samples is based on relative abundance of taxa (**Fig 5A**). However, areas of active investigation include the reconstruction and comparison of counts or biomass for different species present in the sample. The incorporation in the protocol of both the wet-weighing step and the biological spike-in insect additions provide us with two alternative ways of estimating abundances. First, we can take a look at the abundances of different orders per sample adjusted to the wet-weight of the sample (**Fig 5B**). Alternatively, we can use the information from the spike-in insect reads to calibrate our counts table. We performed the corrections and standardization as suggested in Luo et al. 2023 [50] (summarized in their Fig 1, steps a-d). In other words, we adjusted the observed ASV counts table (i.e. the number of reads per ASV in each sample) by dividing each read count by the number of observed spike-in reads. These calculations were performed with reference to a weighted mean as calculated across all added spike-in specimens (e.g. mean[(*S.lateralis**2 specimens) + (*G.bimaculatus**1) + (*G.sigilatus**1) + (*D.serrata**1) + (*D.jambolina**1) + (*D.*

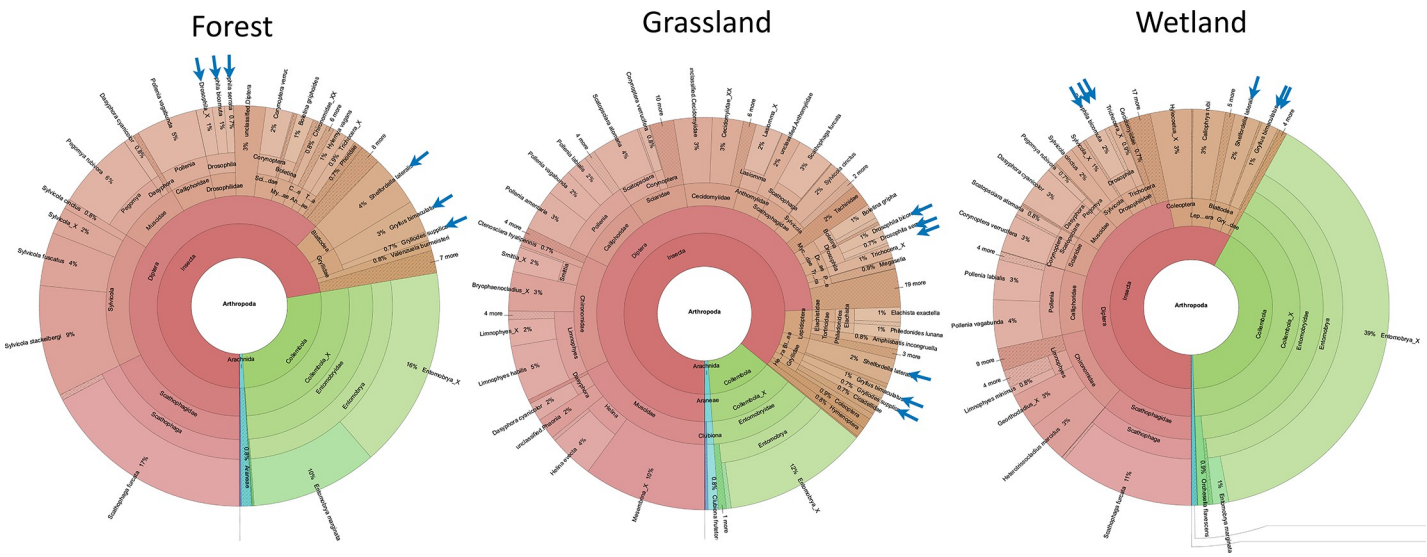


Fig 4. Taxonomic composition of samples collected in three different habitats. The Krona plots show the relative abundance of different taxa in a compound representation of five individual samples pooled per habitat. Blue arrows point to the spike-in species. Interactive versions of the Krona plots are available at GitHub: https://github.com/ela-iwaszkiewicz/Lab_protocols.git.

https://doi.org/10.1371/journal.pone.0286272.g004

*bicornuta**3)]) (Fig 5C). It is important to note that adjusting read counts using wet weight or biological spike-ins can reduce the between-sample variation introduced during lab processing (referred to as "pipeline noise" by Luo et al. 2023 [50]) and allows for improved within-species quantitative comparison across samples, but it does not correct for species-specific biases such as differing DNA yield, preferential PCR amplification etc.

Conclusions

Novel DNA-based methods have the potential to revolutionize biodiversity discovery and monitoring when applied in a high-throughput fashion. Swift processing is crucial for monitoring purposes as well as for informed decision making in conservation efforts.

The metabarcoding protocol described here allows a trained lab technician to process 180 samples (2 x 96-well plates when we include all negative and positive controls), from bulk insect catches to ready-to-sequence libraries, in 7 working days, translating to over 500 Malaise trap samples processed per month (for details see S1 Table). When processing samples at a scale of thousands, an estimated average per-sample reagent cost amounts to about 5 EUR for DNA extraction and purification using homemade magnetic bead solution and 3 EUR for library preparation; additionally, the costs of generating ~1M paired-end reads (2 x 250bp) per sample was about 10 EUR when using NovaSeq 6000 SPrime flow cell. Costs presented here are average costs when implementing the protocol in a high throughput manner, making use of bulk purchase of reagents and consumables, and using home made magnetic beads instead of standardized kits for DNA purification (as described in alternative step 17 in the step-by-step protocol uploaded in protocols.io). Neither these consumables/services costs nor the amount of labor involved (and associated human resources costs) are trivial. However, for large projects addressing grand questions about the biodiversity patterns during times of global change, they are not implausible. Also, there is space for the improvement of time- and cost-efficiency through more extensive use of laboratory automation, or skipping or replacing labor-intensive steps such as agarose gel-based library quality control after both the first and the second PCR.

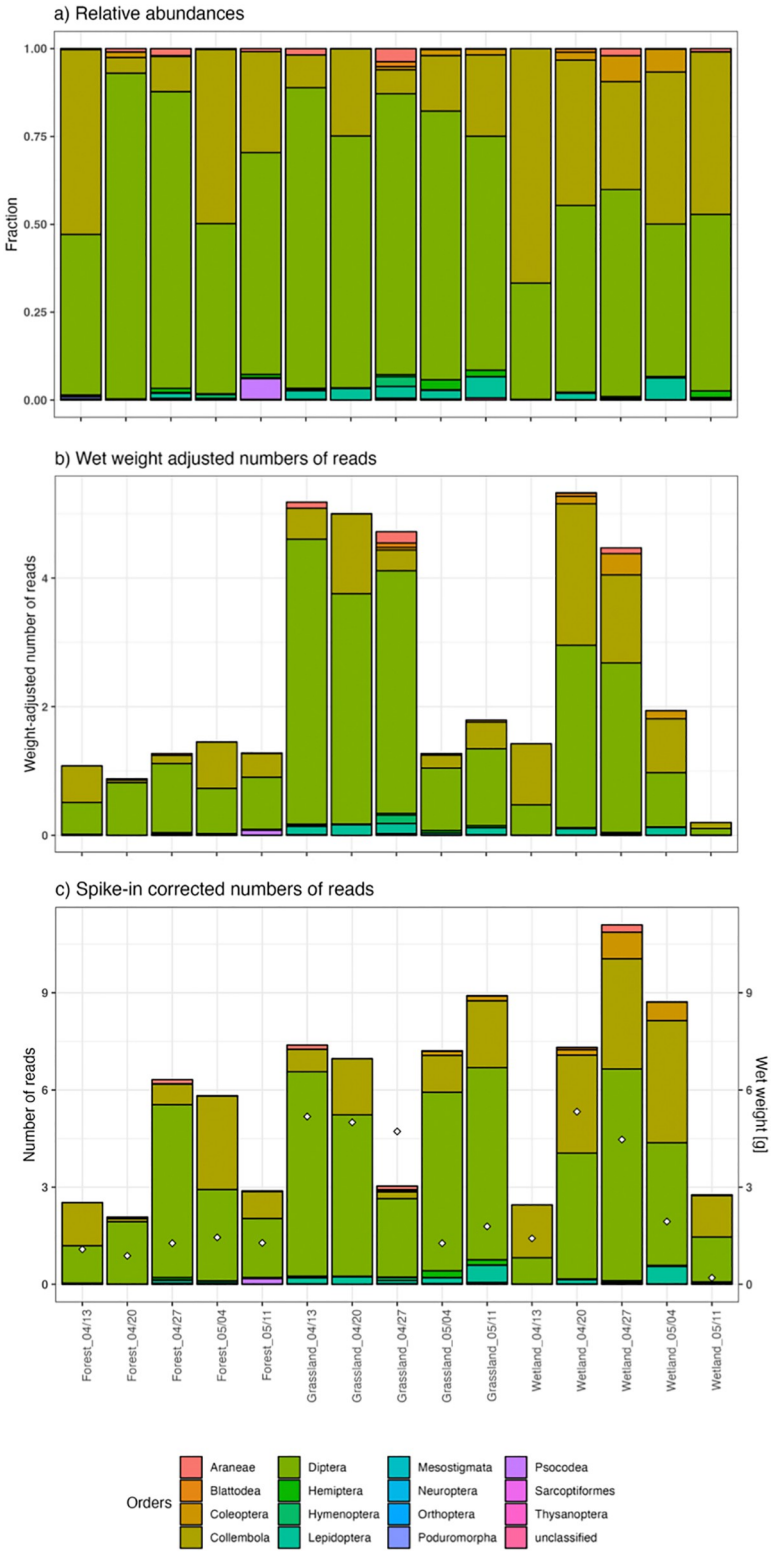


Fig 5. Abundance distributions of insect orders in the samples, using different normalization procedures. a) Relative abundance of ASVs grouped by their taxonomic assignment at the order level. Spike-in ASVs were excluded from the data. **b)** Relative abundance of reads adjusted to each sample's biomass (wet-weight). Spike-in ASVs were excluded from the data. **c)** Spike-in corrected read counts. Colors represent arthropod orders and white diamonds identify the wet weight of the sample (scale shown on the right y-axis).

https://doi.org/10.1371/journal.pone.0286272.g005

We have shown using a small subset of processed samples that using FAVIS results in good quality metabarcoding data that can be used in biodiversity studies and be subject to biological interpretation. While hard to demonstrate in a quantitative manner, we have invested substantial effort in addressing and controlling some of the known methodological challenges including cross-contamination during sample processing and through index hopping, both of which had a measurable effect in our early datasets. The non-destructive nature of the protocol and the retention of specimens post-digestion allows for their future individual characterization using sequencing- or morphology-based studies. At the same time, it is important to pinpoint some of the challenges, likely to become more significant as sample collection and processing accelerates. Among the most important is sample management and tracking. When processing 7000 bulk insect samples from the Insect Biome Atlas project using this protocol, we simplified and streamlined sample management and data recording through the use of QR codes for sample labeling and storage location that are read and registered into a database via a handheld barcode scanner. Another important challenge is the long-term storage of samples and lysates. Those processed as a part of the current project occupy a substantial portion of a custom-build freezer house; but the availability of infrastructure and long-term storage costs could hamper some projects. The third major consideration are the challenges in the analysis and biological interpretation of tremendous amounts of data generated by the project. The bioinformatic workflow presented here is suitable for the analysis of much larger datasets, but dedicated statistical, modeling, and visualization solutions are needed before we can understand the patterns.

Supporting information

S1 Fig. Original, uncropped electrophoresis gel picture underlying Fig 2A from the main text. It shows products of the PCR I—length of the product was ca. 490bp. Yellow frame indicates which parts of the gel were presented in the main text. The Gel was 2.5% agarose and the DNA ladder used was Perfect Ladder 100-1000bp (EurX, Poland).
(PNG)

S2 Fig. Original, uncropped electrophoresis gel picture underlying Fig 2B from the main text. It shows products of the PCR II—length of the product was ca. 563bp. Yellow frame indicates which parts of the gel were presented in the main text. The Gel was 2.5% agarose and the DNA ladder used was Perfect Ladder 100-1000bp (EurX, Poland).
(PNG)

S1 Table. Timeline for the processing of 184 samples using FAVIS protocol.
(PDF)

S1 File. FAVIS protocol downloaded from protocols.io.
(PDF)

S2 File.
(PDF)

Acknowledgments

The authors acknowledge support by NBIS (National Bioinformatics Infrastructure Sweden) as well as from the National Genomics Infrastructure in Stockholm and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

Author Contributions

Conceptualization: Elzbieta Iwaszkiewicz-Eggebrecht, Piotr Łukasik, Anders F. Andersson, Tomas Roslin, Ayco J. M. Tack, Fredrik Ronquist, Andreia Miraldo.

Formal analysis: Elzbieta Iwaszkiewicz-Eggebrecht.

Funding acquisition: Anders F. Andersson, Tomas Roslin, Ayco J. M. Tack, Fredrik Ronquist.

Investigation: Elzbieta Iwaszkiewicz-Eggebrecht, Mateusz Buczek, Harald Havnås, Carina R. Ugarph.

Methodology: Elzbieta Iwaszkiewicz-Eggebrecht, Piotr Łukasik, Mateusz Buczek, Junchen Deng, Emily A. Hartop, Harald Havnås, Monika Prus-Frankowska, Carina R. Ugarph, Paulina Viteri, Andreia Miraldo.

Project administration: Andreia Miraldo.

Supervision: Piotr Łukasik, Andreia Miraldo.

Visualization: Elzbieta Iwaszkiewicz-Eggebrecht.

Writing – original draft: Elzbieta Iwaszkiewicz-Eggebrecht.

Writing – review & editing: Elzbieta Iwaszkiewicz-Eggebrecht, Piotr Łukasik, Mateusz Buczek, Junchen Deng, Emily A. Hartop, Harald Havnås, Carina R. Ugarph, Paulina Viteri, Anders F. Andersson, Tomas Roslin, Ayco J. M. Tack, Fredrik Ronquist, Andreia Miraldo.

References

1. Jordan A, Patch HM, Grozinger CM, Khanna V. Economic Dependence and Vulnerability of United States Agricultural Sector on Insect-Mediated Pollination Service. Environ Sci Technol. 2021 Feb 16; 55(4):2243–53. https://doi.org/10.1021/acs.est.0c04786 PMID: 33496588
2. Hamilton AJ, Basset Y, Benke KK, Grimbacher PS, Miller SE, Novotný V, et al. Quantifying Uncertainty in Estimation of Tropical Arthropod Species Richness. Am Nat. 2010 Jul; 176(1):90–5. https://doi.org/10.1086/652998 PMID: 20455708
3. Novotny V, Basset Y, Miller SE, Weiblen GD, Bremer B, Cizek L, et al. Low host specificity of herbivorous insects in a tropical forest. Nature. 2002 Apr; 416(6883):841–4. https://doi.org/10.1038/416841a PMID: 11976681
4. Stork NE, McBroom J, Gely C, Hamilton AJ. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. Proc Natl Acad Sci. 2015 Jun 16; 112(24):7519–23. https://doi.org/10.1073/pnas.1502408112 PMID: 26034274
5. Stork NE. How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? Annu Rev Entomol. 2018 Jan 7; 63(1):31–45. https://doi.org/10.1146/annurev-ento-020117-043348 PMID: 28938083
6. Hallmann CA, Sorg M, Jongejans E, Siepel H, Hofland N, Schwan H, et al. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. PLOS ONE. 2017 Oct 18; 12(10): e0185809. https://doi.org/10.1371/journal.pone.0185809 PMID: 29045418
7. Klink R van Bowler DE, Gongalsky KB Swengel AB, Gentile A Chase JM. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. Science. 2020 Apr 24; 368 (6489):417–20. https://doi.org/10.1126/science.aax9931 PMID: 32327596
8. Seibold S, Gossner MM, Simons NK, Blüthgen N, Müller J, Ambarlı D, et al. Arthropod decline in grasslands and forests is associated with landscape-level drivers. Nature. 2019 Oct; 574(7780):671–4. https://doi.org/10.1038/s41586-019-1684-3 PMID: 31666721
9. Cardoso P, Barton PS, Birkhofer K, Chichorro F, Deacon C, Fartmann T, et al. Scientists' warning to humanity on insect extinctions. Biol Conserv. 2020 Feb 1; 242:108426.
10. Ballard HL, Robinson LD, Young AN, Pauly GB, Higgins LM, Johnson RF, et al. Contributions to conservation outcomes by natural history museum-led citizen science: Examining evidence and next steps. Biol Conserv. 2017 Apr 1; 208:87–97.

11. Karlsson D, Hartop E, Forshage M, Jaschhof M, Ronquist F. The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. Biodivers Data J [Internet]. 2020 Jan 21 [cited 2021 Mar 29]; 8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6987249/

12. Ronquist F, Forshage M, Häggqvist S, Karlsson D, Hovmöller R, Bergsten J, et al. Completing Linnaeus's inventory of the Swedish insect fauna: Only 5,000 species left? PLOS ONE. 2020 Mar 4; 15(3): e0228561. https://doi.org/10.1371/journal.pone.0228561 PMID: 32130216

13. Basset Y, Cizek L, Cuénoud P, Didham RK, Guilhaumon F, Missa O, et al. Arthropod Diversity in a Tropical Forest. Science. 2012 Dec 14; 338(6113):1481–4. https://doi.org/10.1126/science.1226727 PMID: 23239740

14. van Klink R, August T, Bas Y, Bodesheim P, Bonn A, Fossøy F, et al. Emerging technologies revolutionise insect ecology and monitoring. Trends Ecol Evol. 2022 Oct 1; 37(10):872–85. https://doi.org/10.1016/j.tree.2022.06.001 PMID: 35811172

15. Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, et al. SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Mol Ecol Resour. 2020; 20(1):256–67. https://doi.org/10.1111/1755-0998.13057 PMID: 31293086

16. Kennedy SR, Prost S, Overcast I, Rominger AJ, Gillespie RG, Krehenwinkel H. High-throughput sequencing for community analysis: the promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities. Dev Genes Evol. 2020; 230(2):185–201. https://doi.org/10.1007/s00427-020-00652-x PMID: 32040713

17. Beng KC, Tomlinson KW, Shen XH, Surget-Groba Y, Hughes AC, Corlett RT, et al. The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. Sci Rep. 2016 Apr 26; 6(1):24965. https://doi.org/10.1038/srep24965 PMID: 27112993

18. Liu M, Clarke LJ, Baker SC, Jordan GJ, Burridge CP. A practical guide to DNA metabarcoding for entomological ecologists. Ecol Entomol. 2020; 45(3):373–85.

19. Martoni F, Piper AM, Rodoni BC, Blacket MJ. Disentangling bias for non-destructive insect metabarcoding. PeerJ. 2022 Feb 23; 10:e12981. https://doi.org/10.7717/peerj.12981 PMID: 35228909

20. Cristescu ME. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol Evol. 2014 Oct 1; 29(10):566–71. https://doi.org/10.1016/j.tree.2014.08.001 PMID: 25175416

21. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol. 2012; 21(8):2045–50. https://doi.org/10.1111/j.1365-294X.2012.05470.x PMID: 22486824

22. Hebert PDN, Ratnasingham S, deWaard JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc B Biol Sci. 2003 Aug 7; 270(Suppl 1):S96–9. https://doi.org/10.1098/rsbl.2003.0025 PMID: 12952648

23. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci. 2003 Feb 7; 270(1512):313–21. https://doi.org/10.1098/rspb.2002.2218 PMID: 12614582

24. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol. 1994;(Oct; 3 (5)):294–9. PMID: 7881515

25. Ratnasingham S, Hebert PDN. bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes. 2007; 7(3):355–64.

26. Elbrecht V, Leese F. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. PLOS ONE. 2015 Jul 8; 10(7):e0130324. https://doi.org/10.1371/journal.pone.0130324 PMID: 26154168

27. Morinière J, Araujo BC de, Lam AW, Hausmann A, Balke M, Schmidt S, et al. Species Identification in Malaise Trap Samples by DNA Barcoding Based on NGS Technologies and a Scoring Matrix. PLOS ONE. 2016 May 18; 11(5):e0155497. https://doi.org/10.1371/journal.pone.0155497 PMID: 27191722

28. Zizka VM, Geiger MF, Hörren T, Kirse A, Noll NW, Schäffler L, et al. Recommendations for tissue homogenisation and extraction in DNA metabarcoding of Malaise trap samples [Internet]. bioRxiv; 2022 [cited 2023 Feb 20]. p. 2022.01.25.477667. Available from: https://www.biorxiv.org/content/10.1101/2022.01.25.477667v1

29. Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, et al. Metabarcoding a diverse arthropod mock community. Mol Ecol Resour. 2019; 19(3):711–27. https://doi.org/10.1111/1755-0998.13008 PMID: 30779309

30. Porco D, Rougerie R, Deharveng L, Hebert P. Coupling non-destructive DNA extraction and voucher retrieval for small soft-bodied Arthropods in a high-throughput context: the example of Collembola. Mol Ecol Resour. 2010; 10(6):942–5. https://doi.org/10.1111/j.1755-0998.2010.2839.x PMID: 21565103

31. Steinke D, deWaard SL, Sones JE, Ivanova NV, Prosser SWJ, Perez K, et al. Message in a Bottle—Metabarcoding enables biodiversity comparisons across ecoregions. GigaScience. 2022 Jan 1; 11: giac040. https://doi.org/10.1093/gigascience/giac040 PMID: 35482490

32. Iwaszkiewicz-Eggebrecht E, Granqvist E, Buczek M, Prus M, Kudlicka J, Roslin T, et al. Optimizing insect metabarcoding using replicated mock communities. Methods Ecol Evol. 2023; 14(4):1130–46.

33. Batovska J, Piper AM, Valenzuela I, Cunningham JP, Blacket MJ. Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches. Sci Rep. 2021 Apr 12; 11(1):7946. https://doi.org/10.1038/s41598-021-85855-6 PMID: 33846382

34. Carew ME, Coleman RA, Hoffmann AA. Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? PeerJ. 2018 Jun 13; 6:e4980. https://doi.org/10.7717/peerj.4980 PMID: 29915700

35. Marquina D, Roslin T, Łukasik P, Ronquist F. Evaluation of non-destructive DNA extraction protocols for insect metabarcoding: gentler and shorter is better. Metabarcoding Metagenomics. 2022 Jun 16; 6: e78871.

36. Nielsen M, Gilbert MTP, Pape T, Bohmann K. A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. Environ DNA. 2019; 1(2):144–54.

37. Vesterinen EJ, Ruokolainen L, Wahlberg N, Peña C, Roslin T, Laine VN, et al. What you need is what you eat? Prey selection by the bat Myotis daubentonii. Mol Ecol. 2016; 25(7):1581–94. https://doi.org/10.1111/mec.13564 PMID: 26841188

38. Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, et al. Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ. 2019 Oct 7; 7:e7745. https://doi.org/10.7717/peerj.7745 PMID: 31608170

39. Wu L, Wen C, Qin Y, Yin H, Tu Q, Van Nostrand JD, et al. Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. BMC Microbiol. 2015 Jun 19; 15(1):125. https://doi.org/10.1186/s12866-015-0450-4 PMID: 26084274

40. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2; 17(1):10–2.

41. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul; 13(7):581–3. https://doi.org/10.1038/nmeth.3869 PMID: 27214047

42. Edgar RC. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences [Internet]. bioRxiv; 2016 [cited 2023 Feb 14]. p. 074161. Available from: https://www.biorxiv.org/content/10.1101/074161v1

43. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019 Aug; 37 (8):852–7. https://doi.org/10.1038/s41587-019-0209-9 PMID: 31341288

44. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12(1):385.

45. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

46. Yan L. Ggvenn: Draw Venn Diagram by Ggplot2 [Internet]. 2021. Available from: https://CRAN.R-project.org/package=ggvenn

47. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/

48. Oksanen J, Blanchet G, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package [Internet]. 2020. Available from: https://CRAN.R-project.org/package=vegan

49. Zuur AF, Ieno EN, Smith GM, editors. Principal coordinate analysis and non-metric multidimensional scaling. In: Analysing Ecological Data [Internet]. New York, NY: Springer; 2007 [cited 2023 Feb 6]. p. 259–64. (Statistics for Biology and Health). Available from: https://doi.org/10.1007/978-0-387-45972-1_15

50. Luo M, Ji Y, Warton D, Yu DW. Extracting abundance information from DNA-based data. Mol Ecol Resour. 2023; 23(1):174–89. https://doi.org/10.1111/1755-0998.13703 PMID: 35986714