OXFORD (GIGA)n SCIENCE

GigaScience, 2022, **11**, 1–12
DOI: 10.1093/gigascience/giac065
Review

Toward global integration of biodiversity big data: a harmonized metabarcode data generation module for terrestrial arthropods

Paula Arribas [1,*], Carmelo Andújar [1], Kristine Bohmann [2], Jeremy R. deWaard[3,4], Evan P. Economo[5], Vasco Elbrecht [6], Stefan Geisen [7], Marta Goberna [8], Henrik Krehenwinkel[9], Vojtech Novotny [10,11], Lucie Zinger [12,13], Thomas J. Creedy [14], Emmanouil Meramveliotakis[15], Víctor Noguerales [1], Isaac Overcast [12], Hélène Morlon [12], Anna Papadopoulou [15], Alfried P. Vogler [14,16] and Brent C. Emerson [1]

[1]Island Ecology and Evolution Research Group, Institute of Natural Products and Agrobiology (IPNA-CSIC), 38206 San Cristóbal de la Laguna, Spain
[2]Section for Evolutionary Genomics, Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark
[3]Centre for Biodiversity Genomics, University of Guelph, N1G2W1 Guelph, Canada
[4]School of Environmental Sciences, University of Guelph, N1G2W1 Guelph, Canada
[5]Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology Graduate University, 904-0495 Japan
[6]Centre for Biodiversity Monitoring (ZBM), Zoological Research Museum Alexander Koenig, D-53113 Bonn, Germany
[7]Laboratory of Nematology, Department of Plant Sciences, Wageningen University and Research, 6708PB Wageningen, The Netherlands
[8]Department of Environment and Agronomy, INIA-CSIC, 28040 Madrid, Spain
[9]Department of Biogeography, Trier University, D-54296 Trier, Germany
[10]Biology Centre, Czech Academy of Sciences, Institute of Entomology, 37005 Ceske Budejovice, Czech Republic
[11]Faculty of Science, University of South Bohemia, 37005 Ceske Budejovice, Czech Republic
[12]Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[13]Naturalis Biodiversity Center, 2300 RA Leiden, The Netherlands
[14]Department of Life Sciences, Natural History Museum, SW7 5BD London, UK
[15]Department of Biological Sciences, University of Cyprus, 1678 Nicosia, Cyprus
[16]Department of Life Sciences, Imperial College London, SW7 2AZ London, UK
*Correspondence address. Paula Arribas, Institute of Natural Products and Agrobiology (IPNA-CSIC), 38206 San Cristobal de la Laguna, Spain. E-mail: pauarribas@ipna.csic.es

Abstract

Metazoan metabarcoding is emerging as an essential strategy for inventorying biodiversity, with diverse projects currently generating massive quantities of community-level data. The potential for integrating across such data sets offers new opportunities to better understand biodiversity and how it might respond to global change. However, large-scale syntheses may be compromised if metabarcoding workflows differ from each other. There are ongoing efforts to improve standardization for the reporting of inventory data. However, harmonization at the stage of generating metabarcode data has yet to be addressed. A modular framework for harmonized data generation offers a pathway to navigate the complex structure of terrestrial metazoan biodiversity. Here, through our collective expertise as practitioners, method developers, and researchers leading metabarcoding initiatives to inventory terrestrial biodiversity, we seek to initiate a harmonized framework for metabarcode data generation, with a terrestrial arthropod module. We develop an initial set of submodules covering the 5 main steps of metabarcode data generation: (i) sample acquisition; (ii) sample processing; (iii) DNA extraction; (iv) polymerase chain reaction amplification, library preparation, and sequencing; and (v) DNA sequence and metadata deposition, providing a backbone for a terrestrial arthropod module. To achieve this, we (i) identified key points for harmonization, (ii) reviewed the current state of the art, and (iii) distilled existing knowledge within submodules, thus promoting best practice by providing guidelines and recommendations to reduce the universe of methodological options. We advocate the adoption and further development of the terrestrial arthropod module. We further encourage the development of modules for other biodiversity fractions as an essential step toward large-scale biodiversity synthesis through harmonization.

Keywords: metabarcoding, arthropods, harmonization, data generation, modular structure, biodiversity inventory, biodiversity big data integration, reproducibility, comparability

Background

DNA metabarcoding, involving polymerase chain reaction (PCR)–coupled high-throughput sequencing (HTS) directly from bulk or environmental samples, represents the most cost-efficient approach for obtaining molecular community profiles [1, 2]. Metabarcoding is increasingly being used to characterize and monitor biodiversity and is recognized as a substantial advance leading to a step change in multiple fields of biodiversity science (e.g., [3–5]). Diverse projects, from local to global scales, are currently generating massive quantities of site-based community-level biodiversity inventory data, including hyperdiverse assemblages or groups for which classical sampling and identification is overly complicated and time-consuming. The potential for integrating across such data, from diverse sources and time series,

Received: December 22, 2021. **Revised:** May 4, 2022. **Accepted:** June 2, 2022
© The Author(s) 2022. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

offers new opportunities to better understand how biodiversity is structured in space and time, and the factors that regulate it. Additionally, such integration can be leveraged for better monitoring and the development of holistic biodiversity conservation strategies, in response to global change [4, 6, 7]. However, collective international efforts are required to achieve optimal global integration and synthesis. While integrative efforts for harmonized site-based genomic inventories exist in the microbial realm (e.g., [8–10]), such a framework has yet to be extended to nonmicrobial fractions of biodiversity. However, there is an emerging consensus that such integration can be achieved within an HTS framework, analogous to the Genomic Observatories (GO) concept, first proposed by Davies et al. [11, 12]. If effective strategies can be developed to harmonize the data resulting from metabarcoding studies (i.e., metabarcode inventory data), these can potentially scale up to a noncentralized network within which global patterns and trends of biodiversity can be addressed [13].

There are ongoing efforts to maximize the potential for integrating across independent biodiversity data sets through improved standardization for the reporting of inventory data (Humboldt Core: [14]). In the case of molecular data specifically, the GEOME initiative [15, 16] promotes standardization for the reporting of taxonomic, genomic, and metadata through customizable yet standard-compliant spreadsheets that capture the temporal and geospatial context of a biosample. While recommendations have been made for the harmonization of bioinformatic processing of raw metabarcode read data from metazoan biodiversity fractions [17], harmonization at the stage of generating such metabarcode data has yet to be addressed and thus remains a fundamental impediment for data integration. The success of global microbial diversity assessment initiatives has pivoted on standardized metabarcoding protocols for sampling, DNA extraction, barcode amplification/enrichment, and library generation and sequencing of microbial/planktonic communities (e.g., [18, 19] for the Earth Microbiome Project or [20–22] for the TARA Oceans and the Ocean Sampling Day). Despite pioneering efforts to harmonize metabarcode data generation beyond microbial biodiversity fractions (e.g., see [23, 24]), further efforts are required within this expanding research area.

A Harmonized Framework for the Generation of Metabarcode Data for Terrestrial Animals

Terrestrial metazoans constitute one of the most heterogeneous groups in terms of body size across the tree of life. Metabarcoding is emerging as an important approach for the inventorying of metazoan diversity and is increasingly being used across the fields of community ecology, evolutionary ecology, biogeography, conservation biology, and environmental management, among others. Given the rapid development of data generation in this area, the potential for downstream synthesis across independently generated data sets may be compromised if divergent strategies are being implemented. There is already concern that nuances in metabarcoding workflows make comparisons difficult (e.g., [25–28]). Guidance for the implementation of effective and robust sampling and sample-processing approaches is both timely and essential and will increase the potential for broader benefits to biodiversity science through harmonization. We believe that the overarching goal of a harmonized metabarcode framework for inventorying biodiversity should be to reduce unnecessary heterogeneity in the generation of metabarcode data, thus facilitating comparability and integration among independent metabarcode data sets. The development and implementation of consistent workflows for data generation is a key step for the bottom-up growth of a GO network for global integration and synthesis within biodiversity science, while the challenge is to also allow flexibility to successfully address objectives at the individual project level.

It has previously been argued that a harmonized framework with a "modular" structure for data generation could offer a pathway to navigate through the complex structure of terrestrial metazoan biodiversity, by placing different fractions of terrestrial diversity at the core of each "module" [13]. Within such a framework, best practices and harmonized protocols for the generation of metabarcode data can be developed for different target fractions of biodiversity (e.g., terrestrial arthropods). Within individual modules, submodules serve as the fundamental building blocks that provide guidelines and recommendations for the 5 key steps to generate metabarcode data: (i) sample acquisition; (ii) sample processing; (iii) DNA extraction; (iv) PCR amplification, library preparation, and sequencing; and (v) DNA sequence and metadata deposition. Different data generation pipelines can be configured within a module by choosing among submodule options, allowing for variable requirements of different assemblages within the module (e.g., flying, aquatic, or ground arthropods within a terrestrial arthropod module) and different sample vouchering needs (e.g., destructive vs. nondestructive DNA extraction). Such a modular structure provides a harmonized framework for comparability across independent studies, by reducing redundant efforts and improving reporting and comparability, while retaining flexibility to incorporate additional submodules as the need arises (see Fig. 1, a schematic representation of the proposed modular structure).

Here, through our collective expertise as practitioners of metabarcoding, method developers, and researchers leading metabarcoding initiatives to inventory terrestrial arthropod biodiversity, we seek to initiate a harmonized framework for the generation of terrestrial metazoan metabarcode data. Specifically, we aim to provide an initial set of submodules (black blocks in Fig. 1) covering the 5 main steps of metabarcode data generation (rows 1 to 5 in Fig. 1) that constitute the backbone of a terrestrial arthropod module (red block in Fig. 1). We first (i) identify key points for harmonization within each of the 5 steps, (ii) review the current state of the art within the arthropod metabarcoding literature, and (iii) distill existing information and knowledge within submodules, thus promoting best practice by providing guidelines and recommendations to reduce the universe of methodological options. Standardization or harmonization of methods will, in some contexts, lead to trade-offs against what might be considered perfect methods [29]. Such trade-offs may limit the uptake of harmonized protocols, thus compromising the discovery of unifying principles from analyses synthesizing across comparable studies. Thus, rather than being overly prescriptive, we seek to propose a flexible framework that can be opted into with minimal compromise, to increase the comparative value of metabarcode data.

Harmonization for the Metabarcoding of Terrestrial Arthropods: The Terrestrial Arthropods Module

There are multiple reasons why techniques for inventorying and monitoring terrestrial arthropod biodiversity are urgently needed. First, arthropods comprise the majority of known animal species

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

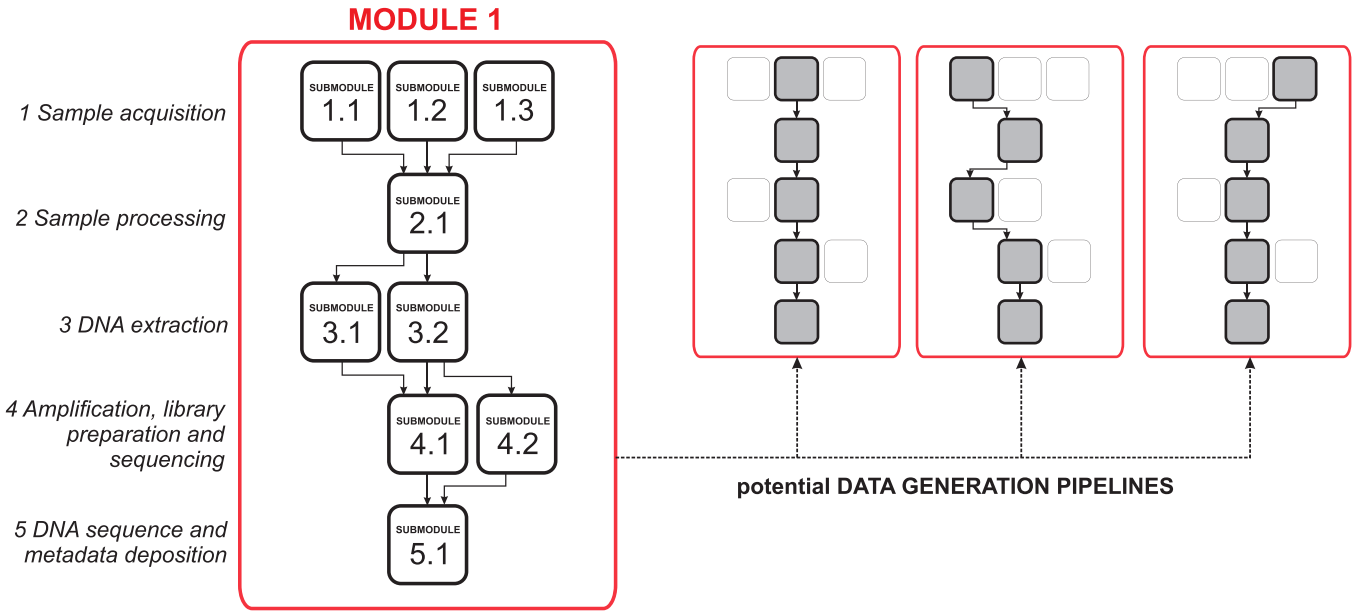


Figure 1: A harmonized framework with a "modular" structure for metazoan metabarcoding. Schematic representation of the modular structure proposed for building a harmonized framework for the generation of metabarcode data for different fractions of terrestrial animals. Different fractions of terrestrial animal diversity are at the core of each "module" (red rectangle, e.g., the terrestrial arthropods module), and within such a framework, best practices and harmonized protocols are developed as submodules (black blocks). Submodules within each module serve as the fundamental building blocks that provide guidelines and recommendations for the 5 well-defined steps for generating metabarcode data (left panel, rows 1 to 5). Within this framework, tailored data generation pipelines can be configured within a module, drawn from the set of alternative submodules.

in terrestrial habitats. It has been estimated that there are 5.5 million insect species on Earth, most yet to be discovered, and up to 6.8 million species (range, 5.9–7.8 million) for all terrestrial arthropods [30]. In addition to this high diversity, arthropods present vast trait variation, which imposes a substantial challenge for assessing their responses to environmental change. We now face the challenge of declining arthropod abundance and richness, a very real and serious threat that society must urgently address [31, 32]. Arthropods are also a key biodiversity fraction for monitoring because they include many invasive species [33], requiring comparable baseline data to study the potential susceptibility and responses of communities to invasion. DNA metabarcoding has emerged as a powerful approach for characterizing complex, and in many cases largely unknown, arthropod assemblages [7, 34]. In response to this, researchers from diverse disciplines are shifting from conventional inventorying of arthropod diversity to DNA metabarcoding, with evidence for exponential growth uptake [17]. Indeed, adaptations of microbial metabarcoding approaches to the macroscopic component of diversity have been heavily influenced by their application to the arthropod fauna (see [1, 35] for pioneering studies). Metabarcoding of DNA extracted from bulk samples of whole organisms (whole-organism community DNA, wocDNA) is (i) the most common and straightforward metabarcoding approach to inventory arthropod biodiversity, (ii) comparable to standard methods of arthropod monitoring, and (iii) has high potential for harmonization [27].

Data generation practices for the metabarcoding of arthropod community samples are still in the early stages. Through the development and adoption of a standardized terrestrial arthropod data generation module, the potential for comparability across future large-scale biodiversity inventorying efforts can be optimized. There is sufficient background from which recommendations can be developed (e.g., [36–40]) to guide methodological decisions within the emerging research community. Recent global initiatives that pivot on arthropod wocDNA also provide a critical mass for developing harmonized data generation, while simultaneously highlighting the relevance and timeliness of a terrestrial arthropod module. These initiatives include the BIOSCAN initiative (https://ibol.org/programs/bioscan/) and its regional extensions such as BIOSCAN Europe (https://www.bioscaneurope.org/), BioAlfa, the Kruger Malaise Program [41], the SITE-100 project (https://www.site100.org/), the Insect Biome Atlas Project (https://insectbiomeatlas.org), LIFEPLAN (https://www.helsinki.fi/en/projects/lifeplan), and the OKEON initiative (https://okeon.unit.oist.jp/).

Identifying Key Points of Harmonization for Submodules within Each Data Generation Step

Sample acquisition step

A starting point for integration across independent biodiversity inventory efforts is a harmonized sample definition. In the case of terrestrial arthropods, sample definition is strongly linked to the sampling technique implemented. There is extensive evidence that different arthropod mass sampling techniques have differing capture efficiencies with regard to total community assemblages within which they are deployed, with no one method detecting the entire arthropod diversity within a site [42]. In this context, with the aim of standardizing insect inventorying and monitoring methods, Montgomery et al. [43] proposed 7 main sampling methods with the aim of maximizing data integration across insect monitoring efforts, including (i) Malaise trapping, (ii) light trapping, (iii) pan trapping, (iv) pitfall trapping, (v) beating sheets, (vi) acoustic monitoring, and (vii) active visual surveys. These complementary sampling methods provide an appropriate platform from which to develop sample acquisition submodules, which could be implemented individually or combined for more complex sampling designs.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

Most implementations of wocDNA metabarcoding to date are Malaise trap based, at scales ranging from local to global (e.g., [44–49]). Additionally, Malaise traps are frequently deployed together with other sampling techniques to generate plot-based arthropod inventory data (e.g., [50], SITE100, ForestGEO arthropod protocol) and are the sampling strategy of the Global Malaise Trap Program/BIOSCAN initiative [44], with more than 10,000 samples already generated worldwide. Malaise traps [51] are primarily effective for sampling flying insects (e.g., [52]) but have gained popularity for assessing terrestrial arthropod communities (e.g., [53]) and have been proposed as ideal for insect biomonitoring using metabarcoding [43, 50]. Once installed, they require limited effort and can yield clean samples comprising almost exclusively arthropods and in very large numbers (up to 10,000 specimens per week in some cases). Moreover, they can remain in place and yield new samples through passive sampling with low handling time, making them suitable for time-resolved monitoring. Given these considerations, Malaise traps are an obvious sampling submodule candidate.

Following the recommendations of Montgomery et al. [43], together with operational procedures adopted within the BIOSCAN initiative (https://biodiversitygenomics.net/resources/bioscan), Townes-style Malaise traps are preferred, with a 165 × 110 cm interception area being most common and 95% ethanol as the preservation agent (see [50]) but propylene glycol (ratio of 50–100% propylene glycol with water is frequently recommended as evaporation is negligible compared to ethanol and adequately preserves DNA [54, 55]). Sampling effort has typically been delimited to 1 week within most metabarcoding studies, representing a compromise between maximizing sampling effort and reducing potential problems with DNA degradation [38]. The Malaise trap should preferably be placed at the center of the habitat patch to be characterized and, when possible, the trap should be positioned at a right angle to the dominant insect flight line. While submodule implementation can be restricted to a single trap, we emphasize that biological replicates (simultaneous Malaise trapping events) are desirable within the same habitat patch [56] and can provide useful information regarding sampling efficiency (see, e.g., [57, 58] for occupancy modeling using some means of sampling replication for insects). Similarly, temporal replication is also desirable, considering the possible variability due to changing environmental conditions for optimal arthropod activity and species-specific idiosyncrasies. If temporal replication is not possible, trapping during maximum activity periods for flying insects is desirable. See Table 1 for a summary of key guidelines and recommendations for the 1.1 Malaise trapping sample acquisition submodule.

Recording metadata associated with sampling is also an important action for harmonization. Our opinion converges on a minimum set of metadata attributes for each sample: (i) the geographical coordinates of the Malaise trap, (ii) the date and time interval for the sampling event, and (iii) photo recording (ideally a 360° photo around each trap) of the habitat patch within which the Malaise trap is placed. In agreement with Montgomery et al. [43], we also recommend metadata reporting for the presence of rainfall, or extreme weather events, during the trapping. Detailed characterization of habitat and microhabitats within sampling sites would require time and resources that may limit module uptake. If needed, environmental characterization of sampling sites can potentially be extracted from remote sensing data (see [4]). For additional information on metadata reporting, see the section on DNA sequence and metadata sharing and storage.

Table 1: Summary of key guidelines and recommendations within the 1.1 Malaise trapping sample acquisition submodule

1.1 Malaise trapping sample acquisition submodule	
Sample definition	Townes-style Malaise trap (165 × 110 cm interception area) One week per sample Collecting fluid: >95% ethanol/50–95% propylene glycol Center in habitat patch location Position perpendicular to natural flight corridor Spatial and temporal replicates
Sampling event metadata	Geographical coordinates Date and period of trapping Photo recording for habitat and microhabitat Extreme weather events during trapping
Sample storage	>95% molecular grade ethanol/propylene glycol Fully submerged biomass Storage conditions of −20°C or −80°C

Sample storage conditions, as the endpoint of the sample acquisition chain, carry implications for downstream data quality and are thus an important focus for harmonization. Sample storage conditions are consequential for the degradation of target DNA and/or the proliferation of nontarget biomass in the sample. As such, they can strongly impact metabarcoding biodiversity profiles [59]. However, the effect of this bias on mock arthropod samples, at least for short-term storage (i.e., <1 month), is of limited importance (see [38]). In the case of longer storage of arthropod community samples, we strongly recommend the use of >95% molecular-grade ethanol as a preservative using leakproof glass or plastic vials or jars [60], ensuring that the entire bulk sample is fully submerged before storage and then storage conditions of −20°C or −80°C. In the case of storage or transport safety constraints, propylene glycol (undiluted) can be used as an alternative to ethanol [61]. Such an approach will limit inherent biases in inventory data due to irregular DNA degradation. The storage of biological replicates is always desirable (Table 1).

While Malaise trapping is notably efficient for aerially active arthropods, species with low mobility are less likely to be sampled (e.g., [62]). In this context, pitfall trapping offers a complementary passive sampling technique for ground active arthropods, and thus we consider it to be an appropriate candidate for the development of a complementary sampling submodule. The joint implementation of Malaise and pitfall trapping represents an appropriate compromise to limit the diversity of sampling techniques implemented, while seeking to capture a broad representation of arthropod biodiversity. Pitfall traps [63] are containers buried in the ground with their rim at surface level to capture ground-dwelling (epigeic) insects. Pitfall traps are the most effective method for sampling ground active arthropods and are an established and popular monitoring technique (e.g., the US National Ecological Observatory Network [NEON] [54]; the UK Environmental Change Network [64]). Pitfall and Malaise traps are highly complementary, sampling largely nonoverlapping fractions of arthropod assemblages with reduced additional effort, and they have

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

Table 2: Summary of key guidelines and recommendations within the 1.2 Pitfall trapping sample acquisition submodule

1.2 Pitfall trapping sample acquisition submodule	
Sample definition	Plastic cups with diameter 11 cm, depth 9–11 cm, and a roof raised 1.5 cm Composite sample (4 pitfall traps, placed at the corners of a square with sides of 25 m) One week per sample Collecting fluid: propylene glycol (50–95%) Spatial and temporal replicates
Sampling event metadata	Geographical coordinates Date and period of trapping Photo recording for habitat and microhabitat Extreme weather events during trapping
Sample storage	>95% molecular-grade ethanol/propylene glycol Fully submerged biomass Storage conditions of −20°C or −80°C

already been jointly applied in several wocDNA metabarcoding studies (e.g., [48]).

Guidelines for standardizing pitfall trapping, based on a review of the existing literature [65], have recommended plastic cups with an 11-cm diameter and a 9- to 11-cm depth, and a roof raised 1.5 cm above the trap entrance. The number of individuals sampled per trap can be limited, and as such, composite samples from multiple pitfall traps can be used to increase the sampling effort. There is some controversy over how far apart traps should be placed to be considered as independent samples (e.g., [66, 67]). We suggest that the NEON protocol [54] provides a suitable framework for harmonization, within which a composite sample is generated using 4 pitfall traps arranged at the corners of a square with sides of 25 m. While submodule implementation can be restricted to a single composite sample (4 pitfall traps), biological replicates are desirable (e.g., [54]) and can be achieved by allowing several meters between replicate traps within each corner. Sampling effort is defined by the trapping interval and varies across studies, typically ranging from 3 days to 4 weeks (e.g., [48, 54, 68]). One week provides an appropriate interval and facilitates coordination with the setting and servicing of Malaise traps. Temporal replication is also desirable, and if not possible, trapping should be targeted toward periods of maximum arthropod activity [54]. Propylene glycol (ratio of 50–100% propylene glycol, with water, for a total volume between 100 and 200 mL, depending upon the dilution ratio) is the most frequently recommended collecting medium, as evaporation is negligible compared to ethanol, it is odorless, and it adequately preserves DNA ([54, 55], Table 2).

Similar to Malaise traps, a minimum set of metadata attributes for each pitfall composite sample should include (i) the geographical coordinates of the trap, (ii) period of the trapping event, and (iii) photo recording (ideally a 360° photo around each trap). Following Montgomery et al. [43], we also recommend metadata reporting for the presence of rainfall or extreme events during sampling. Finally, in order to minimize the degradation of target DNA and/or the proliferation of nontarget biomass in the sample during medium- to long-term storage, we strongly recommend the use of >95% molecular-grade ethanol, or propylene glycol, as described above for Malaise trap samples. See Table 2 for key guidelines and recommendations of the 1.2 Pitfall trapping sample acquisition submodule.

Sample processing step

In contrast to microbial or environmental DNA (eDNA) approaches, where samples can be directly processed for DNA extraction, the macroscopic nature of arthropod community samples has led to a broad range of sample processing protocols, among which size sorting is the most common. Size sorting is often used because larger specimens tend to release more DNA and may dominate the total sequence count in metabarcoding data [69]. Thus, sorting invertebrates into multiple size classes and then pooling the digested tissue according to DNA concentration, abundance, or richness in each class has become common practice (e.g., [1, 70, 71]), and size sorting has revealed improved efficiency in the detection of low biomass species (e.g., [40, 70]). However, increasing sequencing depth can also increase taxon recovery to comparable levels without size sorting [72]. More generally, it has been suggested that with sufficient sequencing depth and within reasonable size ranges, species recovery is not skewed by variable biomass of species and that a size-sorting step need not be carried out [71]. Please see the section on amplification, library preparation, and sequencing steps for details on sequencing depth. In addition to the fact that handling time for size sorting places high logistical constraints for large-scale studies, size-sorting procedures also reduce comparability across independent initiatives if not fully harmonized. Given these considerations, we consider size sorting to be unnecessary for a harmonized approach, but if incorporated, it should be of limited complexity (e.g., wet sieving into 2 size fractions, 4-mm sieve pooled 1:10 to 2:10 [>4 mm/<4 mm] [72]) and properly reported. Removing any form of biomass sorting/sample picking steps will also improve cost-effectiveness and facilitate broad implementation for biomonitoring [27].

Biomass and abundance information is often fundamental for biodiversity analysis, including the global assessment of arthropod decline (see [73]). However, deriving abundance information from metabarcode data remains a challenge, primarily due to inherent biases during PCR amplification, but also because of variation in gene copy number, organelle number, and technical aspects of workflows for sampling, laboratory procedures, sequencing, and bioinformatic processing [5, 69, 74]. Given these considerations, we consider that an arthropod community sample processing submodule should emphasize the importance of (i) providing a wet weight measurement for each sample and (ii) generating arthropod community sample photographs. Wet mass measurement can be used as a surrogate for sample biomass. It can be easily obtained from samples after filtering off excess ethanol using a nylon filtration fabric that retains smaller specimens (e.g., 20-μm filters).

Photographic recording is not a commonly reported practice, but looking forward, we think it is very likely that the integration of quantitative morphological and molecular approaches will be an important area of interest and development [75]. There is potential for image-based specimen identification involving machine learning tools to be applied as an external validation of molecular-based diversity estimations, particularly for arthropod groups with limited cryptic variation between species [75–77]. While obtaining high-quality images of arthropod community samples may be time-consuming, we recommend, as a minimum,

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

Table 3: Summary of key guidelines and recommendations within the 2.1 Arthropod community sample processing submodule

2.1 Arthropod community sample processing submodule	
Sample wet mass weight	20-μm nylon filtration fabric
Sample photography	White background Ethanol submerged (white tray) Photographic scale
Size sorting	Minimize size sorting 4-mm sieve
Vouchering specimens	Random or directed selection of specimens for being individually DNA extracted and barcoded

that such images should be taken at high resolution using a conventional stereoscope equipped with a built-in microscope camera or an external single-lens reflex camera with macro lens, over a white background (ideally submerged under ethanol in a plastic tray), and minimizing the overlap among individuals to provide a physical record of the sample. Vouchering selected specimens may be considered unnecessary when well-parameterized reference libraries are available (e.g., [78]) but is otherwise an important consideration for future taxonomic assignment of metabarcoding reads and for completing reference barcode databases (e.g., following BOLD guidelines; see [50, 79]). Vouchering also provides a resource for potential parallel efforts to generate high-throughput specimen-based genomic resources (i.e., partial or complete genomes, microbiomes, diet) for sites of special interest (SuperGOs [13]; i.e., sites where molecular community data are intensively generated at both the temporal and the genomic axes, consistent with the idea of "model ecosystems"). Vouchered barcode sequences are also of particular relevance for bioinformatic processing of metabarcode reads. It has been demonstrated that such sequences are fundamental for efficient and validated filtering of nuclear copies of mitochondrial sequences and that they control for taxonomically inflated estimates of community composition [80]. While sample processing is not the most problematic step for cross-contamination, contamination issues have been reported (e.g., [81]), and at least basic equipment cleaning between samples is required. See Table 3 for key guidelines and recommendations of the arthropod community sample processing submodule.

DNA extraction step

A fundamental consideration for harmonizing wocDNA extraction concerns whether a preextraction homogenization-grinding step (thus implying destruction of the specimens within an arthropod community sample) is needed. Such a step can facilitate homogeneous digestion across specimens and reduce digestion volumes. It is often achieved through manual grinding in a mortar after freezing in liquid nitrogen, grinding in ethanol, or mechanical bead beating. Nondestructive extraction protocols have been developed for unsorted arthropod samples to maintain exoskeletal integrity (e.g., [61, 70, 82]). Using mock arthropod community samples generated from material collected in Malaise traps, Nielsen et al. [82] found that homogenized samples yielded more DNA but generally produced more inconsistent results when compared to nondestructive extraction. When assessing the recovered taxonomic content of samples using operational taxonomic units (OTUs), intact samples performed at least comparable to, if not better than, homogenized samples. Thus, considering that efficiency seems to be comparable, avoiding a homogenization step will (i) reduce potential heterogeneity among studies, (ii) reduce processing time, (iii) reduce contamination risk, and (iv) maintain a physical archive accessible for future developments in image classification using deep learning for the extraction of additional data, such as abundances (see sample processing section). Given these considerations, nondestructive DNA extraction should be a core feature of the arthropod community sample DNA extraction submodule. When necessary (e.g., soil arthropods where a large fraction have hard exoskeletons; see [83]), semidestructive or destructive extraction submodules will need to be developed. Nondestructive DNA extractions require large volumes of digestion buffer to extract wocDNA. Nielsen et al. [82] have demonstrated that OTU diversity estimates are not influenced by the (sub)volume of digestion buffer that is subsequently purified. Given this consideration, typical commercial kit extraction volumes of 100–200 μL can be considered an appropriate subsampling volume for subsequent purification.

A broad range of DNA extraction protocols are being applied to wocDNA metabarcoding. It remains unclear how different extraction methods might impact downstream results, as there is contrasting evidence on its importance based on eDNA approaches [19, 84]. Manual (column-based) and robotic (bead-based) implementations of the Qiagen (Hilden, Germany) DNeasy Blood & Tissue kit and homologous kits have been widely used for extracting wocDNA from terrestrial invertebrates [34]. There is little evidence for PCR inhibitor issues for DNA extracts from arthropod community samples (but see [85]), and if they occur, they can be appropriately accounted for through dilution of DNA extracts before PCR amplification (see next section). Given these considerations, simple and efficient kit-based protocols that allow sample extraction at scale (e.g., Qiagen DNeasy Blood & Tissue and analogous kits; see [86]) provide an appropriate basis for harmonization. Negative controls and technical replicates are fundamental for quality control and can be used to filter out artifactual sequences [87], and as such, their incorporation in the extraction step will also facilitate validation and integration of data across studies.

Biobanking of DNA from environmental samples has been strongly advocated for long-term biomonitoring [88]. Biobanking of DNA ensures opportunities for reanalysis of past data sets with future technologies, an important consideration given high method turnover and associated comparability issues. Aliquots of purified wocDNA are suitable for archiving, ideally using low-DNA binding tubes and freezers of −80°C or colder, but if this option is unavailable, storage at −20°C in nondefrosting freezers provides an adequate alternative. Several museums are already offering this service with affordable pricing (e.g., Smithsonian & Canadian museum in Ottawa). See Table 4 for key guidelines and recommendations for the arthropod community sample DNA extraction submodule.

Amplification, Library Preparation, and Sequencing Step

There is a clear trend toward the use of the *Cytochrome c oxidase* subunit I barcode region (COI barcode) for wocDNA metabarcoding of arthropods (e.g., [37, 40, 83, 89–92]). This can be largely attributed to (i) the good performance of different COI primers for arthropod community samples, (ii) the availability of large COI barcode reference databases, (iii) sufficient variation to typically allow taxonomic assignment at the species level, and (iv) the potential to identify and remove sequencing errors and spurious sequence assemblies by bioinformatic processing based on the predicted variation in protein-coding regions and the limited

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

Table 4: Summary of key guidelines and recommendations proposed within the 3.1 Arthropod community sample DNA extraction submodule

3.1 Arthropod community sample DNA extraction submodule	
Digestion	No physical homogenization step High volumes of digestion buffer Long digestion (shaking)
Purification	200 μL of digestion buffer Qiagen DNeasy Blood & Tissue type Negative controls and technical extraction replicates
Purified DNA storage	Biobanking of DNA aliquots −80°C, −20°C nondefrosting freezers

Table 5: Summary of key guidelines and recommendations proposed within the 4.1 Arthropod community sample DNA amplification, library preparation, and sequencing submodule

4.1 Arthropod community sample DNA amplification, library preparation, and sequencing submodule	
Target DNA fragments and primers	COI locus Second half (3′) of the COI barcode fragment Degenerate primers (see Elbrecht et al. 2019)
PCR conditions	Minimize number of PCR cycles Dilution of DNA extract Nonproofreading Taq PCR replicates (3), ideally individually labeled Negative controls Technical PCR replication Cross-contamination control practices
Library preparation	Two-step protocol

expected length variation within the COI barcode [89]. Multiple primer sets have been demonstrated to efficiently characterize arthropod community samples, particularly those incorporating degenerate nucleotide positions (i.e., positions that allow for the binding of more than 1 nucleotide) (see Fig. 2 in Elbrecht et al. [37]), with a trend toward using the second half (3′) of the COI barcode for metabarcoding studies (e.g., [40, 93]). The BF3 fragment (418 bp) provides better taxonomic resolution than other overlapping fragments. Furthermore, primers within this region are also unaffected by slippage and provide maximum overlap across already published studies [37]. Given these considerations, choosing primers of demonstrated efficiency within the BF3 region (BF3 + BF2 or III_B_F + Fol-degen-rev, among others; see [37]), or that overlap substantially with it, offer high potential for harmonizing across independent studies.

PCR conditions are strongly dependent on selected primers but also on sample composition and polymerase used. Ideally, PCR annealing temperatures and cycle numbers should be quantitative PCR optimized [94]. However, in the absence of such optimization, steps can be taken to reduce unneeded variability across studies. The number of PCR cycles should be maintained at or below 30 cycles if possible, to limit the formation of intrasample chimeras ([95], reviewed in [5]). Serial dilution is a beneficial strategy, as DNA concentration from arthropod community samples, together with PCR inhibitors, can be high, and potential inhibitors can be effectively diluted out (e.g., [96]). Comparisons of polymerase performance for metabarcoding [97] have revealed that polymerase choice impacts read abundance, but not occurrence. Among 6 commercially available polymerases tested, Qiagen Multiplex Master Mix has been shown to provide the most accurate estimates of relative abundance but also generated the highest error rate [97]. While high-fidelity DNA polymerases can reduce PCR error rates [97, 98], their proofreading activity (3′→5′ exonuclease activity) can increase the rate of chimera formation [99, 100]. PCR volume does not appear to be an important consideration for harmonization as it has been reported that it does not influence downstream results but provides opportunity for cost savings via PCR miniaturization (lower cost from reduced quantities of reaction components [101]; Table 5).

Performing PCR replicates and pooling for library preparation or sequencing is a well-established standard in the metabarcoding literature, particularly for arthropod community samples, with strong recommendations for a minimum pooling of 3 PCR replicates [102, 103]. The use of multiple PCR replicates per sample to be individually sequenced (technical replication) is less common, but their importance has been highlighted. Together with PCR negative controls, technical PCR replicates can provide important quality control for the removal of PCR and sequencing artifacts [87, 94, 104]. Thus, negative controls and technical replication within individual sequencing runs should be considered essential practice to identify potential biases and errors from (i) cross-contamination, (ii) tag-jumping events [105], and (iii) false-negative detection. Given the high potential for cross-contamination within the PCR step, rigorous measures should be taken to minimize this risk (e.g., using filter tips, robotic platforms for plate aliquoting). Cross-contamination can be detected and filtered out by including technical replicates, together with positive and negative controls randomly distributed among different plates to bioinformatically curate data, reducing problems associated with tag switching and/or cross-contamination [106]. These should be included in the laboratory and sequencing workflow (e.g., [107]). An important measure that enables one to filter out potential contamination during data processing is to use different nucleotide tag and/or library index combinations for individual PCR replicates within samples, as this will allow for restrictive sequence processing across each replicate [87, 104]. Similarly, the number of reads assigned to a given tag/library index combination that were not used in the study can provide an estimation of the contamination rate and thus a minimum OTU relative abundance that should be considered as reliable [108]. Mock communities have been investigated as positive controls for estimating recovery bias, and the use of synthetic/exogenous internal standards has also been explored to estimate absolute abundance from metabarcode data [10, 109, 110]. In the context of harmonization across studies, universal positive controls harbor much potential for intercalibration. This has yet to be developed and tested, but could be the basis for further improvement within this submodule.

Library preparation involves the addition of sample-specific nucleotide identifiers to amplicons and nucleotide tails for sequencing, for which there is considerable heterogeneity in the arthropod wocDNA metabarcoding literature. In their recent review, Bohmann et al. [106] identified and reviewed 3 main approaches to achieve sample-specific labeling and library preparation in metabarcoding studies. These include (i) a 1-step PCR

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

approach in which sample DNA extracts are amplified, tagged, and built into sequence libraries in a single PCR reaction with fusion primers, then pooled and sequenced; (ii) a 2-step PCR, in which sample DNA extracts are PCR amplified with 2 primer sets: a first PCR with metabarcoding primers carrying the 5′ sequence overhangs and no nucleotide tags and a second PCR using sequence overhangs, allowing the amplicons to be indexed (i5 and i7 indexes); and (iii) a tagged PCR approach, in which DNA extracts are PCR amplified with metabarcoding primers that carry 5′ nucleotide tags, individually tagged PCR products are then pooled, and PCR-based or ligation-based library preparation is performed for pools of 5′ tagged amplicons.

All 3 labeling strategies have been used for arthropod wocDNA metabarcoding (e.g., [70, 94, 111]). The 2-step approach, which is based on the Illumina 16S ribosomal RNA protocol, originally developed for microbiome studies, appears to be more commonly used. Tests comparing consistency and taxon detection efficiency between 1-step and 2-step PCR protocols (in this case implementing TrueSeq Nano over first untagged PCR) using mock arthropod samples reveal better performance with the 2-step protocol [26]. Ligation-based tagged PCR library preparations have been advocated, to avoid false assignment of sequences to samples by tag jumping [94, 112], a recognized problem within the PCR-based tagged approach [105, 106]. However, no study has yet compared performance between 2-step and ligation-based tagged PCR. Between these two, the 2-step approach is the more frequently used for arthropod metabarcoding and thus provides a suitable approach to minimize heterogeneity across studies (Table 5).

The sequencing depth needed to recover all taxa is strongly dependent on the diversity and complexity of a given sample. A sequencing depth of 60,000 ± 55,000 reads per amplicon per sample is commonly reported [113]. Increasing sequencing depth can increase the detection rate of low-abundance taxa and reduce the impacts of differential processing protocols on perceived diversity [40]. However, increased sequencing depth increases the cost by sample (see Table 2 in Piper et al. [7] for a summary of the costs [2019] and Gb output for each platform) and inherently increases the detection of artifactual sequences, requiring additional procedures for their removal [5, 80, 104]. Distinguishing between sufficient or insufficient sequencing depth can be controlled for by evaluating replicability [40] or by taxa recovery graphs on mock or composition controlled communities of comparable nature [114]. The choice of sequencing platform also has potential to generate variation among data sets. This variation appears to be limited across currently popular platforms, such as Illumina MiSeq, Ion Torrent PGM, and Ion Torrent S5 [40]. However, as future sequencing platforms may present greater variation, it is important to report such details (e.g., sequencing platform, read length). See Table 5 for key guidelines and recommendations for the arthropod community sample DNA amplification, library preparation, and sequencing submodule.

Metadata and DNA sequence sharing and storage step

Metadata associated with the different steps of generating metabarcode data should be reported with DNA sequence data to enhance long-term reuse value (see [115]). The GEOME (Genomic Observatories Metadatabase) initiative [15, 16] offers a very useful platform, facilitating findable, accessible, interoperable, and reusable data archival practices (i.e., FAIR principles). Interoperability is central to GEOME, as metadata follow controlled vocabularies consistent with DarwinCore and MIxS standards [116, 117] and new records on GEOME are incorporated into the Global Biodiversity Information Facility, GBIF (https://www.gbif.org/). A customizable but standard-compliant single spreadsheet for metainformation, including (i) the reference to the submodules implemented within each data acquisition steps (e.g., 1.2 sample acquisition submodule, 2.1 sample processing submodule, etc.) and (ii) all key information highlighted within each of the submodules, will facilitate downstream comparison among data sets. The metadata spreadsheet for the terrestrial arthropod module (GEOME spreadsheet) can be additionally included as supplementary publication material.

Finally, GEOME also facilitates DNA data sharing through the deposition of raw genetic data to the Sequence Read Archive (SRA, www.ncbi.nlm.nih.gov/sra), while maintaining persistent links to standard compliant metadata held in the GEOME database. SRA is thus an ideal platform for the storage of demultiplexed HTS files. Given the continuous development and improvement of bioinformatic tools for HTS data analysis, public archiving of raw DNA data is important to facilitate future synthetic analysis across historical data sets. See Table 6 for key guidelines and recommendations of the arthropod community sample metadata and DNA sequence sharing and storage submodule.

Table 6: Summary of key guidelines and recommendations proposed within the 5.1 Arthropod community sample metadata and DNA sequence sharing and storage submodule

5.1 Arthropod community sample metadata and DNA sequence sharing and storage submodule	
Metadata	GEOME metadata submission GEOME spreadsheet with the key information of the modules performed
DNA sequences	Raw data SRA

Conclusions

Whole-organism community DNA metabarcoding is emerging as a powerful tool to characterize and compare arthropod communities, from the scale of local community composition through to global comparative analyses. For this potential to be fully realized, comparability across data sets generated by independent research groups is a fundamental prerequisite. There are several challenges to achieve this. First, as is the case for many new fields, early development has led to different strategies and tools, among which some will facilitate data comparability, while others will not. Here we have addressed this issue by suggesting a modular framework that seeks to reduce redundant efforts and improve comparability across studies by harmonization of common practice across different research initiatives, where that practice demonstrates utility. We have illustrated this framework with recommendations for a module for the characterization of terrestrial arthropods. A second challenge is that canalization of different practices to optimize comparability at the community level may, inadvertently, limit flexibility at the scale of individual studies. While this is to some extent unavoidable, the flexible structure we presented here seeks to broaden the applicability of a modular framework within the wocDNA metabarcoding community. Finally, unless appropriate data and metadata are provided for a given wocDNA metabarcode study, the opportunities for integrative analyses across historical data sets are likely to be limited. We address this

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

challenge by advocating good reporting practice and highlight that the submodule structure provides a framework for the incorporation of new advances as they emerge within the field of metabarcoding. We advocate the adoption and development of the terrestrial arthropod module that we propose here, as an important step toward harmonization of metabarcode data. We further encourage the development of additional submodules for the terrestrial arthropod module (e.g., soil mesoarthropod sample acquisition, pan trapping for pollinator sample acquisition), as well as modules for other biodiversity fractions that are appropriate targets for wocDNA metabarcoding.

Abbreviations

COI barcode: Cytochrome c oxidase subunit I barcode region; GBIF: Global Biodiversity Information Facility; GEOME: Genomic Observatories Metadatabase; GO: Genomic Observatories; HTS: high-throughput sequencing; OTUs: operational taxonomic units; PCR: polymerase chain reaction; SRA: Sequence Read Archive; wocDNA: whole-organism community DNA.

Data Availability

Not applicable.

Competing Interests

A.P.V. is a cofounder and scientific advisor of NatureMetrics, a private company providing commercial services in DNA-based monitoring. The authors declare that they have no other conflicts of interest.

Funding

The working group "Toward Harmonisation for the Generation of Metabarcoding Data: Soil Biodiversity and Terrestrial Arthropod modules" held in November 2020 (online) was organized by the iBioGen project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 810729. P.A. was funded through a Junior Leader Fellowship (LCF/BQ/PR21/11840006) by "la Caixa" Foundation (ID 100010434) and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 847648.

Authors' Contributions

P.A. and B.C.E. conceptualized the manuscript. All authors contributed to the ideas and discussion of this review. P.A. and B.C.E. coordinated the working group meetings and led the writing with contributions from all authors. All authors read and approved the final manuscript.

References

1. Ji, Y, Ashton, L, Pedley, SM, *et al.* Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol Lett* 2013;**16**(10):1245–57.
2. Porter, TM, Hajibabaei, M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol Ecol* 2018;**27**(2):313–38.
3. Bohan, DA, Vacher, C, Tamaddoni-Nezhad, A, *et al.* Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol Evol* 2017;**32**(7):477–87.
4. Bush, A, Sollmann, R, Wilting, A, *et al.* Connecting Earth observation to high-throughput biodiversity data. *Nat Ecol Evol* 2017;**1**(7):1–9.
5. Taberlet, P, Bonin, A, Zinger, L, *et al. Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press, 2018:1–253. doi: 10.1093/oso/9780198767220.001.0001.
6. Bush, A, Catullo, R, Mokany, K, *et al.* Incorporating existing thermal tolerance into projections of compositional turnover under climate change. *Global Ecol Biogeogr* 2019;**28**(6):851–61.
7. Piper, AM, Batovska, J, Cogan, NOI, *et al.* Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *Gigascience* 2019;**8**(8):1–22.
8. Gilbert, JA, Jansson, JK, Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol* 2014;**12**(1):1–4.
9. Gilbert, JA, Meyer, F, Antonopoulos, D, *et al.* Meeting Report: The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project. *Standards Genomic Sci* 2010;**3**(3):243–8.
10. Ovaskainen, O, Abrego, N, Somervuo, P, *et al.* Monitoring Fungal Communities With the Global Spore Sampling Project. *Front Ecol Evol* 2020;**7**:1–9.
11. Davies, N, Meyer, C, Gilbert, JA, *et al.* A call for an international network of genomic observatories (GOs). *Gigascience* 2012;**1**(1):1–5.
12. Davies, N, Field, D, Amaral-Zettler, L, *et al.* The founding charter of the Genomic Observatories Network. *Gigascience* 2014;**3**(1):1–5.
13. Arribas, P, Andújar, C, Bidartondo, MI, *et al.* Connecting high-throughput biodiversity inventories—opportunities for a site-based genomic framework for global integration and synthesis. *Mol Ecol* 2021;**30**(5):1120–35.
14. Guralnick, R, Walls, R, Jetz, W. Humboldt Core—toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. *Ecography* 2018;**41**(5):713–25.
15. Deck, J, Gaither, MR, Ewing, R, *et al.* The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol* 2017;**15**(8):e2002925.
16. Riginos, C, Crandall, ED, Liggins, L, *et al.* Building a global genomics observatory: using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Mol Ecol Resour* 2020;**20**(6):1458–69.
17. Creedy, TJ, Andújar, C, Meramveliotakis, E, *et al.* Coming of age for COI metabarcoding of whole organism community DNA: towards bioinformatic harmonisation. *Mol Ecol Resour* 2022;**22**(3):847–61.
18. Caporaso, JG, Paszkiewicz, K, Field, D, *et al.* The Western English Channel contains a persistent microbial seed bank. *ISME J* 2012;**6**(6):1089–93.
19. Marotz, C, Amir, A, Humphrey, G, *et al.* DNA extraction for streamlined metagenomics of diverse environmental samples. *BioTechniques* 2017;**62**(6):290–3.
20. Alberti, A. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* 2017;**4**:1700932017.
21. Gorsky, G, Bourdin, G, Lombard, F, *et al.* Expanding Tara Oceans protocols for underway, ecosystemic sampling of the ocean-atmosphere interface during Tara Pacific Expedition (2016–2018). *Front Marine Sci* 2019;**6**:1–20.
22. Kopf, A. The ocean sampling day consortium. *Gigascience* 2015;**4**(1):1–5.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

23. Dickie, IA, Boyer, S, Buckley, HL, *et al.* Towards robust and repeatable sampling methods in eDNA based studies. *Mol Ecol Resour* 2018;**18**(5):940–52.
24. Canonico, G, Buttigieg, PL, Montes, E, *et al.* Global observational needs and resources for marine biodiversity. *Front Marine Sci* 2019;**6**:1–20.
25. Murray, DC, Coghlan, ML, Bunce, M. From benchtop to desktop: important considerations when designing amplicon sequencing workflows. *PLoS One* 2015;**10**(4):e0124671.
26. Zizka, VMA, Elbrecht, V, Macher, JN, *et al.* Assessing the influence of sample tagging and library preparation on DNA metabarcoding. *Mol Ecol Resour* 2019;**19**(4):893–9.
27. Blackman, RC, Mächler, E, Altermatt, F, *et al.* Advancing the use of molecular methods for routine freshwater macroinvertebrate biomonitoring—the need for calibration experiments. *Metabarcoding Metagenomics* 2019;**3**:49–57.
28. Zaiko, A, Greenfield, P, Abbott, C, *et al.* Towards reproducible metabarcoding data: Lessons from an international cross-laboratory experiment. *Mol Ecol Resour* 2022;**22**(2):519–38.
29. Philippot, L, Ritz, K, Pandard, P, *et al.* Standardisation of methods in soil microbiology: progress and challenges. *FEMS Microbiol Ecol* 2012;**82**(1):1–10.
30. Stork, NE, McBroom, J, Gely, C, *et al.* New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc Natl Acad Sci* 2015;**112**(24):7519–23.
31. Goulson, D. The insect apocalypse, and why it matters. *Curr Biol* 2019;**29**(19):R967–71.
32. Harvey, JA, Heinen, R, Armbrecht, I, *et al.* International scientists formulate a roadmap for insect conservation and recovery. *Nat Ecol Evol* 2020;**4**(2):174–6.
33. Seebens, H, Blackburn, TM, Dyer, EE, *et al.* Global rise in emerging alien species results from increased accessibility of new source pools. *Proc Natl Acad Sci* 2018;**115**(10):E2264–73.
34. Liu, M, Clarke, LJLJ, Baker, SC, *et al.* A practical guide to DNA metabarcoding for entomological ecologists. *Ecol Entomol* 2020;**45**(3):373–85.
35. Yu, DW, Ji, Y, Emerson, BC, *et al.* Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol Evol* 2012;**3**(4):613–23.
36. Andújar, C, Arribas, P, Gray, C, *et al.* Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Mol Ecol* 2018;**27**(1):146–66.
37. Elbrecht, V, Braukmann, TW, Ivanova, NV, *et al.* Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* 2019;**7**(e7745):1–23.
38. Krehenwinkel, H, Fong, M, Kennedy, S, *et al.* The effect of DNA degradation bias in passive sampling devices on metabarcoding studies of arthropod communities and their associated microbiota. *PLoS One* 2018;**13**(1):e0189188.
39. Krehenwinkel, H, Kennedy, SR, Rueda, A, *et al.* Scaling up DNA barcoding—primer sets for simple and cost efficient arthropod systematics by multiplex PCR and Illumina amplicon sequencing. *Methods Ecol Evol* 2018;**9**(11):2181–93.
40. Braukmann, TWA, Ivanova, NV, Prosser, SWJ, *et al.* Metabarcoding a diverse arthropod mock community. *Mol Ecol Resour* 2019;**19**(3):711–27.
41. D'Souza, ML, van der Bank, M, Shongwe, Z, *et al.* Biodiversity baselines: tracking insects in Kruger National Park with DNA barcodes. *Biol Conserv* 2021;**256**:109034.
42. Basset, Y, Springate, ND, Aberlenc, HP, *et al.* A review of methods for sampling arthropods in tree canopies. In: NE Strok, J Adis, RK Didham, eds. *Canopy Arthropods*. London: Chapman and Hall.
43. Montgomery, GA, Belitz, MW, Guralnick, RP, *et al.* Standards and best practices for monitoring and benchmarking insects. *Front Ecol Evol* 2021;**8**:1–18.
44. Geiger, MF, Moriniere, J, Hausmann, A, *et al.* Testing the Global Malaise Trap Program—how well does the current barcode reference library identify flying insects in Germany? *Biodiversity Data J* 2016;**4**:e10671.
45. Marquina, D, Esparza-Salas, R, Roslin, T, *et al.* Establishing arthropod community composition using metabarcoding: surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Mol Ecol Resour* 2019;**19**(6):1516–30.
46. Barsoum, N, Bruce, C, Forster, J, *et al.* The devil is in the detail: metabarcoding of arthropods provides a sensitive measure of biodiversity response to forest stand composition compared with surrogate measures of biodiversity. *Ecol Indic* 2019;**101**:313–23.
47. Ritter, CD, Häggqvist, S, Karlsson, D, *et al.* Biodiversity assessments in the 21st century: the potential of insect traps to complement environmental samples for estimating eukaryotic and prokaryotic diversity using high-throughput DNA metabarcoding. *Genome* 2019;**62**(3):147–59.
48. Watts, C, Dopheide, A, Holdaway, R, *et al.* DNA metabarcoding as a tool for invertebrate community monitoring: a case study comparison with conventional techniques. *Austral Entomol* 2019;**58**(3):675–86.
49. Hausmann, A, Segerer, AH, Greifenstein, T, *et al.* Toward a standardized quantitative and qualitative insect monitoring scheme. *Ecol Evol* 2020;**10**(9):4009–20.
50. DeWaard, JR, Levesque-Beaudin, V, DeWaard, SL, *et al.* Expedited assessment of terrestrial arthropod diversity by coupling Malaise traps with DNA barcoding. *Genome* 2019;**62**(3):85–95.
51. Malaise, R. A new insect-trap. *Entomol Tidskr* 1937;**58**:148–60.
52. Schmidt, O, Schmidt, S, Häuser, CL, *et al.* Using Malaise traps for collecting Lepidoptera (Insecta), with notes on the preparation of Macrolepidoptera from ethanol. *Biodiversity Data J* 2019;**7**:1–12.
53. Karlsson, D, Forshage, M, Holston, K, *et al.* The data of the Swedish Malaise Trap Project, a countrywide inventory of Sweden's insect fauna. *Biodiversity Data J* 2020;**8**:1–28.
54. Hoekman, D, Levan, KE, Ball, GE, *et al.* Design for ground beetle abundance and diversity sampling within the National Ecological Observatory Network. *Ecosphere* 2017;**8**(4):1–17.
55. Nakamura, S, Tamura, S, Taki, H, *et al.* Propylene glycol: a promising preservative for insects, comparable to ethanol, from trapping to DNA analysis. *Entomol Exp Appl* 2020;**168**(2):158–65.
56. Steinke, D, Braukmann, TWA, Manerus, L, *et al.* Effects of Malaise trap spacing on species richness and composition of terrestrial arthropod bulk samples. *Metabarcoding Metagenomics* 2021;**5**:43–50.
57. Szewczyk, TM, McCain, CM. Disentangling elevational richness: a multi-scale hierarchical Bayesian occupancy model of Colorado ant communities. *Ecography* 2019;**42**(5):977–88.
58. Powney, GD, Carvell, C, Edwards, M, *et al.* Widespread losses of pollinating insects in Britain. *Nat Commun* 2019;**10**(1):1–6.
59. Weißbecker, C, Buscot, F, Wubet, T. Preservation of nucleic acids by freeze-drying for next generation sequencing analyses of soil microbial communities. *J Plant Ecol.* 2017;**10**(1):81–90.
60. Marquina, D, Buczek, M, Ronquist, F, *et al.* The effect of ethanol concentration on the morphological and molecular preservation of insects for biodiversity studies. *PeerJ* 2021;**9**:e10799.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

61. Martoni, F, Nogarotto, E, Piper, AM, *et al.* Propylene glycol and non-destructive DNA extractions enable preservation and isolation of insect and hosted bacterial DNA. *Agriculture* 2021;**11**(1):77.
62. de Kerdrel, GA, Andersen, JC, Kennedy, SR, *et al.* Rapid and cost-effective generation of single specimen multilocus barcoding data from whole arthropod communities by multiple levels of multiplexing. *Sci Rep* 2020;**10**(1):1–12.
63. Hertz, M. Huomioita petokuoriaisten olinpaikoista. *Luonnon Yst* 1927;**31**:218–22.
64. Brooks, DR, Bater, JE, Clark, SJ, *et al.* Large carabid beetle declines in a United Kingdom monitoring network increases evidence for a widespread loss in insect biodiversity. *J Appl Ecol* 2012;**49**(5):1009–19.
65. Brown, GR, Matthews, IM. A review of extensive variation in the design of pitfall traps and a proposal for a standard pitfall trap design for monitoring ground-active arthropod biodiversity. *Ecol Evol*. 2016;**6**(12):3953–64.
66. Ward, DF, New, TR, Yen, AL. Effects of pitfall trap spacing on the abundance, richness and composition of invertebrate catches. *J Insect Conservation* 2001;**5**(1):47–53.
67. Hohbein, RR, Conway, CJ. Pitfall traps: a review of methods for estimating arthropod abundance. *Wildlife Soc Bull* 2018;**42**(4):597–606.
68. Missa, O, Basset, Y, Alonso, A, *et al.* Monitoring arthropods in a tropical landscape: relative effects of sampling methods and habitat types on trap catches. *J Insect Conservation* 2009;**13**(1):103–18.
69. Deagle, BE, Thomas, AC, McInnes, JC, *et al.* Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Mol Ecol* 2019;**28**(2):391–406.
70. Elbrecht, V, Peinert, B, Leese, F. Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecol Evol* 2017;**7**(17):6918–26.
71. Creedy, TJ, Ng, WS, Vogler, AP. Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. *Ecol Evol* 2019;**9**(6):3105–16.
72. Elbrecht, V, Bourlat, S, Hörren, T, *et al.* Pooling size sorted malaise trap fractions to maximise taxon recovery with metabarcoding. *PeerJ* 2021;**9**:e12177.
73. Cardoso, P, Leather, SR. Predicting a global insect apocalypse. *Insect Conservation Diversity* 2019;**12**(4):263–7.
74. Zinger, L, Bonin, A, Alsos, IG, *et al.* DNA metabarcoding—need for robust experimental designs to draw sound ecological conclusions. *Mol Ecol* 2019;**28**(8):1857–62.
75. Ärje, J, Melvad, C, Jeppesen, MR, *et al.* Automatic image-based identification and biomass estimation of invertebrates. *Methods Ecol Evol* 2020;**11**(8):922–31.
76. Valan, M, Makonyi, K, Maki, A, *et al.* Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Syst Biol* 2019;**68**(6):876–95.
77. Valan, M, Vondráček, D, Ronquist, F. Awakening a taxonomist's third eye: exploring the utility of computer vision and deep learning in insect systematics. *Syst Entomol* 2021;**46**(4):757–66.
78. Ronquist, F, Forshage, M, Häggqvist, S, *et al.* Completing Linnaeus's inventory of the Swedish insect fauna: only 5,000 species left? *PLoS One* 2020;**15**(3):e0228561.
79. deWaard, JR, Ratnasingham, S, Zakharov, EV, *et al.* A reference library for the identification of Canadian invertebrates: 1.5 million DNA barcodes, voucher specimens, and genomic samples. *Sci Data* 2019;**6**(1)1–12.
80. Andújar, C, Creedy, TJ, Arribas, P, *et al.* Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Mol Ecol Resour* 2021;**21**(6):1772–87.
81. Creedy, TJTJ, Norman, H, Tang, CQCQ, *et al.* A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. *Mol Ecol Resour* 2020;**20**(1):40–53.
82. Nielsen, M, Gilbert, MTP, Pape, T, *et al.* A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. *Environ DNA* 2019;**1**(2):144–54.
83. Arribas, P, Andújar, C, Hopkins, K, *et al.* Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods Ecol Evol* 2016;**7**(9):1071–81.
84. Deiner, K, Walser, JC, Mächler, E, *et al.* Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biol Conserv* 2015;**183**:53–63.
85. Majaneva, M, Diserud, OH, Eagle, SHC, *et al.* Choice of DNA extraction method affects DNA metabarcoding of unsorted invertebrate bulk samples. *Metabarcoding Metagenomics* 2018;**2**:1–12.
86. Sellers, GS, Di Muri, C, Gómez, A, *et al.* Mu-DNA: A modular universal DNA extraction method adaptable for a wide range of sample types. *Metabarcoding Metagenomics* 2018;**2**:e24556.
87. Zinger, L, Lionnet, C, Benoiston, A-S, *et al.* metabaR: an R package for the evaluation and improvement of DNA metabarcoding data quality. *Methods Ecol Evol* 2021;**12**(4):586–92.
88. Jarman, SN, Berry, O, Bunce, M. The value of environmental DNA biobanking for long-term biomonitoring. *Nat Ecol Evol* 2018;**2**(8):1192–3.
89. Andújar, C, Arribas, P, Yu, DW, *et al.* Why the COI barcode should be the community DNA metabarcode for the Metazoa. *Mol Ecol* 2018;**27**(20):3968–75.
90. Beng, KC, Tomlinson, KW, Shen, XH, *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;**6**(1):1–13.
91. Elbrecht, V, Leese, F, Nichols, SJ. Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Front Environ Sci* 2017;**5**:1–11.
92. Leray, M, Yang, JY, Meyer, CP, *et al.* A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front Zool* 2013;**10**(1):34.
93. Arribas, P, Andújar, C, Salces-Castellano, A, *et al.* The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. *Mol Ecol* 2021;**30**(1):48–61.
94. Yang, C, Bohmann, K, Wang, X, *et al.* Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods Ecol Evol* 2021;**12**(7):1252–64.
95. Haas, BJ, Gevers, D, Earl, AM, *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 2011;**21**(3):494–504.
96. Elbrecht, V, Vamos, EE, Steinke, D, *et al.* Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* 2018;**6**:e4644.
97. Nichols, RV, Vollmers, C, Newsom, LA, *et al.* Minimizing polymerase biases in metabarcoding. *Mol Ecol Resour* 2018;**18**(5):927–39.
98. Sze, MA, Schloss, PD. The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. *mSphere* 2019;**4**(3):1–13.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024

99. Ahn, JH, Hong, IP, Bok, JI, *et al.* Pyrosequencing analysis of the bacterial communities in the guts of honey bees Apis cerana and Apis mellifera in Korea. *J Microbiol* 2012;**50**(5):735–45.
100. Gury, J, Zinger, L, Gielly, L, *et al.* Exonuclease activity of proofreading DNA polymerases is at the origin of artifacts in molecular profiling studies. *Electrophoresis* 2008;**29**(11):2437–44.
101. Buchner, D, Beermann, AJ, Leese, F, *et al.* Cooking small and large portions of "biodiversity-soup": Miniaturized DNA metabarcoding PCRs perform as good as large-volume PCRs. *Ecol Evol* 2021;**11**(13):9092–9.
102. Ficetola, GF, Pansu, J, Bonin, A, *et al.* Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol Ecol Resour* 2015;**15**(3):543–56.
103. Dopheide, A, Xie, D, Buckley, TR, *et al.* Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods Ecol Evol* 2019;**10**(1):120–33.
104. Alberdi, A, Aizpurua, O, Gilbert, MTP, *et al.* Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol Evol* 2018;**9**(1):134–47.
105. Schnell, IB, Bohmann, K, Gilbert, MTP. Tag jumps illuminated—reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* 2015;**15**(6):1289–303.
106. Bohmann, K, Elbrecht, V, Carøe, C, *et al.* Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Mol Ecol Resour* 2022;**22**(4):1231–46.
107. Bista, I, Carvalho, GR, Tang, M, *et al.* Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Mol Ecol Resour* 2018;**18**(5):1020–34.
108. Esling, P, Lejzerowicz, F, Pawlowski, J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res* 2015;**43**(5):2513–24.
109. Smets, W, Leff, JW, Bradford, MA, *et al.* A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biol Biochem* 2016;**96**:145–51.
110. Ushio, M, Murakami, H, Masuda, R, *et al.* Quantitative monitoring of multispecies fish environmental DNA using high-throughput sequencing. *Metabarcoding Metagenomics* 2018;**2**: 1–15.
111. Elbrecht, V, Leese, F, Rockström, J, *et al.* Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLoS One* 2015;**10**(7):e0130324.
112. Carøe, C, Tagsteady, Bohmann K. A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Mol Ecol Resour* 2020;**20**(6):1620–31.
113. Singer, GAC, Fahner, NA, Barnes, JG, *et al.* Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci Rep* 2019;**9**(1):1–12.
114. Hajibabaei, M, Porter, TM, Wright, M, *et al.* COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS One* 2019;**14**(9):e0220953.
115. Tedersoo, L, Ramirez, KS, Nilsson, RH, *et al.* Standardizing metadata and taxonomic identification in metabarcoding studies. *Gigascience* 2015;**4**(1):1–4.
116. Wieczorek, J, Bloom, D, Guralnick, R, *et al.* Darwin core: an evolving community-developed biodiversity data standard. *PLoS One* 2012;**7**(1):e29715.
117. Yilmaz, P, Kottmann, R, Field, D, *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011;**29**(5):415–20.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac065/6646445 by guest on 28 June 2024