HIDDEN IN PLAIN SIGHT

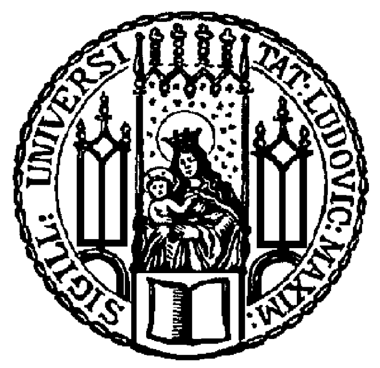
Addressing the hidden diversity of German dark taxa using innovative technologies

by

Caroline Chimeno

Dissertation der Fakultät für Biologie

der Ludwig-Maximilians-Universität München



Munich, 2023

Gutachter 1: Prof. Dr. Gerhard Haszprunar

Gutachter 2: Prof. Dr. Herwig Stibor

Tag der Abgabe: 02.10.2023

Tag der mündlichen Prüfung: 19.02.2024

STATUTORY DECLARATION & STATEMENT

"I herewith declare that I have composed the present thesis myself and without any other than the cited sources and aids. Sentences and parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance."

Munich, October 2, 2023

____

Caroline Chimeno

EIDESSTATTLICHE ERKLÄRUNG

"Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt habe. Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich anderweitig einer Doktorprüfung ohne Erfolg nicht unterzogen habe."

München, den 2. Oktober 2023

____

Caroline Chimeno

To my family

who, I think, is unaware of what I truly do.

CONTENTS

Abbreviations & Acronyms 1
List of Publications 2
List of Figures 4
Summary 5
Zusammenfassung 8
Introduction 11
Not all superheroes wear capes 11
A look at traditional insect research 14
Expediting insect research with molecular methods 18
Peering into the unknown 24
Global Discussion 28
Small, abundant, rare 28
Dealing with the tip of the diversity-iceberg 31
Diversity assessments with COI 33
Integrative taxonomy for hyperabundant & -diverse taxa 35
Yet every method is subjective 40
Trying to bypass specimen destruction in DNA metabarcoding 42
Conclusion 45
References 47
Acknowledgements 79
Appendix 81
Publication I: Abundant, diverse, unknown 83
Publication II: Gallmücken in Bayern 103
Publication III: Peering into the darkness 113
Publication IV: Facing the infinity 133
Publication V: Depicting environmental gradients 165

ABBREVIATIONS & ACRONYMS

ASAP	Assemble Species by Automatic Partitioning
BIN	Barcode Index Number
BIO	Biodiversity Institute of Ontario
BOLD	Barcode of Life Data System
bp	base pair
CBG	Centre for Biodiversity Genomics
COI	Cytochrome C Oxidase Subunit I
DNA	Deoxyribonucleic acid
EtOH	ethanol
EU	European Union
GBOL	German Barcode of Life
HTS	High Throughput Sequencing
MEK	Methyl Ethyl Ketone
MOTU	Molecular Operational Taxonomic Unit
mtDNA	mitochondrial DNA
OTU	Operational Taxonomic Unit
RESL	Refined Single Linkage
SMNS	State Museum of Natural History Stuttgart
SNSB-ZSM	Bavarian State Collection of Zoology
ZFMK	Zoological Research Museum Alexander Koenig

LIST OF PUBLICATIONS

The thesis is based on the following articles, which are referred to in the text by their Roman numerals.

I. Chimeno, C., Schmidt, S., Cancian de Araujo, B., Perez, K., von Rintelen, T., Schmidt, O., Hamid, H., Pramesa Narakusumo, R. & Balke, M. (2023). Abundant, diverse, unknown: Extreme species richness and turnover despite drastic undersampling in two closely placed tropical Malaise traps. *Plos One*, 18(8): e0290173. https://doi.org/10.1371/journal.pone.0290173

II. Chimeno, C., Doczkal, D., Haszprunar, G., Hausmann, A., Jaschhof, M., Kotrba, M., Perez, K., Raupach, M. & Schmidt, S. (2021). Gallmücken in Bayern: DNA Barcoding vermittelt neue Einblicke in die Mega-Vielfalt bislang (zu oft) ignorierter Mikrodipteren. *Nachrichtenblatt Bayerischen Entomologen*, 70 (3/4). *Not peer-reviewed.*

III. Chimeno, C., Hausmann, A., Schmidt, S., Raupach, M. J., Doczkal, D., Baranov, V., Hübner, J., Höcherl, A., Albrecht, R., Jaschhof, M., Haszprunar, G. & Hebert, P. D. (2022). Peering into the Darkness: DNA Barcoding Reveals Surprisingly High Diversity of Unknown Species of Diptera (Insecta) in Germany. *Insects*, 13(1), 82. 10.3390/insects13010082

IV. Chimeno, C., Björn, R., Manfrin, A., Kalinkat, G., Hölker, F. & Baranov, V. (2023). Facing the Infinity: Tackling large samples of challenging Chironomidae (Diptera) with an integrative approach. *PeerJ*, 11: e15336. 10.7717/peerj.15336

V. Chimeno, C., Hübner, J., Seifert, L., Morinière, J., Bozicevic, V., Hausmann, A., Schmidt, S. & Müller, J. (2023). Depicting environmental gradients from Malaise trap samples: Is ethanol based

DNA metabarcoding enough? *Insect Conservation and Diversity*, *16*(1), 47–64.

https://doi.org/10.1111/icad.12609

LIST OF FIGURES

Figure 1. Schematic representation of a Malaise trap. 15

Figure 2. Species identification workflows. 22

Figure 3. Proportion of dark taxa and BIN frequency from samples collected in Sumatra and Germany. 30

SUMMARY

Studies have shown that in reality, we are drastically underestimating species numbers and that a large proportion of the global diversity is still awaiting discovery or description (Engel et al., 2021; González-Oreja, 2008). Yet the bulk of the planet's hidden diversity (and biomass) is found in groups that are difficult to study and have therefore received very little taxonomic attention in the past (so-called "dark taxa") (Hardulak, 2020; Hartop et al., 2022; Hausmann et al., 2020; Meier et al., 2021; Morinière et al., 2019). These groups are so hyperdiverse, that dark taxa have also been referred to as "open-ended" taxa because species number estimates are almost impossible to make (Hartop et al., 2022). The majority of dark taxa are found among two insect orders, Diptera (flies) and Hymenoptera (ants, bees, wasps), and consist in large part of non-brachyceran Diptera such as midges and gnats, as well as parasitoid wasps (Hausmann et al., 2020). Being cryptic diverse, highly abundant, and miniscule (<2 mm), the analysis of these groups is very demanding so often, they are simply cast aside and analysis is limited to better studied and easier-to-handle taxa. At the same time, the fact that dark taxa are so abundant in samples (they can make up more than 70% of specimen numbers) implies that they play an essential in ecological functioning (*GBOL III*, 2023). Conversely, this means that it is all the more important to make these groups tangible to research so that they can be implemented into conservation measures.

If we want to study and describe species more effectively, we need a taxonomic renaissance in descriptive taxonomy (Giangrande, 2003). Morphological methods, which have been used for the last 250 years, alone do not provide sufficient discriminatory information for the tiny, cryptic diverse species of dark taxa. Luckily, recent advances in molecular biology are providing the much-needed means to accelerate species discovery by providing DNA-based discrimination methods (Morinière et al., 2016; Wang et al., 2018). These not only drastically speed-up sample processing and species

identifications, these also enable the analysis of entire insect communities in one go (Srivathsan et al., 2015). Also, ever more scientists are recommending the use of integrative workflows which implement methods from different disciplines for species description and delimitation (Meier et al., 2022). Using such complementary approaches increase scientific vigor, as no single method is error-free (Dayrat, 2005; Schlick-Steiner et al., 2010).

The main goal of this thesis is to develop an integrative framework for the rapid processing of large samples of dark taxa with three specific objectives. These are (1) identifying the dark diversity in temperate regions, (2) developing an integrative methodology to assess samples of dark taxa, and (3) testing the usability of preservative ethanol of insect bulk samples for metabarcoding applications.

The first objective aims at raising awareness for the presence of dark taxa not only in a tropical, but also in a temperate setting *(Publications I-III)*. Data obtained through large-scale DNA barcoding on Malaise trap samples from Padang, Sumatra (2016) and from Germany (2012-2017) were analyzed. The large prevalence of dark taxa in Malaise trap samples (in terms of species diversity and specimen abundance) was demonstrated, and species numbers for four dipteran dark taxa were extrapolated to provide data-based species estimates for Germany.

Second, having raised awareness regarding the hidden diversity of dark taxa in temperate regions, a strategy is formulated for tackling one dark taxon from large samples *(Publication IV)*. Using Chironomidae (non-biting midges; Diptera, Nematocera) as a model group, an integrative approach was proposed which includes (i) a three-level subsampling method to reduce the workload of sample processing, (ii) morphology- and (iii) DNA-based methods in parallel to evaluate species diversity, and (iv) examining possible inconsistencies across methods. Here, the results show that with this integrative framework, more than 90% of all species were detected despite having identified only 7%

of individuals. Also, the results demonstrate that using either identification method on its own would have been prone to errors that would have gone undetected.

Lastly, the usability of ethanol-based DNA for metabarcoding applications is assessed *(Publication V)*. Here, the research question is whether ecological information is conserved in the DNA that is extracted from the collection fluid of bulk samples. If so, this would imply that the usual step of specimen homogenization for DNA extraction can be bypassed because the ethanol of samples can be simply poured out and used for analysis instead. In this manner, all specimens are left intact for further analyzes. Here, the results suggest that ethanol-based DNA does not conserve ecological information and until future research has provided more successful results, it is recommended that researchers dealing with terrestrial ecosystems be careful when using ethanol-based DNA.

In conclusion, this thesis builds a framework that combines different disciplines to efficiently study the immense (hidden) insect diversity that is housed in our temperate environments.

ZUSAMMENFASSUNG

Neueste Studien haben gezeigt, dass die tatsächliche Anzahl rezenter Arten drastisch unterschätzt wird und dass ein großer Teil der Vielfalt noch darauf wartet, entdeckt und beschrieben zu werden (Engel et al., 2021; González-Oreja, 2008). Diese unerforschte Diversität zu erfassen stellt jedoch eine große Herausforderung dar, denn der Großteil der verborgenen Vielfalt (und Biomasse) des Planeten findet sich in Gruppen, die schwer zu untersuchen sind und daher in der Vergangenheit nur wenig taxonomische Aufmerksamkeit erhalten haben (sogenannte "Dark Taxa") (Hardulak, 2020; Hartop et al., 2022; Hausmann et al., 2020; Meier et al., 2021; Morinière et al., 2019). Dark Taxa sind so hyperdivers, dass präzise Artenzahl-Schätzung nur schwierig zu realisierbar sind - daher werden deise auch als "open-ended" Taxa bezeichnet (Hartop et al., 2022).

Der Großteil dieser “Dark Taxa” reiht sich in eine der beiden Ordnungen Diptera (Fliegen) und Hymenoptera (Ameisen, Bienen, Wespen) ein. Diesen Taxa gehören zum Beispiel Mücken, Gnitzen und parasitoide Wespen an (Hausmann et al., 2020). Oftmals kommen Dark Taxa in hohen Individuenzahlen vor, sie sind winzig klein (z.T. <2 mm) und weisen besonders hohe Raten kryptischer Vielfalt auf. Diese Faktoren erschweren die Anwendung traditioneller morphologischer Methoden zur Artenbestimmung und erklären, warum diese Gruppen in der Vergangenheit wissenschaftlich weitestgehend unterrepräsentiert wurden. Gleichzeitig weist die Tatsache, dass Dark Taxa in Proben so häufig vorkommen (sie können mehr als 70 % der Individuen in einer Probe ausmachen) auf eine wesentliche ökologische Rolle hin. Dies bedeutet im Umkehrschluss, dass es umso wichtiger ist, diese Gruppen für die Forschung greifbar zu machen, damit sie in Naturschutzmaßnahmen einbezogen werden können.

Wenn man Arten effektiver studieren und beschreiben will, braucht man eine taxonomische Renaissance der beschreibenden Taxonomie (Giangrande, 2003). Morphologische Methoden, die in

den letzten 250 Jahren angewandt wurden, liefern allein für sich keine ausreichenden Unterscheidungsmerkmale für die winzigen, kryptischen und vielfältigen Arten der dunklen Taxa. Glücklicherweise bieten die jüngsten Fortschritte in der Molekularbiologie die dringend benötigten Möglichkeiten, um die Entdeckung neuer Arten zu beschleunigen (Morinière et al., 2016). Diese treiben nicht nur die Probenbearbeitung und Artenbestimmung drastisch an, sondern ermöglichen auch die Analyse ganzer Insektengemeinschaften in einem Arbeitsgang. Darüber hinaus empfehlen immer mehr Wissenschaftler die Verwendung integrativer Arbeitsabläufe, bei denen Methoden aus verschiedenen Disziplinen zur Artbeschreibung und -abgrenzung eingesetzt werden. Die Anwendung dieses komplementären Ansatzes erhöht die wissenschaftliche Aussagekraft, da keine einzelne Methode für sich fehlerfrei ist (Dayrat, 2005; Schlick-Steiner et al., 2010).

Das Hauptziel dieser Arbeit ist daher die Entwicklung eines integrativen Workflows für die schnelle Verarbeitung großer Proben von Dark Taxa. Zu diesem Zweck werden drei spezifische Ziele verfolgt, nämlich (1) die Identifizierung der unbekannten Diversität in gemäßigten Regionen, (2) die Entwicklung einer integrativen Methode zur Aufarbeitung dieser Insekten und (3) die Prüfung der Verwendbarkeit von Ethanol in Insektensammelproben für Metabarcoding-Anwendungen.

Das erste Ziel ist es, das Bewusstsein zu schärfen für die weite Verbreitung von unbekannter Diversität nicht nur in tropischen, sondern auch in gemäßigten Zonen (Publikationen I-III). Dafür wurden die Metabarcoding-Daten von mehreren Insekten-Aufsammlungen analysiert. Hiermit konnte das hohe Vorkommen von Dark Taxa in Malaisefallen-Proben (im Sinne der Diversität, sowie großer Individuenanzahl pro Spezies) gezeigt werden. Außerdem wurden die Artenzahlen für vier Dipteren-"Dark Taxa" in Deutschland auf der Grundlage der gesammelten Daten extrapoliert.

Nachdem allgemein die verborgene Vielfalt der "Dark Taxa" in gemäßigten Regionen beleuchtet wurde, konnte im nächsten Schritt eine Strategie für die Bearbeitung eines spezifischen Dark Taxons

in großen Sammelproben entwickelt werden (Publikation IV): anhand der Chironomidae als Modellgruppe wird ein integrativer Ansatz vorgeschlagen, der (i) eine dreistufige Unterbeprobung zur Verringerung des Arbeitsaufwands bei der Probenverarbeitung, (ii) die parallel morphologische und (iii) DNA-basierte Methoden zur Bewertung der Artenvielfalt beinhaltet. Im Anschluss (iv) wird die Untersuchung möglicher Unstimmigkeiten zwischen den Methoden untersucht. Die Ergebnisse zeigen, dass sich unter der Verwendung dieses integrativen Ansatzes mehr als 90 % aller Arten nachweisen ließen, nachdem vorab weniger als 10 % aller Individuen morphologisch identifiziert wurden. Zusätzlich wäre die alleinige Anwendung einer der beiden Identifizierungsmethoden anfälliger für Fehler gewesen, die wahrscheinlich unentdeckt geblieben wären.

Abschließend wird die Verwendbarkeit von in Ethanol gelöster DNA für (klassischerweise destruktive) Metabarcoding-Anwendugen bewertet. In dieser methodischen Arbeit wird geprüft, ob ökologische Informationen in der DNA erhalten bleiben, die aus der Sammelflüssigkeit von Insektenproben extrahiert wird. Sollte dies der Fall sein, würde dies bedeuten, dass das Fixativ einfach abgegossen und für die Analyse verwendet werden kann, während alle Individuen für weitere Analysen intakt bleiben. Die Resultate zeigen, dass DNA auf Ethanolbasis keine ökologischen Informationen bewahrt. Bis zukünftige, tiefgreifendere Forschung erfolgreichere Ergebnisse liefert, wird es deshalb empfohlen, bei der Verwendung von Ethanol-DNA vorsichtig walten zu lassen.

Zusammenfassend schafft diese Arbeit einen Rahmen, der verschiedene Disziplinen kombiniert, um die immense (verborgene) Insektenvielfalt, die in unseren gemäßigten Klimazonen beheimatet ist, trotz der wachsenden taxonomischen Hindernisse effizient zu untersuchen.

INTRODUCTION

Not all superheroes wear capes

Despite their small size and often inconspicuous nature, insects account for a large proportion of the animal biomass and diversity in terrestrial systems (Leandro & Jay-Robert, 2019; Wilson, 1987). In fact, more than half of all described species on earth are insects (Mayhew, 2007) and as expressed by Robert M. May (1988), *"To a rough approximation, and setting aside vertebrate chauvinism, it can be said that essentially all organisms are insects*." Today, about one million insect species are recorded worldwide (Morse et al., 2017). Despite this high number, it has become apparent rather early-on that a large proportion of the global insect fauna still remains undiscovered, and estimates in literature range considerably from 2 to 80 million (Erwin, 1983, 1991, 2004; Gaston, 1991; Hodkinson & Casson, 1991; May, 1988; Ødegaard, 2000; Stork, 2018).

We need insects more than they need us

Insects dominate terrestrial systems, not only in terms of species numbers but also in their abundance (Kremen et al., 1993). They play key roles in ecosystem functioning which have a direct effect on all surrounding environments and organisms, including us humans (Leandro & Jay-Robert, 2019; Prather & Laws, 2018; van Huis, 2014; Wilson, 1987). Such functions include soil aeration, enhancing agricultural productivity, plant pollination, seed dispersal, pest regulation, decomposition, and nutrient cycling (Nichols et al., 2008; Ramos et al., 2020). Moreover, insects themselves serve as an important food source for a large range of taxa across different biospheres. For many amphibians, reptiles, birds, and mammals, insects even represent the sole food source, making them key organisms in food chains and food webs (Scudder, 2017; Shurin et al., 2005).

Insects build the foundation of our environments by providing conditions for all organisms to thrive (Duffus et al., 2021; Scudder, 2017) and were they to disappear, humanity would not sustain itself for very long (Morse et al., 2017). In fact, Grimaldi & Engel (2005) describe quite frankly in their book that, *"People gladly imagine a life without insects. But if ants, bees and termites alone were removed from the earth, terrestrial life would probably collapse. Most angiosperms [flowering plants] would die out, the ensuing plant wreckage would molder and ferment for lack of termites, soil depleted of nutrients would barely be able to sustain the remaining plants; erosion would choke waterways with silt. Vast tropical forests of the Amazon, Orinoco, Congo, and other river basins would die off, and the earth's atmosphere and oceans would become toxic."*

We depend on the services provided by insects and contrary to popular belief, the majority of these are provided by wild species (Gill et al., 2016; Jordan et al., 2021; Losey & Vaughan, 2006). Still, insects do not receive the societal acknowledgement and awareness that they truly deserve and instead, are perceived among many as pests, disease vectors, or a nuisance to humans (Cardoso et al., 2020; Fukano & Soga, 2021). With this low popularity and the overall lack of appreciation for these organisms and their interactions, it is not surprising that insects have a lower priority in conservatory actions (Cardoso et al., 2020; Fukano & Soga, 2021; Garibaldi et al., 2014). To raise awareness for the importance of insects, Losey and Vanghan (2006) went ahead and provided estimates for the economic value of four ecological services provided by wild species for which data were available, namely (i) dung burial, (ii) pest control, (iii) pollination, and (iv) wildlife nutrition. Their calculations were based on the projections of losses in the absence of these services, and their results indicate an annual value of at least $57 billion in the United States just for these four services (Losey & Vaughan, 2006). Further calculations reveal that worldwide, insect pollination alone has an annual economic value of $153 billion, and that the value of crops that depend on insect pollination is five-fold that of those that do not (see Gallai et al., 2009).

Saving biodiversity, the wrong way

In 1992, the Convention on Biological Diversity (CBD) was signed by 150 government leaders at the Rio Earth Summit in order to study, halt, and possibly reverse the ongoing negative environmental trends at a global scale (Chandra & Idrisova, 2011; Essl et al., 2020; Liu et al., 2011). Following this convention, various large-scale and long-term initiatives have been launched worldwide with the goal of halting species extinction and promoting sustainable development (Donaldson et al., 2016). Since then, the number of academic publications related to biodiversity research have skyrocketed, with "conservation" being a keyword across works (Liu et al., 2011). In Europe, for example, the Habitats Directive (92/43/EEC) was adopted by the European Union (EU) in 1992 to conserve its wild flora, fauna, and habitats (Mammides, 2019; Mammola et al., 2020; Martín-López et al., 2011). To allocate the EU's conservation budget for the funding of these conservation projects (LIFE projects) across all member states, the LIFE Program was installed (Mammides, 2019; *Life Croaa*, 2023). Since the start of the directive, more than €3.1 billion have been contributed to the protection of the environment and its species (*Life Croaa*, 2023).

Despite these noteworthy efforts directed at conserving biodiversity, studies have demonstrated that the bulk of European species will actually not benefit at all from these measures because there is a strong taxonomic bias in the EU's legislative and funding allocation (Mammides, 2019; Mammola et al., 2020; Sánchez-Fernández et al., 2018). Mammola et al. (2020) were able to show that funding was six-fold higher for vertebrates, although invertebrates represent 79% of all species globally (Mammides, 2019). Instead of targeting groups based on their species numbers, extinction rate, or ecological importance, efforts in the directive were primarily driven by species' popularity among society (Mammola et al., 2020). Unfortunately, these taxonomic biases are not unique to the EU and are in fact very widespread across all conservation efforts worldwide (Donaldson et al., 2016; Mammides, 2019; Restani & Marzluff, 2002). Many studies have addressed this problem and argue that

biodiversity loss will continue if these biases are not addressed soon (Clark & May, 2002; Fazey et al., 2005; Rands et al., 2010; Restani & Marzluff, 2002; Titley et al., 2017). Ultimately, insects need to become a major component of conservation and management planning in order to guarantee a sustainable future for all generations to come (Cardoso et al., 2020; Donaldson et al., 2016), and it is imperative that actions are taken immediately. For that, more funds need to be allocated towards insect research.

A look at traditional insect research

The biggest challenge in insect research, be it the study of insect trends, diversity, abundance or distribution, is the scarcity of baseline information (Cardoso et al., 2011; Cardoso & Leather, 2019; Eisenhauer et al., 2019; Montgomery et al., 2020). This is due to the confluence of several factors, many of which are associated with resource constraints, technological limitations, taxonomic complexity, and research priority (Montgomery et al., 2020).

In general, insect research and/or monitoring consists of (1) insect sampling, (2) sorting for specimens of interest, (3) species identification, and (4) data analysis (Montgomery et al., 2021). Although insect sampling is rather straightforward, subsequent processing and analysis of the samples' contents can be very demanding.

Insect sampling with Malaise traps

The Malaise trap (Malaise, 1937) is one of the most widely used traps among entomologists (Campbell & Hanula, 2007; Uhler et al., 2022; Vårdal & Taeger, 2011). Malaise traps are tent-like structures made of fine mesh netting which capture insects that fly into (or climb up) the tent's intercepting inner wall, ultimately dropping into a collection bottle at the very top that is filled with ethanol (Fig. 1) (Gressitt & Gressitt, 1962; Uhler et al., 2022; Vårdal & Taeger, 2011). This method of trapping takes advantage of

the fact that insects always fly or crawl upwards after encountering an obstacle (Sheikh et al., 2016). The ethanol fumes anesthetize the insects, while the liquid ethanol enables tissue preservation until further steps are taken. Full collection bottles can be unscrewed at any time and easily replaced with the fresh ones, which is often done in a weekly or biweekly rhythm during the high season. Malaise traps are easy to use and most importantly, enable the passive collection of many individuals in a standardized manner, all with relatively low effort from staff (Campbell & Hanula, 2007; Matthews & Matthews, 2017; Skvarla et al., 2021). They are particularly effective for assessments of the local insect communities, especially over longer periods of time (Uhler et al., 2022).

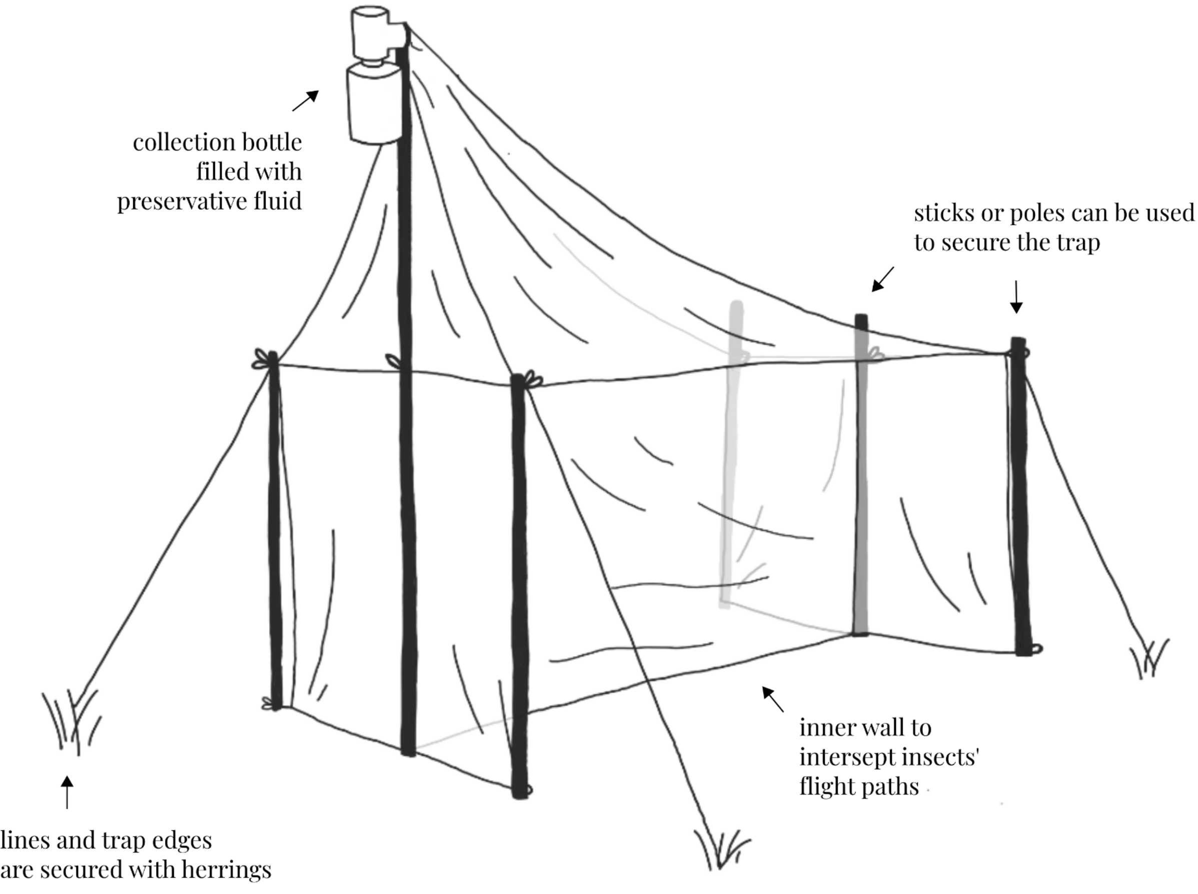


Figure 1. Schematic representation of a Malaise trap. Created by C. Chimeno with Sketchbook.

Processing Malaise trap bulk samples

Malaise traps are very effective at catching insects so consequently, samples will often contain thousands of individuals spanning across a wide range of taxa that need to be looked through (Dewaard et al., 2018; Klink et al., 2022; Morinière et al., 2016). The process of sample sorting is a major obstacle in ecological research because expertise or manpower are often insufficient, making the procedure arduous and inefficient (Miller et al., 2016; Wang et al., 2018). Many scientists therefore employ parataxonomists for support. The term "parataxonomist" was originally coined by Daniel Janzen to describe jobs held by trained individuals with a broad understanding of insect diversity who assist scientist and researchers by collecting, preparing, and sorting specimens for further scientific study (see Janzen, 1991). In this regard, parataxonomists (students, interns) are often incorporated into biodiversity assessments where the workload is especially burdensome. In practice, samples are presorted by parataxonomists at the order- or family-level before being passed on to a respective expert for more detailed sorting and identification (Abadie et al., 2008; Basset et al., 2000; Janzen, 1991; Krell, 2004; Schmiedel et al., 2016). However, using different researchers across laboratories for sorting can be problematic due to subjectivity in identification and sorting criteria. Also, the quality of sorting is highly dependent on the individual's experience, so reproducibility is difficult to achieve.

Species identification

For over 250 years, morphology-based determination techniques have been the golden standard in taxonomy, thus forming the basis for all hypotheses on phylogenetic relationships (Dunn, 2003; Hardulak, 2020; Mallet & Willmott, 2003). Because of this, insect diagnostics tools largely rely on physical characteristics of specimens. When an insect is collected, it is identified by a taxonomist to the lowest taxonomic level by assessing its morphological characters and comparing it to a taxonomic key (Badirli et al., 2023). This procedure can be relatively straightforward when dealing with few individuals or wanting to identify larger, more conspicuous species of well-resolved groups (i.e.,

butterflies, beetles). However, species-level identifications of large insect bulk samples are challenging, time-consuming, costly, and sometimes even impossible (Morinière et al., 2016; Yu et al., 2012). For these reasons, there is a heavy bias towards the analysis of well-resolved groups for which baseline knowledge is available (Meier et al., 2023), whereas others are set aside (Badirli et al., 2023). Ultimately, only a tiny proportion of a sample is analyzed while the rest is either discarded or left on a shelf with the hope that future technologies will one day enable access to species-level information (Bohan et al., 2017; Keck et al., 2017; Mandelik et al., 2010; Porter et al., 2014).

Despite morphology being the go-to-method for a long time, it does have limitations. For example, morphological identifications are only applicable if the physical characters that are needed for identification are well-preserved. This is a problem when working with delicate specimens that are easily damaged during sample processing. Moreover, most identification keys are based on the study of male individuals only, providing no diagnostics for the identification of females (which are often more abundant in populations) (Ekrem et al., 2010). Most importantly, species that are cryptic diverse (visually indistinguishable from each other) can be falsely identified as one species, whereas high visual variations within species can lead to erroneous multiple identifications (Ekrem et al., 2010; Hardulak, 2020). It is also important to keep in mind that the quality of morphology-based identifications relies heavily on the availability and accuracy of identification keys, and the identifier's ability to conduct identifications from an objective point of view (Ekrem et al., 2010). Identifications are subject to disagreement when experts use different methods to distinguish particular traits, or use a different terminology which can be an extra source of confusion (Hardulak, 2020; Pirkola et al., 2001; Vogt et al., 2010).

Expediting insect research with molecular methods

In the last decades, molecular identification technologies have arisen that are now drastically expediting taxonomy and its applications in various fields of science (Elbrecht et al., 2017). These techniques can bring more objectivity into species diagnostics, as delimitations are not based on visible characters but on the information provided by an individual's DNA instead (Bukowski et al., 2022). Just as different species show differences in their morphology, they also show differences in their DNA (Wilson et al., 2019). In general, variations in the DNA are larger between specimens of different species (higher interspecific variation) than among individuals of the same species (lower intraspecific variation), making it possible to distinguish them from one another (Čandek & Kuntner, 2015; Gibbs, 2018). The difference between the greatest intraspecific distance and the smallest interspecific distance is also known as the "barcode gap", which is used to distinguish species from one another (Meier et al., 2008; Phillips et al., 2022).

DNA barcoding

DNA barcoding was first introduced by Paul Hebert in 2003 to provide a standardized and accelerated approach at species identification and discovery (Hebert et al., 2003). It uses a short standardized DNA fragment, also known as a DNA barcode, to identify species (Hollingsworth, 2011; Jinbo et al., 2011; Stoeckle & Hebert, 2008). Hebert proposed the 658 base pair (bp) long DNA sequence of the Cytochrome C Oxidase Subunit I gene (COI) to be used as a universal barcode fragment (Hebert et al., 2003). This gene is especially advantageous as an identification marker because it is located in the mitochondrial DNA (mtDNA) instead of the nuclear DNA (Dawnay et al., 2007; Hashemi-Aghdam et al., 2017). Because of this, it is present in high copy numbers in every body cell of all animals (Castellani et al., 2020; Hardulak, 2020). It is easily accessible and short enough to enable fast and easy sequencing, yet long enough to provide a robust and unique identification marker to distinguish

species from one another (Andújar et al., 2018; Hardulak, 2020; Imtiaz et al., 2017). Since the mitochondrial genome is responsible for the encoding of essential proteins, it is highly conserved among mammals, is haploid, intron-free, and is less susceptible to genetic recombination (Clayton, 1992; Hardulak, 2020; Raffoul et al., 2012).

DNA barcoding is conducted using basic procedures that can be performed in any sterile laboratory (Imtiaz et al., 2017). The main steps include tissue lysis (breaking up the body cells), DNA extraction (isolating the DNA and removing cell debris), amplification (duplicating the DNA), purification (cleaning-up the DNA), and Sanger sequencing (unlocking the DNA code) (Fig. 2) (Imtiaz et al., 2017). Nowadays, sequencing is often outsourced to a commercial company where it is completed in an even more standardized and cost-efficient manner (Coissac et al., 2016; Costion et al., 2011; Touchman, 2009). Sometimes, companies also provide the service of post-sequencing bioinformatics so essentially, end-users have very little to do themselves. Once the sequences are cleaned, they are compared to a reference library for identification (Imtiaz et al., 2017).

One of the greatest advantages of DNA barcoding is the fact that any nature of tissue sample can be used for analysis as long as the DNA quality is sufficiently conserved (Imtiaz et al., 2017). DNA barcoding is easy to use (even for non-experts), widely available, and is nowadays a relatively low-cost technique (Baloğlu et al., 2018; Grant et al., 2021; Hartop et al., 2022; Hausmann et al., 2013; Hebert & Gregory, 2005; Jinbo et al., 2011). Today, it is applied in a wide range of scientific fields, including forensic entomology, food security, biomonitoring of pest and/or invasive species, to monitor poaching, or for border control (see Chimeno et al., 2019; Gorini et al., 2023; Gupta, 1994; Kaur, 2015; Littlefair & Clare, 2016; Rolo, 2010; Schilthuizen et al., 2011; Shadrin, 2021; Wells & Stevens, 2008; Wetton et al., 2004; Wu et al., 2005).

With DNA barcoding successfully identifying species across a large taxonomic and spatial scale, extensive campaigns have followed world-wide (Brown, 2021; Brown et al., 2018; Taberlet et al., 2012) and with ever more data being uploaded to BOLD and other databases, DNA barcoding is becoming more and more robust (DeSalle & Goldstein, 2019). Its application is extensive, providing more in-depth analysis of otherwise very difficult-to-study organisms, so consequently, its use in biomonitoring surveys has gained enormous traction throughout the decades. However, the mass of individuals in insect surveys that need to be processed and identified is still a large constraint, because DNA barcoding via Sanger sequencing only allows for the analysis of one specimen at a time (Cristescu, 2014; Shokralla et al., 2012).

Yet in recent years, sequencing technologies have sustained massive improvements with the development of high throughput sequencing (HTS), which are platforms that drastically increase the sequencing capacity (Qiang-long et al., 2014; Soon et al., 2013; Taberlet et al., 2012). With HTS, billions of sequencing reads are provided in just one single reaction, corresponding to an elevation of at least five orders of magnitude when compared to Sanger sequencing techniques (Soon et al., 2013; Taberlet et al., 2012).

DNA metabarcoding

DNA metabarcoding is an extension of DNA barcoding that uses this HTS platform (Chimeno et al., 2019; Elbrecht et al., 2017; Piper et al., 2019; Taberlet et al., 2012). Now, instead than analyzing specimens one at a time, multiple bulk samples each containing hundreds of specimens can be processed simultaneously (Fig. 2) (Aylagas et al., 2016, 2018; Cristescu, 2014; Yu et al., 2012). Instead of targeting a single species, metabarcoding aims at identifying the species composition of a sample (Beng et al., 2016; Bush et al., 2019; Liu et al., 2011). With this technology, entire communities can be screened for taxonomic diversity holistically in a highly standardized, reliable, and cost-efficient

manner, enabling more comprehensive analysis of entire samples should adequate reference libraries be available (Hajibabaei et al., 2012; Hardulak, 2020; Morinière et al., 2016; Yu et al., 2012). At the same time, metabarcoding is relaxing the bottleneck of ecological research because there is no need to individually sort samples prior to processing because the biomass of the sample is processed as an entirety (Beermann et al., 2018; Elbrecht & Steinke, 2019).

In the laboratory, the sample's contents are dried, and the biomass is (traditionally) homogenized into a fine powder to release as much DNA as possible (Hardulak, 2020; Mata et al., 2021; Zizka et al., 2022). An aliquot of the powder is taken for tissue lysis which is followed by DNA extraction, DNA amplification, and high throughput sequencing (Elbrecht & Steinke, 2019; M. Liu et al., 2020; Morinière et al., 2016). After sequencing, a bioinformatic pipeline is applied for filtering and quality procedures, and the sequences are clustered based on their similarity into Molecular Operational Taxonomic Units (MOTUs) (Boyer et al., 2016; Brandt et al., 2021; Buchner et al., 2022; Elbrecht et al., 2017). These MOTUs are compared to a reference library for taxonomic assignment (Keck et al., 2018) and at the end, an MOTU-(or taxon, after reference library comparison)-by-sample matrix is created for data analysis (Ji et al., 2013). Essentially, the end-user obtains a dataset that indicates which MOTUs (or taxa) were detected in which sample, and the number of reads recovered for each. In this case, abundance information is not conserved in the output meaning that scientists work with incidence data (Deagle et al., 2019; Lamb et al., 2019).

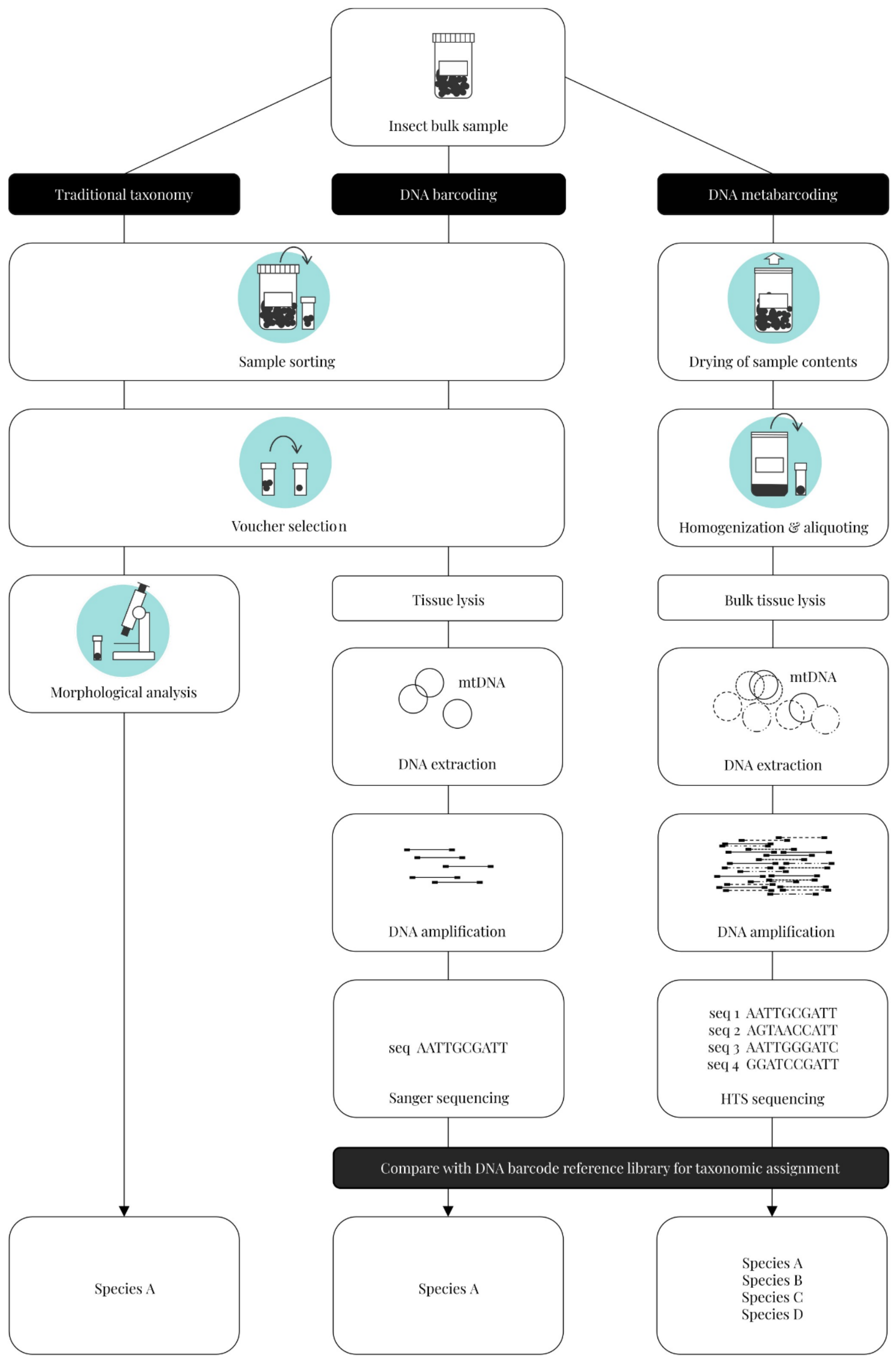


Figure 2. Species identification workflows. Schematic representation of the species identification workflows using (left) traditional morphology, (middle) DNA barcoding and (right) DNA metabarcoding.

Understanding BOLD & the BIN system

Accurate identifications via DNA barcoding are only guaranteed if a comprehensive and accurate reference library is at hand (Wilson et al., 2019). Therefore, the first and most crucial step in any DNA barcoding application is the establishment of a reference library with high species coverage (Chimeno et al., 2019; Liu et al., 2020; Morinière et al., 2016). Extensive public reference databases already exist, such as GenBank (Benson et al., 2012) and the Barcode of Life Data System (BOLD; Ratnasingham & Hebert, 2007) which are freely accessible to any researcher interested in DNA barcoding. BOLD is most widely used for COI-based data and includes barcode data not only for animals, but also for plants and fungi (Ratnasingham & Hebert, 2007). BOLD was officially established at the Centre for Biodiversity Genomics (CBG; Guelph, Ontario, Canada) in 2007 after Paul Hebert's plea for the wide-scale use of DNA barcoding (see Ratnasingham & Hebert, 2007) and today, BOLD contains ~12 million sequences belonging to ~250,000 formally described animal species (http://www.boldsystems.org/).

BOLD provides users with their very own private workbench to create projects, download and/or use public barcode records, upload their barcoding data, and conduct analyses using user-friendly tools that are directly integrated into the interface (Frézal & Leblois, 2008; Ratnasingham & Hebert, 2007). Upon upload, all sequences go through extensive quality checks before being made public (Ratnasingham & Hebert, 2007). In order for them to obtain a "formal" barcode status, further conditions need to be met (i.e., species name, identifier name, voucher storing information, collection data, length >500 bp, primer information, and raw sequence files) (Meiklejohn et al., 2019; Ratnasingham & Hebert, 2007).

Once uploaded, BOLD applies an algorithm, Refined Single Linkage (RESL), to cluster sequences based on their similarity with a taxon-specific threshold (Ratnasingham & Hebert, 2013). Clusters of similar sequences are then assigned an Operational Taxonomic Unit (OTU), which acts as a species proxy.

However, in order to use OTUs universally, they are given a unique identifier name, the Barcode Index Number (BIN) (Ratnasingham & Hebert, 2013). Every sequence ≥300 bp is automatically assigned to a Barcode Index Number (BIN) that is already in BOLD if sequence similarity based on the RESL-algorithm is below a given threshold (Ratnasingham & Hebert, 2013). Sequences with ≥500 bp which do not find a match, serve as founders of new BINs. The BIN system is a dynamic algorithm, and as more sequences are uploaded providing more intra- and interspecific information, the algorithm is rerun so that BIN assignments become more concordant (Ratnasingham & Hebert, 2013).

Peering into the unknown

Severe Taxonomic Gaps

With the increase in global DNA-based applications, ever more barcode sequences are being regularly uploaded to reference libraries, producing a wealth of computable data that is accessible to any and every researcher world-wide (Page, 2016). However, many sequences cannot be linked to a traditional species name because reference libraries are incomplete (Morinière et al., 2016; Page, 2016; Virgilio et al., 2010). In order to close this taxonomic gap, traditional species descriptions need to happen at a faster pace than that of sequence data generation. However, not only do taxonomic procedures require more time, taxonomic expertise is overall in continuous decline (Hausmann et al., 2020; Keck et al., 2018; Page, 2016). This is in part due to the fact that taxonomy is viewed among many as a marginal science, so too few funds are allocated for taxonomic work or the training of new taxonomists (de Carvalho, 2005; de Carvalho et al., 2007). This imbalance between the high number of species awaiting description and/or identification and that of very few experts capable of doing so is called the taxonomic impediment (Engel et al., 2021; Giangrande, 2003; Hardulak, 2020; Hausmann et al., 2020; Morinière et al., 2019). Active discussions about the taxonomic impediment have been

ongoing for almost three decades, concluding that in reality, we have very little knowledge about the true diversity of species on Earth (Engel et al., 2021; González-Oreja, 2008).

The main issue with recording Earth's unknown biodiversity is that the bulk of it is found in groups that have received very little taxonomic attention in the past, either because they are difficult to study or because it is assumed that they have little economic importance or applied research value (Hartop et al., 2022). However, some of these groups are so species-rich and abundant, that they conversely must conduct essential functions in ecosystems and thus need to be included in biodiversity research (Hartop et al., 2022).

Dark Taxa

Originally used by Page (2016) to refer to all unnamed sequences in DNA reference libraries, the term "dark taxa" has evolved to address all of these species-rich groups for whom most species are left undescribed (Hartop et al., 2022; Hausmann et al., 2020). It is suspected that a large proportion of these "open-ended" dark taxa are found within two insect orders, Diptera (flies) and Hymenoptera (ants, wasps, bees) (Hausmann et al., 2020). Within these orders, dark taxa mostly include families of non-brachyceran Diptera (primarily mosquitoes, gnats, midges) and nearly all families of parasitoid Hymenoptera (wasps). One large-scale study conducted by Paul Hebert and his team demonstrates the magnitude of diversity among such groups (Hebert et al., 2016). After barcoding more than one million insects, they were able to conclude that just for one family of flies (Cecidomyiidae; gall midges), a total of two million species are estimated to occur world-wide (Hebert et al., 2016), which already exceeds the number of currently described animal species.

Making dark taxa tangible to science using traditional morphology-based methods alone is not feasible, and as just recently stated by Meier et al. (2023), they require a completely different set of taxonomic protocols for assessment. One major constraint is the lack of taxonomic expertise for these

groups (Hardulak, 2020). The distribution of expertise among insect taxa is highly uneven, with (too) many working on groups that are well-resolved, and too few working on dark taxa (Boero, 1996; Giangrande, 2003). Yet even with available experts, procedures would be much too time-consuming and demanding (Hartop et al., 2022). Most specimens of dark taxa are miniscule (<2 mm), so identifications require meticulous preparation of individuals such as dissection and mounting on microscopic slides (Ekrem et al., 2010; Jaschhof & Jaschhof, 2022). The abundance of these tiny specimens in samples makes the workload especially overwhelming; in environmental- and bulk samples, dark taxa can make up to 70% of all specimen numbers, representing thousands of small-bodied, similar-looking individuals (*GBOL III*, 2023).

As mentioned by Giangrande (2003), we are in dire need of a taxonomic renaissance in descriptive taxonomy if we want to study species diversity more effectively. We cannot apply the same methods that have been used for the last 250 years unchanged because morphological methods alone do not provide sufficient discriminatory information for the tiny, cryptic diverse species of dark taxa. As highlighted by Hartop (2022), who is dedicated to tackling the Phoridae (also a dark taxon of flies), studying hyperdiverse groups requires a multilevel approach (Hartop et al., 2022). Scientists have been proposing the use of integrative taxonomy in recent years, which consists of applying multiple methods in parallel to accurately delimitate and identify species (Dayrat, 2005; Meier et al., 2006; Schlick-Steiner et al., 2010; Will et al., 2005). Instead of dismissing morphology-based methodologies, these are complemented with methods from different disciplines in one workflow. Integrative workflows increase vigor because ultimately, no single methodology is error-free (Dayrat, 2005; Meier et al., 2006; Schlick-Steiner et al., 2010; Will et al., 2005). In recent years, various integrative pipelines have been developed in order to tackle a much larger number of specimens and/or species in a rapid and cost-efficient manner (Morinière et al., 2016; Srivathsan et al., 2018; Wang et al., 2018).

The German Barcode of Life Project

With the launch of its third phase in 2020, the national German Barcode of Life (GBOL) project is currently dedicated to studying several dark taxa of Diptera and Hymenoptera using the latest methods of integrative taxonomy. The consortium includes researchers, PhD students, and students from the Zoological Research Museum Alexander Koenig (ZFMK) in Bonn, the State Museum of Natural History Stuttgart (SMNS) in Stuttgart, the Bavarian State Collection of Zoology in Munich (SNSB-ZSM), the Department of Animal Ecology and Tropical Biology of the University of Würzburg, and the Entomological Society of Krefeld. Moreover, the project makes use of countless intra- and international cooperations from external experts to implement as much taxonomic and ecological expertise as possible into its research. It is in the framework of GBOL III that research for this thesis has been conducted, also including the analysis of data that were collected in the previous phases of the project.

GLOBAL DISCUSSION

As expressed by Meier et al. (2006), “several of the biggest challenges in taxonomy and systematics are related to a toxic mixture of small size, abundance, and rarity”, making the analysis of species-rich arthropod bulk samples especially difficult. This “toxic mixture” refers to non-other than the bulk of cryptic and megadiverse dark taxa that were targeted in this thesis. Essentially, very little is known about their true species numbers and ecological functions, so effectively tackling these insects requires different approaches than those that have been used for the last 250 years (Hartop et al., 2022; Meier et al., 2022). Instead of applying conventional morphology-based sorting and identification protocols, large-scale DNA barcoding and metabarcoding techniques were applied to assess the sampled biodiversity. In this manner, the means were provided to conduct species delimitations, identifications, and conversely, diversity assessments in a rapid, cost-efficient, and more objective manner.

Small, abundant, rare

In Germany, taxonomic research in entomology has a long history, with notable contributions from renowned scientists spanning several centuries, including naturalists Carl Linnaeus (1707–1778) and Johann Christian Fabricius (1745–1808), entomologists Johann Wilhelm Meigen (1764–1845) and Ernst Heeger (1783–1866), and zoologist Willi Hennig (1913–1976) (Boventer, 1960; Carpenter, 1945; Dupuis, 1984; Manktelow, 2010; Sneli et al., 2009). Today, Germany houses extensive insect collections in various national museums and research institutes, as well as entomological societies and organizations that promote the research and taxonomy of insects, also among citizen scientists. Overall, its fauna is considered as being well studied (Hausmann et al., 2020), yet the results of this thesis contradict this statement as they show that there are severe knowledge gaps that need to be

addressed. Interestingly, this work shows that almost as many species are awaiting description in Germany as in the tropics – the tropics, which are renowned for their very high species richness while being severely understudied (Basset et al., 2012). For both settings (Sumatra and Germany), almost identical proportions of dark taxa were obtained from samples in terms of abundance and biodiversity (Fig. 3). These findings evince that so much less is known about the true diversity of the German insect fauna than often assumed. Therefore, it is imperative for research efforts to become more evenly distributed across taxa, also in relation to their estimated diversity.

In *Publication III*, a closer look was taken at just four dark taxa of flies (Cecidomyiidae, Chironomidae, Phoridae, and Sciaridae) to provide data-based species number estimates for Germany. Based on the calculations, 1,800 new species are expected (at the very least) just for these four taxa in Germany. This number already increases the current species count of Diptera in Germany by almost 20%. Moreover, these results demonstrate that contrary to popular belief, researchers don't have to travel to remote areas to witness this remarkable array of insects – they are present in both urban and rural areas, right on our doorsteps. For example, in *Publication II*, an extreme diversity of Cecidomyiidae (gall midges) was recovered from samples collected in an urban setting: The institute's premises which is nested in a residential neighborhood adjacent to a busy street that leads to a large interstate. Because only half of the collected samples were processed, it is believed that the true diversity of Cecidomyiidae is in fact much higher at this site and that the true diversity is in fact even higher. Among the four taxa that were tackled in *Publication III*, Cecidomyiidae displayed *by far* the highest recovery ratios. Merging data from *Publications I-III* revealed that 13% of all processed specimens and 20% of all recovered BINs were associated with just this one taxon. Considering these and similar findings by Morinière et al. (2019) and Hebert et al. (2016) alike, Cecidomyiidae may be by far the most diverse and species-rich taxon among all Diptera (Ševčík et al., 2016). In fact, following their analysis of over one million Canadian insects, Hebert et al. (2016) concluded that at least two million species

of Cecidomyiidae may be found world-wide. This estimate already exceeds the number of all described species on Earth.

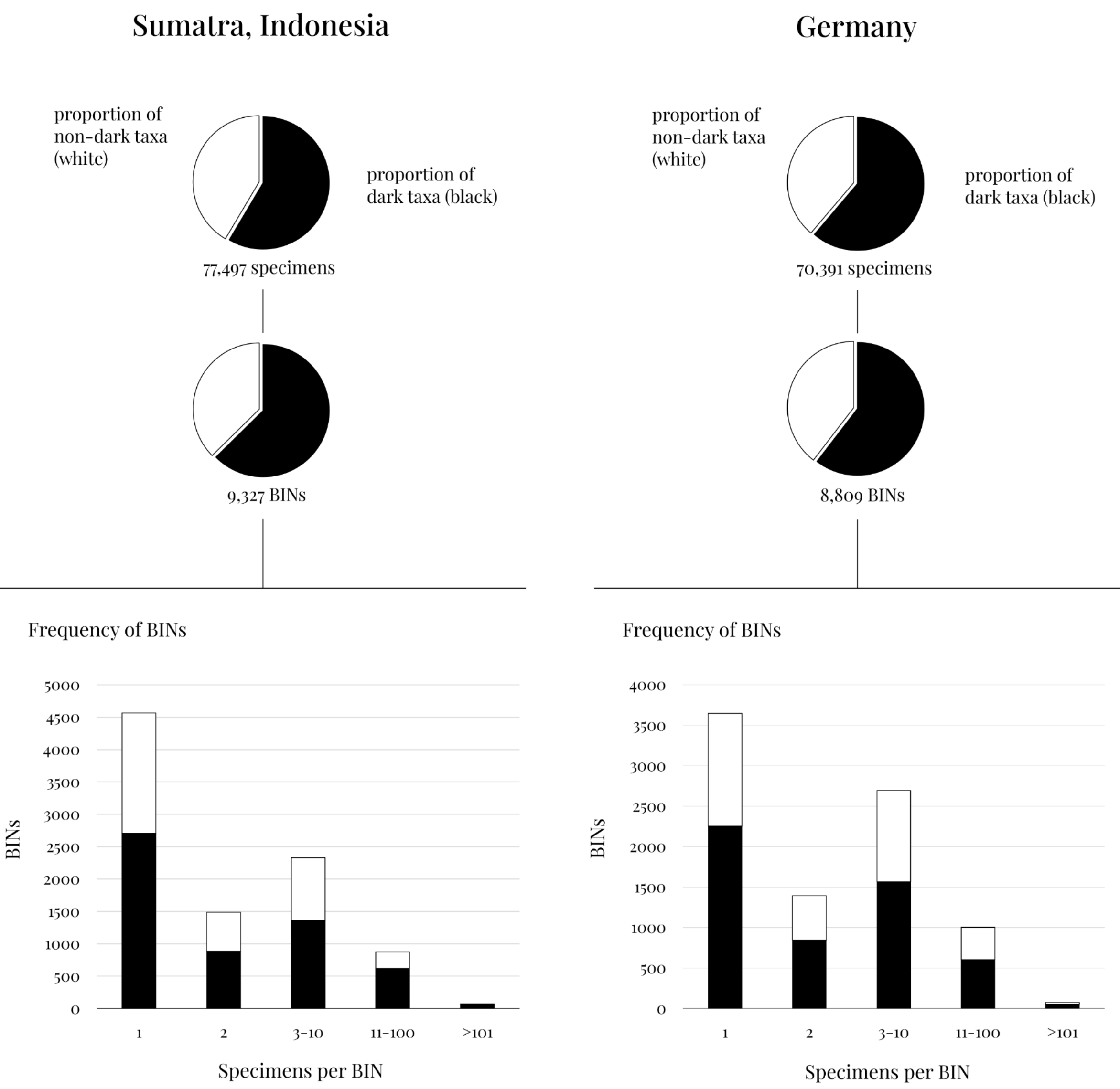


Figure 3. Proportion of dark taxa and BIN frequency from samples collected in Sumatra and Germany.

To understand the driving force behind this hyperdiversity, Morinière et al. (2019) examined the average interspecific variations of COI among various families of Diptera. Here, the authors were able to show that interspecific variations are much lower among species of Cecidomyiidae than among

those of other families, which implies an increased evolutionary rate. This is thought to be caused by haplodiploidy (Hebert et al., 2016), a form of reproduction in which females develop from fertilized eggs and are diploid and males from unfertilized eggs and are haploid (Lohse & Ross, 2015; Normark, 2004). According to Patten et al. (2015), haplodiploidy leads to a decrease in nuclear introgression in relation to mitochondrial introgression, so this "seemingly increased" rate of mitochondrial introgression facilitates the generation and maintenance of new species (Lohse & Ross, 2015). This unusual form of reproduction has originated at least ten times in insects (Normark, 2003; Otto & Jarne, 2001) and is also present among the taxon Sciaridae (Normark, 2004), another megadiverse dark taxon for which a high recovery ratio was obtained in this thesis.

Dealing with the tip of the diversity-iceberg

Across all studies, results suggest that the trap sites were drastically undersampled. This fact, in combination with the high proportion of rare species that was recovered, implies that the true insect diversity at the sampling sites is in fact much higher. Ultimately, this thesis is just dealing with the tip of the diversity-iceberg and much work lies ahead to uncover the rest.

Examining BIN-frequencies across all datasets revealed that more than half of the recovered dark taxa (61%) are represented by one or two specimens only (Fig. 3). As stressed by Lim et al. (2012), 25-30% of all species are so rare that they have only been collected once (Bickel, 1999; Coddington et al., 2009; Novotny et al., 2007). On top of that, these are often hidden in species-rich samples that are filled to the brim with specimens of common species (Longino, 2009; Meier et al., 2016). With sequencing costs having drastically decreased in the last decade, DNA barcoding is more widely accessible (Coissac et al., 2016; Meier et al., 2016), yet applying large-scale Sanger sequencing to hundreds of thousands of specimens to uncover rare species remains costly. Against the backdrop that essentially only a fraction of the sample contents was processed, it is therefore assumed that a

multitude of rare species have been overlooked in this thesis. For example, in *Publication I*, some samples from Malaise trap 1 were processed completely, while for the rest of samples (including those from Malaise trap 2) a limit of 15 96-well-plates were filled which corresponds to 1,475 specimens each. From the premises of the institute, only every fourth sample was chosen for sequencing, and from these, 32,592 specimens were barcoded, which is only about 25% of collected specimens *(Publication II)*. In *Publication III*, every second sample from the Bavarian Forest National Park was selected for processing, and of the estimated 52,000 specimens that were inside, 29,481 specimens were barcoded, corresponding to 56% of all specimens. Sampling in the Allgäu Alps recovered well over a million specimens, but here, only 20,250 specimens were processed (2%).

Yet even if every single specimen were to have been analyzed from every single sample, the results indicate that the actual amplitude of diversity would have been underestimated regardless due to the limited sampling that was conducted in space (few traps) and time (short sampling season). In *Publication I*, the Malaise traps were operated for three months only. In tropical settings, where insects are present and mobile all year round, a three-month-sampling period provides very limited coverage of the temporal diversity. Also, sampling was conducted with just two Malaise traps which is very little considering that the tropics are spatially very heterogeneous, providing a wide array of microhabitats in x-dimensions (Basset et al., 2012). As demonstrated by Basset et al. (2012), the canopy is by far the most species-rich habitat of tropical forests. While Malaise traps are very effective at collecting arthropods (Matthews & Matthews, 2017), no additional sampling methods were applied to sample the diverse canopy communities that have been observed in other studies (Basset et al., 2012; Ozanne et al., 2003). With these traps, only arthropods that are found in the litter and understory habitats were targeted yet still, the results suggest that here, +38% (Malaise trap 1) and +44% (Malaise trap 2) more species could have been recovered with a doubled sampling effort. Interestingly, only 24% of species were shared between trap sites. The fact that more than 80% sample coverage was

recovered throughout this study indicates that arthropod diversity patchiness may be at play, which is common for tropical forests (see Milton and Kaspari, 2007).

In *Publication III*, sampling was done with 14 Malaise traps: One was placed on the premises of the Bavarian State Collection on Zoology (see also *Publication II*), two were operated in the Bavarian Forest National Park, and twelve were operated in the Allgäu Alps in an altitudinal transect (1,036–2,160 m asl) ranging from the Oytal to the Schochen and Nebelhorn Mountains. Using fourteen Malaise traps in total provides limited spatial coverage, especially because in *Publication III*, the goal was to analyze this data to provide species estimates at the national level. All traps were situated in Bavaria, limiting the faunistic coverage of samples to Southern Germany only.

Diversity assessments with COI

In all studies, molecular methods were applied to conduct COI-based diversity assessments. Using BINs as species proxies, extensive insect community analyses were performed despite the fact that less than 25% of all processed specimens (and 21% of all BINs) were linked to a named species on BOLD. In fact, a large proportion of sequences in the datasets were only identifiable to the family- or order-level (Publication I: 93% of specimens; Publication II: 30% of specimens; Publication III: 30% of specimens), underlining the presence of taxonomic gaps within BOLD. Because the majority of the uploaded sequences were assigned to a BIN on BOLD, it was possible to bypass this lack of taxonomic information (see Hebert et al., 2016; Ortiz et al., 2017).

While BINs are commonly used as species proxies in diversity research (Blagoev et al., 2016; Hebert et al., 2010; Ortiz et al., 2017; Ratnasingham & Hebert, 2007; Reid et al., 2011), there are varying opinions regarding the feasibility of doing so (see Meier et al., 2022). The greatest concern regarding the use of BINs is that the RESL-algorithm on BOLD is proprietary, meaning that it is not 100% transparent (Meier et al., 2022). Users on BOLD are unable to see which parameters are used for initial clustering,

merging, and refinement of COI-sequences. Because of this, both Meier et al. (2022) and Cranston et al. (2013) recommend using more than one species delimitation algorithm in order to test the robustness of clusters. Following this recommendation, sequence data was analyzed with three separate clustering algorithms (see *Publication I and V*). With ASAP (Assemble Species by Automatic Partitioning), hierarchical clustering is done using pairwise genetic distances of sequences (Puillandre et al., 2021). The program builds numerous partitions ranked by scores, with the best ones provided in the output to be used for analysis. SpeciesIdentifier version 1.9 (Meier et al., 2006) was used for objective clustering using a preset threshold (3%) for comparative purposes and to increase confidence regarding the relative extent of diversity in the traps. Although cluster counts were overall more conservative with ASAP than with SpeciesIdentifier or RESL, the recovered diversity patterns were overall very similar across methods. Still, it should not be assumed that OTU counts are equal to species numbers in a 1:1 ratio, because there are various molecular factors (such as heteroplasmy, hybridization, recent speciation, NUMTs sequencing, introgression or homogenization of mtDNA haplotypes) that can challenge COI-based delimitation techniques (Buhay, 2009; Dobson, 2007; Duron et al., 2008; Hazkani-Covo et al., 2010; Kmiec et al., 2006; Pamilo et al., 2007). These can lead to algorithms assigning members of a single species into different clusters (thus overestimating diversity) or several species into one (thus underestimating diversity). Despite this, numerous studies have been able to show that there is an overall strong correspondence between BINs- and species numbers (see Hebert et al., 2016; Raupach et al., 2010), and when analyzing more than a million insect specimens, Hebert et al. (2016) recovered patterns of species richness that were overall congruent with expectations.

Integrative taxonomy for hyperabundant and -diverse taxa

Tackling hyperabundant and hyperdiverse taxa from large bulk samples is challenging when applying conventional methods of taxonomy (Meier et al., 2023), so in this thesis, an integrative and complementary workflow was developed in order to study such groups more effectively. This was done using the model taxon Chironomidae. Despite chironomids being crucial organisms in aquatic ecosystems, many biodiversity studies or biomonitoring surveys of aquatic habitats are ignoring them (Dorić et al., 2021; Raunio et al., 2011). This is due to several factors: (i) Chironomids are relatively difficult to identify (Cranston, 2008; Proulx et al., 2013), (ii) only few taxonomists with the required expertise are available for species-level identification (Chan et al., 2014; Cranston et al., 2013), (iii) traditional morphology-based species delimitations often require laborious dissection and mounting of specimens on microscope slides (Ekrem et al., 2010; Gadawski et al., 2022), and (iv) they can be extremely species-rich even in relative low-diversity, temperate, and boreal ecosystems (Lundström et al., 2010). In an ecosystem, their abundance and diversity can be higher than that of all other macroinvertebrates combined (Gratton & Zanden, 2009; Karima, 2021; Marziali et al., 2010). The workload associated with the processing of these organisms is therefore immense when applying conventional identification methods (Brodin et al., 2012; Rosenberg, 1992), so it is simply easier to exclude them from analysis.

The results in *Publication IV* suggest that the established multi-level sorting system was successful: A coverage of over 90 % in species- and cluster counts was obtained, and Chao1 statistics indicate that substantially more species would not have been captured by increasing the sampling effort. This is interesting, because after sorting chironomids into morphotype groups, ultimately only 7% of all specimens were processed and identified. This suggests that the method of grouping specimens into morphotypes, then selecting vouchers for subsequent analysis can be easily delegated to

parataxonomists. Overall, in-depth knowledge of chironomids morphology is not essential for this stage of sample processing, because sorting is based on phenotypic traits such as size, coloration, venation, setation, and shapes of antennae which simply require having a good "eye" and patience (Ekrem et al., 2010; Krell, 2004). This approach was also applied by Ekrem et al. (2010) to subsample non-biting midges for analysis in their study. The identifications of voucher specimens recovered up to seven taxonomic entities per single morphotype, indicating that when in doubt, it is simply easier to merge more specimens into one larger morphotype and compensate by increasing the number of selected vouchers.

Applying traditional morphology to identify vouchers revealed that no-species level information could be provided for almost 20% of them, which is not surprising considering that chironomids are difficult to identify. Morphological identifications require extensive knowledge which can generally only be provided by an expert, but still, the availability of an expert taxonomist does not automatically guarantee error-free species identifications, as demonstrated in this and other studies (see Failla et al., 2016). Not only did the analyses reveal a 9% error rate among morphological identifications, six of the "single species morphotypes" that were said to be distinguishable enough under the stereo microscope for direct species assignment were incorrectly identified. For another 9% of specimens, only identifications to the family or to the genus-level could be provided. False identifications were almost always within a given genus, hence, between closely related species whose morphological differences are often very subtle and therefore require dissection and meticulous analysis (Ekrem et al., 2010). For diverse morphotypes, the number of taxonomic entities recovered using morphology was often over- or underestimated. This reflects the fact that on one hand, these taxa can display high levels of intraspecific morphological variation (Carew et al., 2007, 2011), and on the other hand, closely related species exhibit strong similarities, leading to the erroneous synonymization of species (Anderson et al., 2013). Despite having drastically reduced the taxonomist's workload by analyzing

only a small portion of all collected individuals, they still spent about 500 active working hours processing, mounting, and identifying specimens, which was prone to errors over time (person. comment V. Baranov). Ultimately, these results clearly demonstrate the difficulty in meeting the requirements of ecological studies when using morphology alone, especially when working with cryptic species.

The DNA-barcoding workflow that was applied in parallel was much faster. In total, about 63 working hours were invested in this pipeline from tissue lysis and DNA extraction, to the recovery and editing of sequences, to the upload to BOLD. Identifications were straightforward using BOLD, and examining BIN assignments revealed possible cases of unresolved cryptic diversity: Seven species were involved in multiples BIN assignments and four BINs were linked to interim species names. According to recent research, the genera that these species belong to, namely *Cladopelma*, *Polypedium*, *Pscetrocladius*, and *Smittia* overall display very high intraspecific variations in the COI barcode region across species, making it very difficult to identify a barcode gap for species discrimination (Cranston et al., 2012; Pillot, 2008; Tang et al., 2022). These genera also include species complexes whose taxonomic position is yet unresolved, so many traditionally described species are suspected to comprise more than one cryptic species that are awaiting formal description (Carew et al., 2005; Lehmann, 1970; Saether, 1971). As mentioned, interim species names were linked to four BINs in the dataset, which happens when a genetic difference is detected and/or has been confirmed within one traditional species, yet no species name can be provided at the given time. This can be an indication for the need of a taxonomic revision or a formal species description (Ekrem et al., 2019; Morinière et al., 2019). In other words: Interim species names provide species with an “intermediate name” until they obtain a formal species name (see Geiger et al., 2016; Morinière et al., 2016).

Despite the DNA barcoding pipeline accelerating identifications, researchers must take the time to examine their accuracy. Because even if a high quality reference library is used as a backbone to

analysis (Chimeno et al., 2019; Ekrem, 2007), it does not mean that is is 100% error-free. Several studies have examined the accuracy and reliability of voucher sequences (see Bridge et al., 2003; Meiklejohn et al., 2019; Vilgalys, 2003), concluding that one of the greatest sources for errors is the misidentification or mislabeling of voucher specimens. The problem is caused by the fact that public databases such as BOLD or GenBank are inclusive, enabling any user to upload voucher sequences (Vilgalys, 2003). Without curation (which is demanding against the backdrop that uploads are ongoing and increasing), misidentifications can go unnoticed and quickly inflate the database if sequences with unknown identifications are matched to these. Although Vilgalys (2003) specified inclusivity as being the driving factor behind the upload of "junk data", inclusivity in itself is a regulator because the more data is being uploaded, the more obvious misidentifications become, even for non-experts. With this in mind, every molecular-based identification in *Publication IV* was double-checked for accuracy and not blindly trusted. For this, a neighbor-joining tree was created of publicly available sequences of vouchers that have been morphologically identified by a taxonomic expert prior to being uploaded to BOLD. Sequence records that were either identified using the "BIN taxonomy match" tool on BOLD, or that did not provide any information on the method of voucher identification whatsoever were excluded from analysis. As addressed by Brodin et al. (2012), reference databases need to be expanded as best as possible in order to provide a better taxonomic coverage of species and their intraspecific variation. Quantity, however, should not come at a cost of quality. Because it is not feasible for taxonomic experts to review large batches of records on BOLD, it should become mandatory that all records are provided with an identifier and identification method upon their upload to BOLD so that less-experienced researchers can rely on the data even when no taxonomic expert is available. Ultimately, it is in our own interest as scientists that reference libraries be as accurate as possible, so in order to achieve this, every user needs to take responsibility and check the integrity of their data before making it public.

Applying two identification pipelines in parallel on the same vouchers revealed discrepant results for a third of cases. After meticulous review of the data, the cause for these discrepancies were traced back to (i) misidentifications or the lack of species-level assignments by the taxonomic expert via morphology, (ii) misidentifications of voucher sequences on BOLD, (iii) the lack of public data on BOLD (no match), or (iv) multiple BIN-assignments or BIN-sharing cases on BOLD. By comparing the outputs of each methodology, it became obvious that each method has its own limitations. Meticulous revision revealed that 9% of vouchers were morphologically misidentified by the taxonomic expert. For another 9% of vouchers, morphology could not provide identifications at the species-level, meaning that for 18% of vouchers, morphology did not provide accurate or comprehensive species-level information. On the other hand, morphology did provide more comprehensive species information for a total of 40 specimens (14%). Here, species-level IDs were provided for five BINs that were not yet on BOLD, and for six BINs that were linked to discrepant identifications by expert taxonomists.

These results demonstrate that while each method has its own drawbacks, using them in combination increases rigor. DNA barcoding does not require difficult-to-acquire taxonomic knowledge and drastically fast-forwards the process of identification of difficult-to-study organisms. At the same time, barcode registries are only as good as the quality of their vouchers, which have been identified using morphology. When tackling megadiverse dark taxa, no method should be applied on its own, as each has its own repertoire of biases and limitations (Engel et al., 2021; Hartop et al., 2022; Krell, 2004).

Yet every method is subjective

No matter which method - or better yet - methods one chooses to sample, delimit, or describe species, it (or they) will always provide an arbitrary result. Scientists create models and frameworks to simplify and describe complex systems in nature, yet as good as these can be, there will always be a level of

incongruence between the resulting output and reality itself. Therefore, as underlined in this thesis, the best approach at tackling something that is unknown is to apply different methodologies from different disciplines to obtain independent outputs that together, provide a more comprehensive picture.

Morphology has been viewed as the golden standard in taxonomy for a long time, yet too few discuss the fact that morphology is a subjective method in itself. Not only do accurate morphological identifications highly rely on the availability and accuracy of determination keys, they also rely on the identifier's ability to conduct these identifications from an objective perspective (Ekrem et al., 2019). Snodgrass (1935) stated that *"anatomy is what you can see with your eyes, morphology is what you think you see with your mind. In other words, morphology is based on a larger frame of reference and one's subjective judgment. Or, put very simply, morphology is the sum total of one's imagination or perceived notion of a particular structure, organ, or organ system."* As expressed by Ayodhya P. Gupta (1994), *"…morphology is a subjective product of the morphologist's mind"* adding that *"in the decades ahead, arthropod anatomy-morphology will be pursued vigorously at the molecular level because, to be on the cutting edge, it will have to remain in the mainstream of modern biology. This view should in no way be construed to mean that studies of classical anatomy-morphology at the gross, macro, micro, and ultrastructural and subcellular levels will not be needed or become irrelevant; on the contrary, such studies will serve as prerequisites to those at the molecular level."* Just like morphology, molecular identification methods themselves suffer from a wide range of biases (Tedersoo et al., 2019). Sources of bias range from molecular factors affecting the genetic variability of the specimens in question (i.e., numts, recent speciation, prevalence of paraphyly, high degree of infection by endosymbiotic bacteria), to laboratory protocols used (i.e., nature of sample, extraction protocols, primers, sequencing), to bioinformatic pipeline (clustering algorithms) and the reliability of identifications in reference databases (Virgilio et al., 2010). Essentially, scientists are always working with alternative

representations of reality and conversely, every framework that scientists work with is nothing but a subset. A sampling site, for example, is nothing but a subset of a region and/or habitat, and sampling methods will always only collect a proportion of a site's true diversity. Depending on trap type, a bias favoring the sampling of some taxa over others is always present, meaning that the community captured will never depict the true insect community of a sampled site (Karlsson et al., 2020). Furthermore, the setup of a trap in terms of site choice, orientation, and above-ground-level is another source of bias, and these factors strongly influence sampling results (Chan-Canché et al., 2020). Then, once samples are collected, only a fraction of these are processed, from which only a few vouchers are selected for analysis. Each step of this process narrows our perspective of reality further until ultimately, only the tiniest aspect of a very complex world is examined.

Against this backdrop, different methodologies will always provide discrepant results. The mere purpose of models is to break down the complexity of natural systems so that scientists can study specific aspects in isolation and understand driving mechanisms (Parker, 2010). The greater the simplification of a model is, the greater is the uncertainty of its output (Loucks et al., 2017). Yet discordant results do not necessarily challenge the rigor of another method (Stegenga, 2009). Instead, these provide more in-depth knowledge and often, we simply do not have the ability to interpret the results as they should. As expressed by Stegenga (2009), *"Hypotheses are better supported with evidence generated by multiple techniques that rely on different background assumptions."* The least we can do is apply as many different models to the same framework to minimize levels of uncertainty, thus obtaining a more resilient representation of what we think is reality.

Trying to bypass specimen destruction in DNA metabarcoding

DNA metabarcoding has become a well-established method for large-scale diversity assessments (Shum & Palumbi, 2021). Still, a consensus workflow is lacking in some fields of research (Elbrecht &

Leese, 2015), with one subject of debate being the nature of sample used for analysis. Traditionally, sample contents are homogenized into a fine powder from which the DNA is extracted in one extraction step (Yu et al., 2012). While homogenization of tissue releases the most DNA, this comes at the cost of losing the structural integrity of specimens, preventing the recovery of abundance data or a-posteriori morphological analysis and/or verification of specimens' identity (Aylagas et al., 2016, 2018). Therefore, in *Publication V*, the collection medium of a sample (EtOH), which is otherwise discarded, was tested for its usability as a DNA source for analysis. If so, specimens can be left intact, providing scientists with a wealth of ecological information that can be analyzed at any given time. In contrast to other studies, the aim was not to compare differences in OTUs recovered with tissue-based and with ethanol-based DNA because it has already been shown (and it is expected) that these recover different subsets. Instead, we examined whether ecological gradients were conserved across methods. In ecology, where researchers always work with subsets of communities, identical taxonomic recovery may not always be as crucial as the conservation of ecological and environmental information.

As expected, very different arthropod communities were obtained with each method (see also (Elbrecht et al., 2017; Kirse et al., 2022; Marquina et al., 2019), and the reasons for this are discussed in *Publication V*. Regarding the depiction of ecological gradients, information was only partly conserved in the collection medium of samples. Testing for significant differences in community compositions based on three factors (trap site, habitat type, and seasonality) recovered that habitats and sites had no effect on community compositions, but seasonality did. Statistical analysis of the individual insect orders (Diptera, Hymenoptera, Lepidoptera, Coleoptera, and Hemiptera), almost always indicated significant differences that were driven by seasonality (adonis2 $p = 0.001$). These seasonal gradients were strongest among Hymenoptera and Coleoptera. It is not clear as to why seasonal trends in ethanol are better conserved among some groups and lesser so among others. However, it is speculated that a group's trophic level may have a meaningful impact, as arthropod specimens that

fall prey to other arthropods are introduced into the ethanol as gut content (Marquina et al., 2019). Differing temporal-based factors (e.g. predator–prey interactions, predator metabolic rates, time elapsed since prey consumption) would skew natural patterns of abundances because gut-based DNA of the same species is introduced into the ethanol at random points of time. In addition, there are numerous methodological, environmental and biological/physiological constraints that have a direct influence on the success rates of gut content sequencing (see Eitzinger et al., 2013; Greenstone et al., 2010; von Berg et al., 2008). With too many sources of bias that are introduced into the analysis of the ethanol-based DNA, and no possibility of discriminating between ingested and captured arthropods, seasonal patterns are especially prone to distortion among groups that include many prey species. This is reinforced by the fact that seasonal gradients were best depicted among Hymenoptera, Coleoptera, and Hemiptera, which encompass species that are less susceptible to falling prey to other arthropods, so they are also less likely to be introduced into the ethanol of samples as gut content. Typical predators of Coleoptera, Hymenoptera, and Hemiptera are, for example, birds, bats, and frogs (*Britannica*, 2022). Other arthropods that predate these taxa include Odonata and Araneae, both of which are less represented in the dataset. In contrast, predators of Diptera and Lepidoptera (which depicted highly skewed gradients) were very abundant in the trap samples, as these include many taxa of Hymenoptera, Coleoptera, Diptera, and Araneae (Flint et al., 1998).

Consistent with previous findings, alpha-diversity assessments demonstrated that the ethanol-based DNA (1) failed at discriminating between the terrestrial and riparian habitats and (2) underrepresented the magnitude of arthropod diversity within every single habitat (see Erdozain et al., 2019; Linard et al., 2016). Recently, (Zenker et al., 2020) conducted DNA metabarcoding exclusively on the preservative ethanol of automatic light trap samples to compare the alpha and beta diversity of arthropod communities in Brazil. Unfortunately, they did not examine or process the tissue of these samples at all, so no reference was available as a guideline to their interpretations. Observing the alpha-diversity

curves that were obtained, it is strongly believed that the sole use of preservative ethanol can clearly lead to false conclusions, and therefore, the sole use of ethanol-based DNA should be discouraged until further research has been conducted.

To conclude, ethanol-based DNA sequencing did not provide information on ecological gradients, except for the case of seasonal patterns. The conserved seasonality among some taxa is an interesting starting point for further investigations, but until more research has provided successful results, it is recommended that researchers covering terrestrial ecosystems be careful when using ethanol-based DNA. It is important to mention that in this study, 80% ethanol (1 vol% MEK) was used for arthropod sampling. DNA extractions were conducted in spring 2020 following the collection season (April–October 2019). According to (Marquina et al., 2020), this concentration of ethanol is too low for optimal DNA preservation over time. Therefore, it is highly encouraged for others to use 95% ethanol for sampling to guarantee optimal DNA preservation.

CONCLUSION

This thesis addresses the difficulties associated with the research of the small, cryptic diverse, and highly abundant dark taxa. One major constraint is the inability to link these insects to species names or ecological functions. Here, the use of innovative molecular approaches enabled in-depth analysis of dark taxa despite the majority of specimens not being associated to a species name. An integrative workflow (combining morphology and molecular biology) was developed with the goal of accelerating bulk sample processing with a more targeted approach at voucher selection for subsequent analysis. This workflow rendered successfully, with results depicting that more than 90% of the collected diversity was recovered although essentially less than 7% of specimens were actually assessed in detail. At the same time, this thesis depicts the limitations (and thus difficulties) associated with the use of one single method for the analysis of challenging groups.

Technological advances are currently expediting biodiversity research, and the next decade will most certainly be witness to substantial breakthroughs in the field of taxonomy. Recent developments include a sorting robot (the DiversityScanner) which has the ability to recognize and sort insect specimens based on overview images of bulk samples (see Wührl et al., 2022). Especially interesting is the fact that it is able to process very small specimens, including many dark taxa (<2 mm) (Wührl et al., 2022). In another very recent study, a workflow was developed that combines HotSHOT with MinION (Oxford, Nanopore, Oxford, UK) technologies to conduct fast and accurate species-level sorting of ecological samples (see Vasilita et al., 2023). With a modest amount of equipment, manpower, and training, the authors were able to conduct species-level sorting within hours, which came down to 2.5 minutes per specimen (Vasilita et al., 2023). Fresh off the press, Meier et al., (2023) propose a new protocol termed “dark taxonomy” which – as suggested in this thesis – uses integrative taxonomy for overcoming the taxonomic impediments for dark taxa. Commendably, the authors have

chosen to first tackle species from samples that are most relevant in biomonitoring frameworks, which is in stark contrast to the taxonomy biases prevailing in other studies. Of course, accurate identifications are only possible if identified sequences are present in databases, however, coupling these approaches with other workflows, such as the one developed in this thesis (also see Hartop et al., 2022), could drastically expedite the work for taxonomists.

Ultimately, the greatest hurdle in taxonomy is the mindset of those conducting the research and with taxonomists ever more embracing the implementation of non-conventional, future-oriented, and/or AI-based technologies half the battle is already won.

REFERENCES

Abadie, J.-C., Andrade, C., Machon, N., & Porcher, E. (2008). On the use of parataxonomy in biodiversity monitoring: A case study on wild flora. *Biodiversity and Conservation*, 17(14), 3485–3500. https://doi.org/10.1007/s10531-008-9354-z

Anderson, A. M., Stur, E., & Ekrem, T. (2013). Molecular and morphological methods reveal cryptic diversity and three new species of Nearctic Micropsectra (Diptera:Chironomidae). *Freshwater Science*, 32(3), 892–921. https://doi.org/10.1899/12-026.1

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, 27(20), 3968–3975. https://doi.org/10.1111/mec.14844

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Frontiers in Marine Science*, 3, 96.https://doi.org/10.3389/fmars.2016.00096

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

Badirli, S., Picard, C. J., Mohler, G., Richert, F., Akata, Z., & Dundar, M. (2023). Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution*, 14(6), 1515–1530. https://doi.org/10.1111/2041-210X.14104

Baloğlu, B., Clews, E., & Meier, R. (2018). NGS barcoding reveals high resistance of a hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs. *Frontiers in Zoology*, 15(1), 31. https://doi.org/10.1186/s12983-018-0276-7

Basset, Y., Cizek, L., Cuénoud, P., Didham, R. K., Guilhaumon, F., Missa, O., Novotny, V., Ødegaard,

F., Roslin, T., Schmidl, J., Tishechkin, A. K., Winchester, N. N., Roubik, D. W., Aberlenc, H.-P., Bail, J., Barrios, H., Bridle, J. R., Castaño-Meneses, G., Corbara, B., … Leponce, M. (2012). Arthropod Diversity in a Tropical Forest. *Science*, 338(6113), 1481–1484. https://doi.org/10.1126/science.1226727

Basset, Y., Novotny, V., Miller, S. E., & Pyle, R. (2000). Quantifying Biodiversity: Experience with Parataxonomists and Digital Photography in Papua New Guinea and Guyana. *BioScience*, 50(10), 899–908. https://doi.org/10.1641/0006-3568(2000)050[0899:QBEWPA]2.0.CO;2

Beermann, A. J., Zizka, V. M. A., Elbrecht, V., Baranov, V., & Leese, F. (2018). DNA metabarcoding reveals the complex and hidden responses of chironomids to multiple stressors. *Environmental Sciences Europe*, 30(1), 26. https://doi.org/10.1186/s12302-018-0157-x

Beng, K. C., Tomlinson, K. W., Shen, X. H., Surget-Groba, Y., Hughes, A. C., Corlett, R. T., & Slik, J. W. F. (2016). The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. *Scientific Reports*, 6(1), 24965. https://doi.org/10.1038/srep24965

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, 41(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

Bickel, D. (1999). What museum collections reveal about species accumulation, richness, and rarity: An example from the Diptera. In *The Other 99%: The Conservation and Biodiversity of Invertebrates* (pp. 174–181). Royal Zoological Society of New South Wales.

Blagoev, G. A., Dewaard, J. R., Ratnasingham, S., Dewaard, S. L., Lu, L., Robertson, J., Telfer, A. C., & Hebert, P. D. (2016). Untangling taxonomy: A DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources*, 16(1), 325–341. DOI: 10.1111/1755-0998.12444

Boero, F. (1996). Specie, biodiversità e comunità. *Biologia Marina Mediterranea*, 3, 13–20.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., & Woodward, G. (2017). Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of

Ecological Networks. *Trends in Ecology & Evolution*, 32(7), 477–487. https://doi.org/10.1016/j.tree.2017.03.001

Boventer, K. (1960). Johann Wilhelm Meigen 1764-1845: Entomologe und Botaniker, Begründer der Dipterologie. *Sudhoffs Archiv Für Geschichte Der Medizin Und Der Naturwissenschaften*, *44*(1), 45–53. https://www.jstor.org/stable/20774623

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, 16(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, 21(6), 1904–1921. https://doi.org/10.1111/1755-0998.13398

Bridge, P. D., Roberts, P. J., Spooner, B. M., & Panchal, G. (2003). On the unreliability of published DNA sequences. *New Phytologist*, 160(1), 43–48. https://doi.org/10.1046/j.1469-8137.2003.00861.x

Britannica. (2022, March 17). Britannica. https.//www.britannica.com

Brodin, Y., Ejdung, G., Strandberg, J., & Lyrholm, T. (2012). Improving environmental and biodiversity monitoring in the Baltic Sea using DNA barcoding of Chironomidae (Diptera). *Molecular Ecology Resources*, 996–1004. https://doi.org/10.1111/1755-0998.12053

Brown, B. V. (2021). Sampling Methods for Adult Flies (Diptera). In J. C. Santos & G. W. Fernandes (Eds.), *Measuring Arthropod Biodiversity: A Handbook of Sampling Methods* (pp. 187–204). Springer International Publishing. https://doi.org/10.1007/978-3-030-53226-0_7

Brown, B. V., Borkent, A., Adler, P. H., Amorim, D. de S., Barber, K., Bickel, D., Boucher, S., Brooks, S. E., Burger, J., Burington, Z. L., Capellari, R. S., Costa, D. N. R., Cumming, J. M., Curler, G., Dick, C. W., Epler, J. H., Fisher, E., Gaimari, S. D., Gelhaus, J., … Zumbado, M. A. (2018).

Comprehensive inventory of true flies (Diptera) at a tropical site. *Communications Biology*, 1(1), 21. https://doi.org/10.1038/s42003-018-0022-x

Buchner, D., Macher, T.-H., & Leese, F. (2022). APSCALE: Advanced pipeline for simple yet comprehensive analyses of DNA metabarcoding data. *Bioinformatics*, 38(20), 4817–4819. https://doi.org/10.1093/bioinformatics/btac588

Buhay, J. E. (2009). "COI-like" Sequences Are Becoming Problematic in Molecular Systematic and DNA Barcoding Studies. *Journal of Crustacean Biology*, 29(1), 96–110. https://doi.org/10.1651/08-3020.1

Bukowski, B., Ratnasingham, S., Hanisch, P. E., Hebert, P. D. N., Perez, K., deWaard, J., Tubaro, P. L., & Lijtmaer, D. A. (2022). DNA barcodes reveal striking arthropod diversity and unveil seasonal patterns of variation in the southern Atlantic Forest. *PLOS ONE*, 17(4), e0267390. https://doi.org/10.1371/journal.pone.0267390

Bush, A., Compson, Z. G., Monk, W. A., Porter, T. M., Steeves, R., Emilson, E., Gagne, N., Hajibabaei, M., Roy, M., & Baird, D. J. (2019). Studying Ecosystems With DNA Metabarcoding: Lessons From Biomonitoring of Aquatic Macroinvertebrates. *Frontiers in Ecology and Evolution*, 7, 434. https://www.frontiersin.org/articles/10.3389/fevo.2019.00434

Campbell, J. W., & Hanula, J. L. (2007). Efficiency of Malaise traps and colored pan traps for collecting flower visiting insects from three forested ecosystems. *Journal of Insect Conservation*, 11(4), 399–408. https://doi.org/10.1007/s10841-006-9055-4

Čandek, K., & Kuntner, M. (2015). DNA barcoding gap: Reliable species identification over morphological and geographical scales. *Molecular Ecology Resources*, 15(2), 268–277. https://doi.org/10.1111/1755-0998.12304

Cardoso, P., Barton, P. S., Birkhofer, K., Chichorro, F., Deacon, C., Fartmann, T., Fukushima, C. S., Gaigher, R., Habel, J. C., Hallmann, C. A., Hill, M. J., Hochkirch, A., Kwak, M. L., Mammola, S., Ari Noriega, J., Orfinger, A. B., Pedraza, F., Pryke, J. S., Roque, F. O., … Samways, M. J. (2020).

Scientists' warning to humanity on insect extinctions. *Biological Conservation*, 242, 108426. https://doi.org/10.1016/j.biocon.2020.108426

Cardoso, P., Erwin, T. L., Borges, P. A. V., & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, 144(11), 2647–2655. https://doi.org/10.1016/j.biocon.2011.07.024

Cardoso, P., & Leather, S. R. (2019). Predicting a global insect apocalypse. *Insect Conservation and Diversity*, 12(4), 263–267. https://doi.org/10.1111/icad.12367

Carew, M. E., Marshall, S. E., & Hoffmann, A. A. (2011). A combination of molecular and morphological approaches resolves species in the taxonomically difficult genus Procladius Skuse (Diptera: Chironomidae) despite high intra-specific morphological variation. *Bulletin of Entomological Research*, 101(5), 505–519. https://doi.org/10.1017/S000748531100006X

Carew, M. E., Pettigrove, V., Cox, R. L., & Hoffmann, A. A. (2007). DNA identification of urban Tanytarsini chironomids (Diptera:Chironomidae). *Journal of the North American Benthological Society*, 26(4), 587–600. https://doi.org/10.1899/06-120.1

Carew, M. E., Pettigrove, V., & Hoffmann, A. A. (2005). The Utility of DNA Markers in Classical Taxonomy: Using Cytochrome Oxidase I Markers to Differentiate Australian Cladopelma (Diptera: Chironomidae) Midges. *Annals of the Entomological Society of America*, 98(4), 587–594. https://doi.org/10.1603/0013-8746(2005)098[0587:TUODMI]2.0.CO;2

Carpenter, M. M. (1945). Bibliography of Biographies of Entomologists. *The American Midland Naturalist*, *33*(1), 1–116.

Castellani, C. A., Longchamps, R. J., Sun, J., Guallar, E., & Arking, D. E. (2020). Thinking outside the nucleus: Mitochondrial DNA copy number in health and disease. *Mitochondrion*, 53, 214–223. https://doi.org/10.1016/j.mito.2020.06.004

Chan, A., Chiang, L.-P., Hapuarachchi, H. C., Tan, C.-H., Pang, S.-C., Lee, R., Lee, K.-S., Ng, L.-C., & Lam-Phua, S.-G. (2014). DNA barcoding: Complementing morphological identification of

mosquito species in Singapore. *Parasites & Vectors*, 7, 569. https://doi.org/10.1186/s13071-014-0569-4

Chan-Canché, R., Ballina-Gómez, H., Leirana-Alcocer, J., Bordera, S., & González-Moreno, A. (2020). Sampling of parasitoid Hymenoptera: Influence of the height on the ground. *Journal of Hymenoptera Research*, 78, 19–31. https://doi.org/10.3897/jhr.78.54309

Chandra, A., & Idrisova, A. (2011). Convention on Biological Diversity: A review of national challenges and opportunities for implementation. *Biodiversity and Conservation*, 20(14), 3295–3316. https://doi.org/10.1007/s10531-011-0141-x

Chimeno, C., Morinière, J., Podhorna, J., Hardulak, L., Hausmann, A., Reckel, F., Grunwald, J. E., Penning, R., & Haszprunar, G. (2019). DNA Barcoding in Forensic Entomology – Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Specie,. *Journal of Forensic Sciences*, 64(2), 593–601. https://doi.org/10.1111/1556-4029.13869

Clark, J. A., & May, R. M. (2002). Taxonomic Bias in Conservation Research. *Science*, 297(5579), 191–192. https://doi.org/10.1126/science.297.5579.191b

Clayton, D. A. (1992). Transcription and Replication of Animal Mitochondrial DNAs. In D. R. Wolstenholme & K. W. Jeon (Eds.), *International Review of Cytology* (Vol. 141, pp. 217–232). Academic Press. https://doi.org/10.1016/S0074-7696(08)62067-7

Coddington, J. A., Agnarsson, I., Miller, J. A., Kuntner, M., & Hormiga, G. (2009). Undersampling Bias: The Null Hypothesis for Singleton Species in Tropical Arthropod Surveys. *Journal of Animal Ecology*, 78(3), 573–584. http://www.jstor.org/stable/27696402

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. *Molecular Ecology*, 25(7), 1423–1428. https://doi.org/10.1111/mec.13549

Costion, C., Ford, A., Cross, H., Crayn, D., Harrington, M., & Lowe, A. (2011). Plant DNA Barcodes Can Accurately Estimate Species Richness in Poorly Known Floras. *PLOS ONE*, 6(11), e26841.

https://doi.org/10.1371/journal.pone.0026841

Cranston, P. (2008). Keys to The Adult Male Chironomidae of Britain and Ireland—Edited by P.H. Langton and L.C.V. Pinder and A Guide to The Identification of Genera of Chironomid Pupal Exuviae Occurring in Britain and Ireland (Including Common Genera From Northern Europe) and Their Use In Monitoring Lotic and Lentic Fresh Waters—Edited by R.S. Wilson and L.P. Ruse. *Systematic Entomology*, 33(1), 215–217. https://doi.org/10.1111/j.1365-3113.2007.00412_5.x

Cranston, P. S., Ang, Y. C., Heyzer, A., Lim, R. B. H., Wong, W. H., Woodford, J. M., & Meier, R. (2013). *The nuisance midges (Diptera: Chironomidae) of Singapore's Pandan and Bedok reservoirs*. 61(2), 779–793.

Cranston, P. S., Hardy, N. B., & Morse, G. E. (2012). A dated molecular phylogeny for the Chironomidae (Diptera). *Systematic Entomology*, 37(1), 172–188. https://doi.org/10.1111/j.1365-3113.2011.00603.x

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29(10), 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Dawnay, N., Ogden, R., McEwing, R., Carvalho, G. R., & Thorpe, R. S. (2007). Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International*, 173(1), 1–6. https://doi.org/10.1016/j.forsciint.2006.09.013

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, *85*(3), 407–415. https://doi.org/10.1111/j.1095-8312.2005.00503.x

de Carvalho, M. R. (2005). Revisiting the Taxonomic Impediment. *Science*, 307(5708), 353b–353b. https://doi.org/10.1126/science.307.5708.353b

de Carvalho, M. R., Bockmann, F. A., Amorim, D. S., Brandão, C. R. F., de Vivo, M., de Figueiredo, J. L., Britski, H. A., de Pinna, M. C. C., Menezes, N. A., Marques, F. P. L., Papavero, N., Cancello,

E. M., Crisci, J. V., McEachran, J. D., Schelly, R. C., Lundberg, J. G., Gill, A. C., Britz, R., Wheeler, Q. D., … Nelson, G. J. (2007). Taxonomic Impediment or Impediment to Taxonomy? A Commentary on Systematics and the Cybertaxonomic-Automation Paradigm. *Evolutionary Biology*, 34(3), 140–143. https://doi.org/10.1007/s11692-007-9011-6

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., Kartzinel, T. R., & Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, 28(2), 391–406. https://doi.org/10.1111/mec.14734

DeSalle, R., & Goldstein, P. (2019). Review and Interpretation of Trends in DNA Barcoding. *Frontiers in Ecology and Evolution*, 7, 302. https://www.frontiersin.org/articles/10.3389/fevo.2019.00302

Dewaard, J., Levesque-Beaudin, V., deWaard, S., Ivanova, N., Mckeown, J., Miskie, R., Naik, S., Perez, K., Ratnasingham, S., Sobel, C., Sones, J., Steinke, C., Telfer, A., Young, A., Young, M., Zakharov, E., & Hebert, P. (2018). Expedited Assessment of Terrestrial Arthropod Diversity by Coupling Malaise Traps with DNA Barcoding. *Genome*, 62(3), 85–96. https://doi.org/10.1139/gen-2018-0093

Dobson, S. L. (2007). Evolution of Wolbachia cytoplasmic incompatibility. *Evolution*, *58*(10), 2156–2166.

Donaldson, M. R., Burnett, N. J., Braun, D. C., Suski, C. D., Hinch, S. G., Cooke, S. J., & Kerr, J. T. (2016). Taxonomic bias and international biodiversity conservation research. In *Facets* (Vol. 1, Issue 1, pp. 105–113). Canadian Science Publishing 65 Auriga Drive, Suite 203, Ottawa, ON K2E 7W6.

Dorić, V., Pozojević, I., Vučković, N., Ivković, M., & Mihaljević, Z. (2021). Lentic chironomid performance in species-based bioassessment proving: High-level taxonomy is not a dead end in monitoring. *Ecological Indicators*, 121, 107041. https://doi.org/10.1016/j.ecolind.2020.107041

Duffus, N. E., Christie, C. R., & Morimoto, J. (2021). Insect cultural services: How insects have

changed our lives and how can we do better for them. *Insects*, 12(5), 377. https://doi.org/10.3390/insects12050377

Dunn, C. P. (2003). Keeping taxonomy based in morphology. *Trends in Ecology & Evolution*, 18(6), 270–271. https://doi.org/10.1016/S0169-5347(03)00094-6

Dupuis, C. (1984). Willi Hennig's Impact on Taxonomic Thought. *Annual Review of Ecology and Systematics*, *15*(1), 1–25. https://doi.org/10.1146/annurev.es.15.110184.000245

Duron, O., Bouchon, D., Boutin, S., Bellamy, L., Zhou, L., Engelstädter, J., & Hurst, G. D. (2008). The diversity of reproductive parasites among arthropods: Wolbachiado not walk alone. *BMC Biology*, 6(1), 27. https://doi.org/10.1186/1741-7007-6-27

Eisenhauer, N., Bonn, A., & A. Guerra, C. (2019). Recognizing the quiet extinction of invertebrates. *Nature Communications*, 10(1), 50. https://doi.org/10.1038/s41467-018-07916-1

Eitzinger, B., Micic, A., Körner, M., Traugott, M., & Scheu, S. (2013). Unveiling soil food web links: New PCR assays for detection of prey DNA in the gut of soil arthropod predators. *Soil Biology and Biochemistry*, 57, 943–945. https://doi.org/10.1016/j.soilbio.2012.09.001

Ekrem, T. (2007). A taxonomic revision of the genus Stempellinella (Diptera: Chironomidae). *Journal of Natural History*, 41(21–24), 1367–1465. https://doi.org/10.1080/00222930701437360

Ekrem, T., Stur, E., & Hebert, P. D. N. (2010). Females do count: Documenting Chironomidae (Diptera) species diversity using DNA barcoding. *Organisms Diversity & Evolution*, 10(5), 397–408. https://doi.org/10.1007/s13127-010-0034-y

Ekrem, T., Stur, E., Orton, M. G., & Adamowicz, S. J. (2019). DNA barcode data reveal biogeographic trends in Arctic non-biting midges. *Trends in DNA Barcoding and Metabarcoding*, 01(01), 787–796. https://doi.org/10.1139/gen-2018-0100@gen-dna.issue01

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLOS ONE*, 10(7), e0130324.

https://doi.org/10.1371/journal.pone.0130324

Elbrecht, V., Peinert, B., & Leese, F. (2017). *Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding*. *Ecology and Evolution*, 7(17), 6918–6926. https://doi.org/10.7287/peerj.preprints.2561

Elbrecht, V., & Steinke, D. (2019). Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. *Freshwater Biology*, 64(2), 380–387. https://doi.org/10.1111/fwb.13220

Engel, M. S., Ceríaco, L. M. P., Daniel, G. M., Dellapé, P. M., Löbl, I., Marinov, M., Reis, R. E., Young, M. T., Dubois, A., Agarwal, I., Lehmann A., P., Alvarado, M., Alvarez, N., Andreone, F., Araujo-Vieira, K., Ascher, J. S., Baêta, D., Baldo, D., Bandeira, S. A., … Zacharie, C. K. (2021). The taxonomic impediment: A shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society*, 193(2), 381–387. https://doi.org/10.1093/zoolinnean/zlab072

Erdozain, M., Thompson, D. G., Porter, T. M., Kidd, K. A., Kreutzweiser, D. P., Sibley, P. K., Swystun, T., Chartrand, D., & Hajibabaei, M. (2019). Metabarcoding of storage ethanol vs. Conventional morphometric identification in relation to the use of stream macroinvertebrates as ecological indicators in forest management. *Ecological Indicators*, 101, 173–184. https://doi.org/10.1016/j.ecolind.2019.01.014

Erwin, T. L. (1983). Tropical Forest Canopies: The Last Biotic Frontier. *Bulletin of the Entomological Society of America*, 29(1), 14–20. https://doi.org/10.1093/besa/29.1.14

Erwin, T. L. (1991). How Many Species Are There?: Revisited. *Conservation Biology*, 5(3), 330–333.

Erwin, T. L. (2004). The biodiversity question: How many species of terrestrial arthropods are there. In M. D. Lowman & H. B. Rinker (Eds.), *Forest Canopies* (pp. 259–269). Elsevier. https://doi.org/10.1016/B978-012457553-0/50019-8

Essl, F., Latombe, G., Lenzner, B., Pagad, S., Seebens, H., Smith, K., Wilson, J. R., & Genovesi, P. (2020). The Convention on Biological Diversity (CBD)'s Post-2020 target on invasive alien

species–what should it include and how should it be monitored? *NeoBiota*, 62, 99–121. https://doi.org/10.3897/neobiota.62.53972

Failla, A. J., Vasquez, A. A., Hudson, P., Fujimoto, M., & Ram, J. L. (2016). Morphological identification and COI barcodes of adult flies help determine species identities of chironomid larvae (Diptera, Chironomidae). *Bulletin of Entomological Research*, 106(1), 34–46. https://doi.org/10.1017/S0007485315000486

Fazey, I., Fischer, J., & Lindenmayer, D. B. (2005). What do conservation biologists publish? *Biological Conservation*, 124(1), 63–73. https://doi.org/10.1016/j.biocon.2005.01.013

Flint, M. L., Dreistadt, S. H., & Program, U. of C. I. P. M. (1998). *Natural Enemies Handbook: The Illustrated Guide to Biological Pest Control*. University of California Press.

Frézal, L., & Leblois, R. (2008). Four years of DNA barcoding: Current advances and prospects. *Infection, Genetics and Evolution*, 8(5), 727–736. https://doi.org/10.1016/j.meegid.2008.05.005

Fukano, Y., & Soga, M. (2021). Why do so many modern people hate insects? The urbanization–disgust hypothesis. *Science of The Total Environment*, 777, 146229. https://doi.org/10.1016/j.scitotenv.2021.146229

Gadawski, P., Montagna, M., Rossaro, B., Giłka, W., Pešić, V., Grabowski, M., & Magoga, G. (2022). DNA barcoding of Chironomidae from the Lake Skadar region: Reference library and a comparative analysis of the European fauna. *Diversity and Distributions*, 1–20. https://doi.org/10.1111/ddi.13504

Gallai, N., Salles, J.-M., Settele, J., & Vaissière, B. E. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics*, 68(3), 810–821. https://doi.org/10.1016/j.ecolecon.2008.06.014

Garibaldi, L. A., Carvalheiro, L. G., Leonhardt, S. D., Aizen, M. A., Blaauw, B. R., Isaacs, R., Kuhlmann, M., Kleijn, D., Klein, A. M., Kremen, C., Morandin, L., Scheper, J., & Winfree, R. (2014). From research to action: Enhancing crop yield through wild pollinators. *Frontiers in Ecology and*

the Environment, 12(8), 439–447. https://doi.org/10.1890/130330

Gaston, K. J. (1991). The Magnitude of Global Insect Species Richness. *Conservation Biology*, 5(3), 283–296. https://doi.org/10.1111/j.1523-1739.1991.tb00140.x

GBOL III: Dark Taxa – Dark Taxa. (2023). https://gbol.bolgermany.de/gbol3/

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

Giangrande, A. (2003). Biodiversity, conservation, and the 'Taxonomic impediment.' *Aquatic Conservation: Marine and Freshwater Ecosystems*, 13(5), 451–459. https://doi.org/10.1002/aqc.584

Gibbs, J. (2018). DNA barcoding a nightmare taxon: Assessing barcode index numbers and barcode gaps for sweat bees. *Genome*, 61(1), 21–31. https://doi.org/10.1139/gen-2017-0096

Gill, R. J., Baldock, K. C. R., Brown, M. J. F., Cresswell, J. E., Dicks, L. V., Fountain, M. T., Garratt, M. P. D., Gough, L. A., Heard, M. S., Holland, J. M., Ollerton, J., Stone, G. N., Tang, C. Q., Vanbergen, A. J., Vogler, A. P., Woodward, G., Arce, A. N., Boatman, N. D., Brand-Hardy, R., … Potts, S. G. (2016). Protecting an Ecosystem Service. In *Advances in Ecological Research* (Vol. 54, pp. 135–206). Elsevier. https://doi.org/10.1016/bs.aecr.2015.10.007

González-Oreja, J. A. (2008). The encyclopedia of life vs. the brochure of life: Exploring the relationships between the extinction of species and the inventory of life on Earth. *Zootaxa*, *1965*(1), 61–68.

Gorini, T., Mezzasalma, V., Deligia, M., De Mattia, F., Campone, L., Labra, M., & Frigerio, J. (2023). Check Your Shopping Cart: DNA Barcoding and Mini-Barcoding for Food Authentication. *Foods*, 12(12), 2392. https://doi.org/10.3390/foods12122392

Grant, D. M., Brodnicke, O. B., Evankow, A. M., Ferreira, A. O., Fontes, J. T., Hansen, A. K., Jensen, M.

R., Kalaycı, T. E., Leeper, A., Patil, S. K., Prati, S., Reunamo, A., Roberts, A. J., Shigdel, R., Tyukosova, V., Bendiksby, M., Blaalid, R., Costa, F. O., Hollingsworth, P. M., … Ekrem, T. (2021). The Future of DNA Barcoding: Reflections from Early Career Researchers. *Diversity*, 13(7), 313. https://doi.org/10.3390/d13070313

Gratton, C., & Zanden, M. J. V. (2009). Flux of aquatic insect productivity to land: Comparison of lentic and lotic ecosystems. *Ecology*, 90(10), 2689–2699. https://doi.org/10.1890/08-1546.1

Greenstone, M. H., Szendrei, Z., Payton, M. E., Rowley, D. L., Coudron, T. C., & Weber, D. C. (2010). Choosing natural enemies for conservation biological control: Use of the prey detectability half-life to rank key predators of Colorado potato beetle: Ranking predators for biocontrol. *Entomologia Experimentalis et Applicata*, 136(1), 97–107. https://doi.org/10.1111/j.1570-7458.2010.01006.x

Gressitt, J. L., & Gressitt, M. K. (1962). An improved Malaise trap. *Pacific Insects*, 4(1), 87–90.

Grimaldi, D., & Engel, M. S. (2005). *Evolution of the Insects*. Cambridge University Press.

Gupta, A. P. (1994). Insect Anatomy-Morphology: Quo Vadis? *Annals of the Entomological Society of America*, 87(2), 147–156. https://doi.org/10.1093/aesa/87.2.147

Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, 12(1), 28. https://doi.org/10.1186/1472-6785-12-28

Hardulak, L. A. (2020). Development and improvement of next generation sequencing pipelines for mixed and bulk samples of German fauna. Dissertation, LMU München: Fakultät für Biologie. https://doi.org/10.5282/edoc.27000

Hartop, E., Srivathsan, A., Ronquist, F., & Meier, R. (2022). Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa. *Systematic Biology*, 71(6), 1404–1422. https://doi.org/10.1093/sysbio/syac033

Hashemi-Aghdam, S. S., Rafie, G., Akbari, S., & Oshaghi, M. A. (2017). Utility of mtDNA-COI Barcode Region for Phylogenetic Relationship and Diagnosis of Five Common Pest Cockroaches. *Journal of Arthropod-Borne Diseases*, 11(2), 182–193. PMID: 29062843; PMCID: PMC5641607.

Hausmann, A., Godfray, H. C. J., Huemer, P., Mutanen, M., Rougerie, R., Nieukerken, E. J. van, Ratnasingham, S., & Hebert, P. D. N. (2013). Genetic Patterns in European Geometrid Moths Revealed by the Barcode Index Number (BIN) System. *PLOS ONE*, 8(12), e84518. https://doi.org/10.1371/journal.pone.0084518

Hausmann, A., Krogmann, L., Peters, R., Rduch, V., & Schmidt, S. (2020). GBOL III: DARK TAXA. *iBOL Barcode Bulletin*, 10(1). https://doi.org/10.21083/ibol.v10i1.6242

Hawlitschek, O., Morinière, J., Lehmann, G. U. C., Lehmann, A. W., Kropf, M., Dunz, A., Glaw, F., Detcharoen, M., Schmidt, S., Hausmann, A., Szucsich, N. U., Caetano-Wyler, S. A., & Haszprunar, G. (2017). DNA barcoding of crickets, katydids and grasshoppers (Orthoptera) from Central Europe with focus on Austria, Germany and Switzerland. *Molecular Ecology Resources*, 17(5), 1037–1053. https://doi.org/10.1111/1755-0998.12638

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLOS Genetics*, 6(2), e1000834. https://doi.org/10.1371/journal.pgen.1000834

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., deWaard, J. R., & Landry, J.-F. (2010). DNA barcodes for 1/1000 of the animal kingdom. *Biology Letters*, 6(3), 359–362. https://doi.org/10.1098/rsbl.2009.0848

Hebert, P. D. N., & Gregory, T. R. (2005). The Promise of DNA Barcoding for Taxonomy. *Systematic Biology*, 54(5), 852–859. https://doi.org/10.1080/10635150500354886

Hebert, P. D. N., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Levesque-Beaudin, V., Milton, M. A.,

Pedersen, S., Jannetta, P., & deWaard, J. R. (2016). Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702), 20150333. https://doi.org/10.1098/rstb.2015.0333

Hendrich, L., Morinière, J., Haszprunar, G., Hebert, P. D. N., Hausmann, A., Köhler, F., & Balke, M. (2015). A comprehensive DNA barcode database for Central European beetles with a focus on Germany: Adding more than 3500 identified species to BOLD. *Molecular Ecology Resources*, 15(4), 795–818. https://doi.org/10.1111/1755-0998.12354

Hodkinson, D., & Casson, D. (1991). A lesser predilection for bugs: Hemiptera (Insecta) diversity in tropical rain forests. *Biological Journal of the Linnean Society*, *43(2*), 101–109. https://doi.org/10.1111/j.1095-8312.1991.tb00587.x

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *108(4*9), 19451–19452. https://doi.org/10.1073/pnas.1116812108

Imtiaz, A., Nor, S. A. M., & Naim, D. M. (2017). Review: Progress and potential of DNA barcoding for species identification of fish species. *Biodiversitas Journal of Biological Diversity*, 18(4), 1394–1405. https://doi.org/10.13057/biodiv/d180415

Janzen, D. H. (1991). How to Save Tropical Biodiversity: The National Biodiversity Institute of Costa Rica. *American Entomologist*, 37(3), 159–171. https://doi.org/10.1093/ae/37.3.159

Jaschhof, M., & Jaschhof, C. (2022). A Review of New Zealand Porricondylinae (Diptera: Cecidomyiidae). *Proceedings of the Entomological Society of Washington*, 124(3), 459–478. https://doi.org/10.4289/0013-8797.124.3.459

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16(10), 1245–1257. https://doi.org/10.1111/ele.12162

Jinbo, U., Kato, T., & Ito, M. (2011). Current progress in DNA barcoding and future implications for entomology. *Entomological Science*, 14(2), 107–124. https://doi.org/10.1111/j.1479-8298.2011.00449.x

Jordan, A., Patch, H. M., Grozinger, C. M., & Khanna, V. (2021). Economic Dependence and Vulnerability of United States Agricultural Sector on Insect-Mediated Pollination Service. *Environmental Science & Technology*. 55(4), 2243-2253. https://doi.org/10.1021/acs.est.0c04786

Karima, Z. (2021). Chironomidae: Biology, Ecology and Systematics. In F. K. Perveen (Ed., The *Wonders of Diptera—Characteristics, Diversity, and Significance for the World's Ecosystems*. IntechOpen. https://doi.org/10.5772/intechopen.95577

Karlsson, D., Forshage, M., Holston, K., & Ronquist, F. (2020). The data of the Swedish Malaise Trap Project, a countrywide inventory of Sweden's insect fauna. *Biodiversity Data Journal*, 8, e56286. https://doi.org/10.3897/BDJ.8.e56286

Kaur, S. (2015). DNA Barcoding and Its Applications. *International Journal of Engineering Research and General Science,* 3(2), 602–604.

Keck, F., Vasselon, V., Rimet, F., Bouchez, A., & Kahlert, M. (2018). Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. *Molecular Ecology Resources*, 18(6), 1299–1309. https://doi.org/10.1111/1755-0998.12919

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, 15(5), 266–274. https://doi.org/10.1002/fee.1490

Kirk-Spriggs, A. H., & Sinclair, B. J. (2017). Manual of Afrotropical Diptera Volume 1. SANBI Publishing. http://opus.sanbi.org:80/jspui/handle/20.500.12143/6089

Kirse, A., Bourlat, S., Langen, K., Zapke, B., & Zizka, V. (2022). *Comparison of destructive and non-destructive DNA extraction methods for the metabarcoding of arthropod bulk samples*. Molecular Ecology Resources, 23(1), 92–105.https://doi.org/10.1111/1755-0998.13694

Klink, R. van, August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F., Høye, T. T., Jongejans, E., Menz, M. H. M., Miraldo, A., Roslin, T., Roy, H. E., Ruczyński, I., Schigel, D., Schäffler, L., Sheard, J. K., Svenningsen, C., Tschan, G. F., Wäldchen, J., … Bowler, D. E. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*, 37(10), 872–885. https://doi.org/10.1016/j.tree.2022.06.001

Kmiec, B., Woloszynska, M., & Janska, H. (2006). Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. *Current Genetics*, 50(3), 149–159. https://doi.org/10.1007/s00294-006-0082-1

Krell, F.-T. (2004). Parataxonomy vs. Taxonomy in biodiversity studies—Pitfalls and applicability of "morphospecies" sorting. *Biodiversity and Conservation*, 13(4), 795–812. https://doi.org/10.1023/B:BIOC.0000011727.53780.63

Kremen, C., Colwell, R. K., Erwin, T. L., Murphy, D. D., Noss, R. F., & Sanjayan, M. A. (1993). Terrestrial Arthropod Assemblages: Their Use in Conservation Planning. *Conservation Biology*, 7(4), 796–808. https://doi.org/10.1046/j.1523-1739.1993.740796.x

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, 28(2), 420–430. https://doi.org/10.1111/mec.14920

Leandro, C., & Jay-Robert, P. (2019). Perceptions and representations of animal diversity: Where did the insects go? *Biological Conservation*, 237, 400–408. https://doi.org/10.1016/j.biocon.2019.07.031

Lehmann, J. (1970). Revision der Europäischen Arten (Imagines ♂♂) der Gattung Parachironomus Lenz (Diptera, Chironomidae). *Hydrobiologia*, 36(1), 129–158. https://doi.org/10.1007/BF00751287

Lim, G. S., Balke, M., & Meier, R. (2012). Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods. *Systematic Biology*, 61(1), 165–169.

https://doi.org/10.1093/sysbio/syr030

Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A., & Vogler, A. P. (2016). Lessons from genome skimming of arthropod-preserving ethanol. *Molecular Ecology Resources*, 16(6), 1365–1377. https://doi.org/10.1111/1755-0998.12539

Littlefair, J. E., & Clare, E. L. (2016). Barcoding the food chain: From Sanger to high-throughput sequencing. *Genome*, 59(11), 946–958. https://doi.org/10.1139/gen-2016-0028

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, 45(3), 373–385. https://doi.org/10.1111/een.12831

Liu, X., Zhang, L., & Hong, S. (2011). Global biodiversity research during 1900–2009: A bibliometric analysis. *Biodiversity and Conservation*, 20, 807–826.

Lohse, K., & Ross, L. (2015). What haplodiploids can teach us about hybridization and speciation. *Molecular Ecology*, 24(20), 5075–5077. https://doi.org/10.1111/mec.13393

Longino, J. T. (2009). Tropical Insect Diversity—How to Sample It. *Tropical Biology and Conservation Management, Volume XI: Case Studies*, 35.

Losey, J. E., & Vaughan, M. (2006). The economic value of ecological services provided by insects. *Bioscience*, 56(4), 311–323. https://doi.org/10.1641/0006-3568(2006)56[311:TEVOES]2.0.CO;2

Loucks, D. P., Van Beek, E., Loucks, D. P., & van Beek, E. (2017). System sensitivity and uncertainty analysis. In D. P. Loucks & E. van Beek (Eds.),*Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications*(pp. 331–374). Springer. https://doi.org/10.1007/978-3-319-44234-1_8

Lundström, J. O., Brodin, Y., Schäfer, M. L., Vinnersten, T. Z. P., & Östman, Ö. (2010). High species richness of Chironomidae (Diptera) in temporary flooded wetlands associated with high species turn-over rates. *Bulletin of Entomological Research*, 100(4), 433–444. https://doi.org/10.1017/S0007485309990472

Malaise, R. (1937). A new insect-trap. *Entomologisk Tidskrift*, 58, 148–160.

Mallet, J., & Willmott, K. (2003). Taxonomy: Renaissance or Tower of Babel?*Trends in Ecology & Evolution*, 18(2), 57–59. https://doi.org/10.1016/S0169-5347(02)00061-7

Mammides, C. (2019). European Union's conservation efforts are taxonomically biased. *Biodiversity and Conservation*, 28(5), 1291–1296. https://doi.org/10.1007/s10531-019-01725-8

Mammola, S., Riccardi, N., Prié, V., Correia, R., Cardoso, P., Lopes-Lima, M., & Sousa, R. (2020). Towards a taxonomically unbiased European Union biodiversity strategy for 2030. *Proceedings of the Royal Society B*, 287(1940), 20202166. https://doi.org/10.1098/rspb.2020.2166

Mandelik, Y., Roll, U., & Fleischer, A. (2010). Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *Journal of Applied Ecology*, 47(6), 1179–1188. https://doi.org/10.1111/j.1365-2664.2010.01864.x

Manktelow, M. (2010). History of taxonomy. *Lecture from Dept. of Systematic Biology, Uppsala University*, *29*. http://www.atbi.eu/summerschool/files/summerschool/Manktelow_Syllabus.pdf

Marquina, D., Buczek, M., Ronquist, F., & Łukasik, P. (2020). The effect of ethanol concentration on the morphological and molecular preservation of insects for biodiversity studies. *PeerJ*, 9, e10799. https://doi.org/10.1101/2020.03.05.978288

Marquina, D., Esparza Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources*, 19(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

Martín-López, B., García-Llorente, M., Palomo, I., & Montes, C. (2011). The conservation against development paradigm in protected areas: Valuation of ecosystem services in the Doñana social–ecological system (southwestern Spain). *Ecological Economics*, 70(8), 1481–1491. https://doi.org/10.1016/j.ecolecon.2011.03.009

Marziali, L., Armanini, D. G., Cazzola, M., Erba, S., Toppi, E., Buffagni, A., & Rossaro, B. (2010). Responses of Chironomid larvae (insecta, Diptera) to ecological quality in Mediterranean river mesohabitats (South Italy). *River Research and Applications*, 26(8), 1036–1051. https://doi.org/10.1002/rra.1303

Mata, V. A., Ferreira, S., Campos, R. M., da Silva, L. P., Veríssimo, J., Corley, M. F. V., & Beja, P. (2021). Efficient assessment of nocturnal flying insect communities by combining automatic light traps and DNA metabarcoding. *Environmental DNA*, 3(2), 398–408. https://doi.org/10.1002/edn3.125

Matthews, R., & Matthews, J. (2017). The Malaise trap: Its utility and potential for sampling insect populations. *The Great Lakes Entomologist*, *4*(4). https://scholar.valpo.edu/tgle/vol4/iss4/4

May, R. M. (1988). How Many Species Are There on Earth? *Science*, 241(4872), 1441–1449. https://doi.org/10.1126/science.241.4872.1441

Mayhew, P. J. (2007). Why are there so many insect species? Perspectives from fossils and phylogenies. *Biological Reviews of the Cambridge Philosophical Society*, 82(3), 425–454. https://doi.org/10.1111/j.1469-185X.2007.00018.x

Meier, R., Blaimer, B. B., Buenaventura, E., Hartop, E., Rintelen, T., Srivathsan, A., & Yeo, D. (2022). A re analysis of the data in Sharkey et al.'s (2021) minimalist revision reveals that BINs do not deserve names, but BOLD Systems needs a stronger commitment to open science. *Cladistics*, 38(2), 264–275. https://doi.org/10.1111/cla.12489

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, 55(5), 715–728. https://doi.org/10.1080/10635150600969864

Meier, R., Srivathsan, A., Oliveira, S., Balbi, M., Ang, Y., Yeo, D., & Amorim, D. (2023). "Dark taxonomy": A new protocol for overcoming the taxonomic impediments for dark taxa and broadening the taxon base for biodiversity assessment. *bioRxiv*, 2023-08.

https://doi.org/10.1101/2023.08.31.555664

Meier, R., Wong, W., Srivathsan, A., & Foo, M. (2016). $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics*, 32(1), 100–110. https://doi.org/10.1111/cla.12115

Meier, R., Zhang, G., & Ali, F. (2008). The Use of Mean Instead of Smallest Interspecific Distances Exaggerates the Size of the "Barcoding Gap" and Leads to Misidentification. *Systematic Biology*, 57(5), 809–813. https://doi.org/10.1080/10635150802406343

Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLOS ONE*, 14(6), e0217084. https://doi.org/10.1371/journal.pone.0217084

Miller, S. E., Hausmann, A., Hallwachs, W., & Janzen, D. H. (2016). Advancing taxonomy and bioinventories with DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702), 20150339. https://doi.org/10.1098/rstb.2015.0339

Montgomery, G. A., Belitz, M. W., Guralnick, R. P., & Tingley, M. W. (2021). Standards and Best Practices for Monitoring and Benchmarking Insects. *Frontiers in Ecology and Evolution*, 8, 513. https://www.frontiersin.org/articles/10.3389/fevo.2020.579193

Montgomery, G. A., Dunn, R. R., Fox, R., Jongejans, E., Leather, S. R., Saunders, M. E., Shortall, C. R., Tingley, M. W., & Wagner, D. L. (2020). Is the insect apocalypse upon us? How to find out. *Biological Conservation*, 241, 108327. https://doi.org/10.1016/j.biocon.2019.108327

Morinière, J., Araujo, B. C. de, Lam, A. W., Hausmann, A., Balke, M., Schmidt, S., Hendrich, L., Doczkal, D., Fartmann, B., Arvidsson, S., & Haszprunar, G. (2016). Species identification in Malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLOS ONE*, 11(5), e0155497. https://doi.org/10.1371/journal.pone.0155497

Morinière, J., Balke, M., Doczkal, D., Geiger, M. F., Hardulak, L. A., Haszprunar, G., Hausmann, A., Hendrich, L., Regalado, L., Rulik, B., Schmidt, S., Wägele, J.-W., & Hebert, P. D. N. (2019). A

DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring. *Molecular Ecology Resources*, 19(4), 900–928. https://doi.org/10.1111/1755-0998.13022

Morse, J. C., Footitt, R. G., & Adler, P. H. (2017). *Insect biodiversity: Science and society*.

Nichols, E., Spector, S., Louzada, J., Larsen, T., Amezquita, S., & Favila, M. E. (2008). Ecological functions and ecosystem services provided by Scarabaeinae dung beetles. *Biological Conservation*, 141(6), 1461–1474. https://doi.org/10.1016/j.biocon.2008.04.011

Normark, B. B. (2003). The Evolution of Alternative Genetic Systems in Insects. *Annual Review of Entomology*, 48(1), 397–423. https://doi.org/10.1146/annurev.ento.48.091801.112703

Normark, B. B. (2004). Haplodiploidy as an outcome of coevolution between male-killing cytoplasmic elements and their hosts. *Evolution*, 58(4), 790–798. https://doi.org/10.1111/j.0014-3820.2004.tb00412.x

Novotny, V., Miller, S. E., Hulcr, J., Drew, R. A. I., Basset, Y., Janda, M., Setliff, G. P., Darrow, K., Stewart, A. J. A., Auga, J., Isua, B., Molem, K., Manumbor, M., Tamtiai, E., Mogia, M., & Weiblen, G. D. (2007). Low beta diversity of herbivorous insects in tropical forests. *Nature*, *448*(7154), 692-695. https://doi.org/10.1038/nature06021

Ødegaard, F. (2000). How many species of arthropods? Erwin's estimate revised. *Biological Journal of the Linnean Society*, 71(4), 583–597. https://doi.org/10.1111/j.1095-8312.2000.tb01279.x

Ortiz, A. S., Rubio, R. M., Guerrero, J. J., Garre, M. J., Serrano, J., Hebert, P. D. N., & Hausmann, A. (2017). Close congruence between Barcode Index Numbers (bins) and species boundaries in the Erebidae (Lepidoptera: Noctuoidea) of the Iberian Peninsula. *Biodiversity Data Journal*, 5, e19840. https://doi.org/10.3897/BDJ.5.e19840

Otto, S. P., & Jarne, P. (2001). Haploids–Hapless or Happening? *Science*, 292(5526), 2441–2443. https://doi.org/10.1126/science.1062890

Ozanne, C. M. P., Anhuf, D., Boulter, S. L., Keller, M., Kitching, R. L., Körner, C., Meinzer, F. C.,

Mitchell, A. W., Nakashizuka, T., Dias, P. L. S., Stork, N. E., Wright, S. J., & Yoshimura, M. (2003). Biodiversity Meets the Atmosphere: A Global View of Forest Canopies. *Science*, 301(5630), 183–186. https://doi.org/10.1126/science.1084507

Page, R. D. M. (2016). DNA barcoding and taxonomy: Dark taxa and dark texts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702), 20150334. https://doi.org/10.1098/rstb.2015.0334

Pamilo, P., Viljakainen, L., & Vihavainen, A. (2007). Exceptionally High Density of NUMTs in the Honeybee Genome. *Molecular Biology and Evolution*, 24(6), 1340–1346. https://doi.org/10.1093/molbev/msm055

Parker, W. S. (2010). Scientific Models and Adequacy-for-Purpose. *The Modern Schoolman*, 87(3/4), 285–293. https://doi.org/10.5840/schoolman2010873/410

Patten, M. M., Carioscia, S. A., & Linnen, C. R. (2015). Biased introgression of mitochondrial and nuclear genes: A comparison of diploid and haplodiploid systems. *Molecular Ecology*, 24(20), 5200–5210. https://doi.org/10.1111/mec.13318

Phillips, J. D., Gillis, D. J., & Hanner, R. H. (2022). Lack of Statistical Rigor in DNA Barcoding Likely Invalidates the Presence of a True Species' Barcode Gap. *Frontiers in Ecology and Evolution*, 10, 859099. https://doi.org/10.3389/fevo.2022.859099

Pillot, H. K. M. (2008). Identification and ecology of the genus Smittia Holmgren in the Netherlands (Diptera: Chironomidae). *Tijdschrift Voor Entomologie*, 151, 245–270.

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

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, *4*, 209–230. https://doi.org/10.1023/A:1011994105352

Porter, T. M., Gibson, J. F., Shokralla, S., Baird, D. J., Golding, G. B., & Hajibabaei, M. (2014). Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Molecular Ecology Resources*, 14(5), 929–942. https://doi.org/10.1111/1755-0998.12240

Prather, C. M., & Laws, A. N. (2018). Insects as a piece of the puzzle to mitigate global problems: An opportunity for ecologists. *Basic and Applied Ecology*, 26, 71–81. https://doi.org/10.1016/j.baae.2017.09.009

Proulx, I., Martin, J., Carew, M., & Hare, L. (2013). Using various lines of evidence to identify Chironomus species (Diptera: Chironomidae) in eastern Canadian lakes. *Zootaxa*, 3741, 401–458. https://doi.org/10.11646/zootaxa.3741.4.1

Puillandre, N., Brouillet, S., & Achaz, G. (2021). ASAP: Assemble species by automatic partitioning. *Molecular Ecology Resources*, 21(2), 609–620. https://doi.org/10.1111/1755-0998.13281

Qiang-long, Z., Shi, L., Peng, G., & Fei-shi, L. (2014). High-throughput Sequencing Technology and Its Application. *Journal of Northeast Agricultural University (English Edition)*, 21(3), 84–96. https://doi.org/10.1016/S1006-8104(14)60073-8

Raffoul, J. J., Heydari, A. R., & Hillman, G. G. (2012). DNA Repair and Cancer Therapy: Targeting APE1/Ref-1 Using Dietary Agents. *Journal of Oncology*, *2012*, e370481. https://doi.org/10.1155/2012/370481

Ramos, D. L., Cunha, W. L., Evangelista, J., Lira, L. A., Rocha, M. V. C., Gomes, P. A., Frizzas, M. R., & Togni, P. H. B. (2020). Ecosystem services provided by insects in Brazil: What do we really know? *Neotropical Entomology*, 49, 783–794. https://doi.org/10.1007/s13744-020-00781-y

Rands, M. R. W., Adams, W. M., Bennun, L., Butchart, S. H. M., Clements, A., Coomes, D., Entwistle, A., Hodge, I., Kapos, V., Scharlemann, J. P. W., Sutherland, W. J., & Vira, B. (2010). Biodiversity Conservation: Challenges Beyond 2010. *Science*, 329(5997), 1298–1303. https://doi.org/10.1126/science.1189138

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. *PLOS ONE*, *8*(7), e66213. https://doi.org/10.1371/journal.pone.0066213

Raunio, J., Heino, J., & Paasivirta, L. (2011). Non-biting midges in biodiversity conservation and environmental assessment: Findings from boreal freshwater ecosystems. *Ecological Indicators*, 11(5), 1057–1064. https://doi.org/10.1016/j.ecolind.2010.12.002

Raupach, M. J., Astrin, J. J., Hannig, K., Peters, M. K., Stoeckle, M. Y., & Wägele, J.-W. (2010). Molecular species identification of Central European ground beetles (Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Frontiers in Zoology*, 7(1), 26. https://doi.org/10.1186/1742-9994-7-26

Reid, B. N., Le, M., McCord, W. P., Iverson, J. B., Georges, A., Bergmann, T., Amato, G., Desalle, R., & Naro-Maciel, E. (2011). Comparing and combining distance-based and character-based approaches for barcoding turtles. *Molecular Ecology Resources*, 11(6), 956–967. https://doi.org/10.1111/j.1755-0998.2011.03032.x

Restani, M., & Marzluff, J. M. (2002). Funding Extinction? Biological Needs and Political Realities in the Allocation of Resources to Endangered Species Recovery. *BioScience*, 52(2), 169. https://doi.org/10.1641/0006-3568(2002)052[0169:FEBNAP]2.0.CO;2

Rolo, E. S. A. (2010). DNA barcoding and forensic entomology: A molecular approach for Diptera species' identification. Dissertation. University of Lisbon.

Rosenberg, D. M. (1992). Freshwater biomonitoring and Chironomidae. *Netherland Journal of Aquatic Ecology*, 26(2), 101–122. https://doi.org/10.1007/BF02255231

Saether, O. A. (1971). Nomenclature and phylogeny of the genus *Harnischia* (Diptera: Chironomidae).

The Canadian Entomologist, 103(3), 347–362. https://doi.org/10.4039/Ent103347-3

Sánchez-Fernández, D., Abellán, P., Aragón, P., Varela, S., & Cabeza, M. (2018). Matches and mismatches between conservation investments and biodiversity values in the European Union. *Conservation Biology: The Journal of the Society for Conservation Biology*, 32(1), 109–115. https://doi.org/10.1111/cobi.12977

Schilthuizen, M., Scholte, C., van Wijk, R. E., Dommershuijzen, J., van der Horst, D., Zu Schlochtern, M. M., Lievers, R., & Groenenberg, D. S. (2011). Using DNA-barcoding to make the necrobiont beetle family Cholevidae accessible for forensic entomology.*Forensic Science International*, 210(1–3), 91–95. https://doi.org/10.1016/j.forsciint.2011.02.003

Schlick-Steiner, B. C., Steiner, F. M., Seifert, B., Stauffer, C., Christian, E., & Crozier, R. H. (2010). Integrative Taxonomy: A Multisource Approach to Exploring Biodiversity. *Annual Review of Entomology*, 55(1), 421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Schmiedel, U., Araya, Y., Bortolotto, M. I., Boeckenhoff, L., Hallwachs, W., Janzen, D., Kolipaka, S. S., Novotny, V., Palm, M., Parfondry, M., Smanis, A., & Toko, P. (2016). Contributions of paraecologists and parataxonomists to research, conservation, and social development. *Conservation Biology*, 30(3), 506–519. https://doi.org/10.1111/cobi.12661

Scudder, G. G. E. (2017). The Importance of Insects. In R .G. Foottit & P. H. Adler (Eds.) *Insect Biodiversity: Science and Society* (pp. 9–43). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118945568.ch2

Ševčík, J., Kaspřák, D., Mantič, M., Fitzgerald, S., Ševčíková, T., Tóthová, A., & Jaschhof, M. (2016). Molecular phylogeny of the megadiverse insect infraorder Bibionomorpha sensu lato (Diptera). *PeerJ*, 4, e2563. https://doi.org/10.7717/peerj.2563

Shadrin, D. M. (2021). DNA Barcoding: Applications. *Russian Journal of Genetics*, *57*(4), 489–497. https://doi.org/10.1134/S102279542104013X

Sheikh, A. H., Thomas, M., Bhandari, R., & Meshram, H. (2016). Malaise trap and insect sampling:

Mini Review. *Bio Bulletin* 2(2), 35–40.

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21(8), 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Shum, P., & Palumbi, S. R. (2021). Testing small scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. *Molecular Ecology*, 30(13), 3355–3373. https://doi.org/10.1111/mec.15851

Shurin, J. B., Gruner, D. S., & Hillebrand, H. (2005). All wet or dried up? Real differences between aquatic and terrestrial food webs. *Proceedings of the Royal Society B: Biological Sciences*, 273(1582), 1–9. https://doi.org/10.1098/rspb.2005.3377

Skvarla, M. J., Larson, J. L., Fisher, J. R., & Dowling, A. P. G. (2021). A Review of Terrestrial and Canopy Malaise Traps. *Annals of the Entomological Society of America*, 114(1), 27–47. https://doi.org/10.1093/aesa/saaa044

Sneli, J.-A., Knudsen, J., & Vedelsby, A. (2009). Johan Christian Fabricius and his molluscan species, Acesta excavata (JC Fabricius, 1779). *Steenstrupia*, *30*(2), 153–162.

Snodgrass, R. E. (1935). The abdominal mechanisms of a grasshopper. *Smithsonian Miscellaneous Collections*.

Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, 9(1), 640. https://doi.org/10.1038/msb.2012.61

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W. X., Bertrand, D., Ng, A. H. Q., Boey, E. J. H., Koh, J. J. Y., Nagarajan, N., & Meier, R. (2018). A MinION™-based pipeline for fast and cost-effective DNA barcoding. *Molecular Ecology Resources*, 18(5), 1035–1049. https://doi.org/10.1111/1755-0998.12890

Srivathsan, A., Sha, J. C. M., Vogler, A. P., & Meier, R. (2015). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (Pygathrix

nemaeus). *Molecular Ecology Resources*, 15(2), 250–261. https://doi.org/10.1111/1755-0998.12302

Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, 76(5), 650–661. https://doi.org/10.1086/605819

Stoeckle, M. Y., & Hebert, P. D. N. (2008). Barcode of Life. *Scientific American*, 299(4), 82–89. http://www.jstor.org/stable/26000849.

Stork, N. E. (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, 63(1), 31–45. https://doi.org/10.1146/annurev-ento-020117-043348

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tang, H., Cheng, Q., Han, W., & Cranston, P. S. (2022). Integrative taxonomy: Molecular phylogenetics of Polypedilum (Cerobregma) and revisited morphology of Yaethauma and Collartomyia (Diptera: Chironomidae) reveals synonymy and supports new classification. *Zoological Journal of the Linnean Society*, 194(1), 102–119. https://doi.org/10.1093/zoolinnean/zlaa187

Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations. *Molecular Ecology Resources*, 19(1), 47–76. https://doi.org/10.1111/1755-0998.12959

Titley, M. A., Snaddon, J. L., & Turner, E. C. (2017). Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. *PloS One*, 12(12), e0189577. https://doi.org/10.1371/journal.pone.0189577

Touchman, J. W. (2009). DNA Sequencing: An Outsourcing Guide. *Current Protocols Essential Laboratory Techniques*, 2(1), 12. https://doi.org/10.1002/9780470089941.et1201s02

Uhler, J., Haase, P., Hoffmann, L., Hothorn, T., Schmidl, J., Stoll, S., Welti, E. A. R., Buse, J., & Müller, J. (2022). A comparison of different Malaise trap types. *Insect Conservation and Diversity*, 15(6), 666–672. https://doi.org/10.1111/icad.12604

van Huis, A. (2014). *The Global Impact of Insects*. Wageningen University. https://edepot.wur.nl/410394

Vårdal, H., & Taeger, A. (2011). The life of René Malaise: From the wild east to a sunken island. *Zootaxa*, 3127, 38–52.

Vasilita, C., Feng, V., Hansen, A. K., Hartop, E., Srivathsan, A., Struijk, R., & Meier, R. (2023). *Express barcoding with NextGenPCR and MinION for species-level sorting of ecological samples* (p. 2023.04.27.538648). bioRxiv. https://doi.org/10.1101/2023.04.27.538648

Vilgalys, R. (2003). Taxonomic misidentification in public DNA databases. *New Phytologist*, *160*(1), 4–5. https://doi.org/10.1046/j.1469-8137.2003.00894.x

Virgilio, M., Backeljau, T., Nevado, B., & De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics*, 11(1), 206. https://doi.org/10.1186/1471-2105-11-206

Vogt, L., Bartolomaeus, T., & Giribet, G. (2010). The linguistic problem of morphology: Structure versus homology and the standardization of morphological data. *Cladistics*, 26(3), 301–325. https://doi.org/10.1111/j.1096-0031.2009.00286.x

von Berg, K., Traugott, M., Symondson, W. O. C., & Scheu, S. (2008). The effects of temperature on detection of prey DNA in two species of carabid beetle. *Bulletin of Entomological Research*, 98(3), 263–269. https://doi.org/10.1017/S0007485308006020

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., & Meier, R. (2018). Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. *Molecular Ecology Resources*, 18(3), 490–501. https://doi.org/10.1111/1755-0998.12751

Wells, J. D., & Stevens, J. R. (2008). Application of DNA-based methods in forensic entomology. *Annu.*

Rev. Entomol., 53, 103–120. https://doi.org/10.1146/annurev.ento.52.110405.091423

Wetton, J. H., Tsang, C. S. F., Roney, C. A., & Spriggs, A. C. (2004). An extremely sensitive species-specific ARMs PCR test for the presence of tiger bone DNA. *Forensic Science International*, 140(1), 139–145. https://doi.org/10.1016/j.forsciint.2003.11.018

What is the LIFE programme? (2023). Life Croaa. https://www.life-croaa.eu/en/what-is-the-life-programme/

Wiederholm, T. (1989). Chironomidae of the Holarctic region. Keys and diagnoses. Part 3-Adult males. *Entomologica Scandinavian Supplements*, 34, 1–532.

Will, K. W., Mishler, B. D., & Wheeler, Q. D. (2005). The Perils of DNA Barcoding and the Need for Integrative Taxonomy. *Systematic Biology*, 54(5), 844–851. https://doi.org/10.1080/10635150500354878

Wilson. (1987). The Little Things That Run the World (The Importance and Conservation of Invertebrates). *Conservation Biology*, 1(4), 344–346. https://www.jstor.org/stable/2386020

Wilson, J.-J., Sing, K.-W., & Jaturas, N. (2019). DNA Barcoding: Bioinformatics Workflows for Beginners. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 985–995). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20468-8

Wu, H., Wan, Q.-H., Fang, S.-G., & Zhang, S.-Y. (2005). Application of mitochondrial DNA sequence analysis in the forensic identification of Chinese sika deer subspecies. *Forensic Science International*, 148(2), 101–105. https://doi.org/10.1016/j.forsciint.2004.04.072

Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M., Schmidt, S., Cerretti, P., & Meier, R. (2022). DiversityScanner: Robotic handling of small invertebrates with machine learning methods. *Molecular Ecology Resources*, 22(4), 1626–1638. https://doi.org/10.1111/1755-0998.13567

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods*

in Ecology and Evolution, 3(4), 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zahiri, R., Lafontaine, J. D., Schmidt, B. C., deWaard, J. R., Zakharov, E. V., & Hebert, P. D. N. (2014). A Transcontinental Challenge—A Test of DNA Barcode Performance for 1,541 Species of Canadian Noctuoidea (Lepidoptera). *PLOS ONE*, 9(3), e92797. https://doi.org/10.1371/journal.pone.0092797

Zenker, M. M., Specht, A., & Fonseca, V. G. (2020). Assessing insect biodiversity with automatic light traps in Brazil: Pearls and pitfalls of metabarcoding samples in preservative ethanol. *Ecology and Evolution*, 10(5), 2352–2366. https://doi.org/10.1002/ece3.6042

Zizka, V. M., Geiger, M. F., Hörren, T., Kirse, A., Noll, N. W., Schäffler, L., Scherges, A. M., & Sorg, M. (2022). *Recommendations for tissue homogenisation and extraction in DNA metabarcoding of Malaise trap samples* (p. 2022.01.25.477667). bioRxiv. https://doi.org/10.1101/2022.01.25.477667

ACKNOWLEDGEMENTS

Thank you to my family, ANIA, EMILIO, IVAN & ARTHUR, and to my extended family members, MIO & JESSY, for being the best support system that one can ask for, and for always having my back throughout my crazies of ideas.

To VIKTOR BARANOV, who was not only my supervisor and mentor (the best!) throughout this journey, but who has also become a very dear friend. Thank you for your blunt honesty and unfiltered opinions in all fields of discussion, and for guiding me through the horrors of academia. May you drink much sangria and soak up much sun in Seville, Spain.

To the team at the Bavarian State Collection of Zoology: DR. STEFAN SCHMIDT, DR. AXEL HAUSMANN, PD DR. MICHAEL RAUPACH, DR. MICHAEL BALKE & PROF. DR. GERHARD HASZPRUNAR, for kindly pushing me to my limits so that I could evolve into the scientist that I have now become. To my colleagues ROSA ALBRECHT & AMELIE HÖCHERL for making good food a central part of our time at the institute, and to FABIAN DEISTER for unveiling my passion for numbers, data analysis and programming.

To ISABELLA STÖGER, for helping me keep my sanity in the laboratory and to STEFAN FILSER, for maintaining my sanity in the office. And lastly, to JÉRÔME MORINIÈRE, who is my Jedi Master in the world of DNA barcoding.

APPENDIX

Publication I 83

Publication II 103

Publication III 113

Publication IV 133

Publication V 165

Publication I:

Abundant, diverse, unknown: Extreme species richness despite drastic undersampling in two closely placed Malaise traps

Published in PLOS One

Chimeno, C., Schmidt, S., Cancian de Araujo, B., Perez, K., von Rintelen, T., Schmidt, O., Hamid, H., Pramesa Narakusumo, R. & Balke, M. (2023). Abundant, diverse, unknown: Extreme species richness and turnover despite drastic undersampling in two closely placed tropical Malaise traps. *Plos One*, 18(8): e0290173. https://doi.org/10.1371/journal.pone.0290173

PLOS ONE

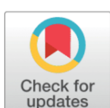
RESEARCH ARTICLE

Abundant, diverse, unknown: Extreme species richness and turnover despite drastic undersampling in two closely placed tropical Malaise traps

Caroline Chimeno[1]*, Stefan Schmidt[1], Bruno Cancian de Araujo[1,2], Kate Perez[3], Thomas von Rintelen[4], Olga Schmidt[1], Hasmiandy Hamid[5], Raden Pramesa Narakusumo[6], Michael Balke[1]

1 Zoologische Staatssammlung München (SNSB-ZSM), Munich, Germany, **2** Entomological Biodiversity Laboratory, Federal University of Espirito Santo, Vitoria, Brazil, **3** Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada, **4** Center for Integrative Biodiversity Discovery, Museum für Naturkunde—Leibniz-Institut fur Evolutions- und Biodiversitätsforschung, Berlin, Germany, **5** Department of Plant Protection, Faculty of Agriculture, Universitas Andalas, Padang, Indonesia, **6** Research Center for Biosystematics and Evolution, National Research and Innovation Agency (BRIN), Museum Zoologicum Bogoriense, Cibinong, Indonesia

* chimeno@snsb.de



OPEN ACCESS

Citation: Chimeno C, Schmidt S, Cancian de Araujo B, Perez K, von Rintelen T, Schmidt O, et al. (2023) Abundant, diverse, unknown: Extreme species richness and turnover despite drastic undersampling in two closely placed tropical Malaise traps. PLoS ONE 18(8): e0290173. https://doi.org/10.1371/journal.pone.0290173

Editor: Pierfilippo Cerretti, Universita degli Studi di Roma La Sapienza, ITALY

Received: April 27, 2023

Accepted: August 3, 2023

Published: August 16, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0290173

Copyright: © 2023 Chimeno et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All original data spreadsheets that were downloaded from BOLD for analysis have been uploaded to Figshare (https://

Abstract

Arthropods account for a large proportion of animal biomass and diversity in terrestrial systems, making them crucial organisms in our environments. However, still too little is known about the highly abundant and megadiverse groups that often make up the bulk of collected samples, especially in the tropics. With molecular identification techniques ever more evolving, analysis of arthropod communities has accelerated. In our study, which was conducted within the Global Malaise trap Program (GMP) framework, we operated two closely placed Malaise traps in Padang, Sumatra, for three months. We analyzed the samples by DNA barcoding and sequenced a total of more than 70,000 insect specimens. For sequence clustering, we applied three different delimitation techniques, namely RESL, ASAP, and SpeciesIdentifier, which gave similar results. Despite our (very) limited sampling in time and space, our efforts recovered more than 10,000 BINs, of which the majority are associated with “dark taxa”. Further analysis indicates a drastic undersampling of both sampling sites, meaning that the true arthropod diversity at our sampling sites is even higher. Regardless of the close proximity of both Malaise traps (< 360 m), we discovered significantly distinct communities.

Introduction

In the age of rapid biodiversity decline, taxonomists find themselves in a race against time to discover and describe new species before they become extinct [1–5]. However, identifying species in several megadiverse groups of organisms requires in-depth taxonomic expertise, which

doi.org/10.6084/m9.figshare.21815034). The R script and all input data sets are also deposited on Figshare (R code: https://doi.org/10.6084/m9.figshare.21806370.v2; BIN dataset: https://doi.org/10.6084/m9.figshare.21815142; ASAP dataset: https://doi.org/10.6084/m9.figshare.21815064.v1). The datasets on BOLD can be found under doi.org/10.5883/DS-GMTINDO1 and doi.org/10.5883/DS-GMTINDO2.

Funding: The project was supported by the Deutsche Forschungsgemeinschaft and the Bundesministerium für Bildung und Forschung (BMBF) within the bilateral "Biodiversity and Health" funding program (Project numbers: 16GW0111K, 16GW0112) with additional support from DIPA PUSLIT Biologi LIPI 2015-2016. The sequence analyses for this study were supported, in part, by Genome Canada through the Ontario Genomics Institute, while informatics support was provided through a grant from the Ontario Ministry of Research and Innovation. The funders did not play any role in the 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.

is either in decline or very limited, the latter being the case in the so-called dark taxa [6, 7]. This mismatch between high species numbers awaiting discovery and few researchers available for doing so is also known as the "taxonomic impediment". It is prominent among arthropods [8], and considering that arthropods account for a large proportion of the animal biomass and diversity in terrestrial ecosystems [9–11], is a direct constraint to global biodiversity research. Often, ecological surveys must limit their analyses to a subset of known species (e.g., flagship indicator species) because there is not enough know-how to analyze the highly abundant, often minute specimens that make up the bulk of the sample [12, 13].

As a potential remedy, molecular identification techniques have greatly evolved in the last decade, providing accelerated sample processing methodologies in various fields of research [14]. DNA barcoding, for example, is a method that uses a short DNA sequence of the COI gene in the mitochondrial DNA to identify and distinguish species from one another [15–17]. Paul Hebert and colleagues first introduced it in 2003, and today, it is a standard approach for molecular identification or presorting species [18]. DNA barcoding is easy to use [even for non-experts], widely available, and nowadays economic [16, 19–22].

In 2012, the Global Malaise trap Program was initiated by the Centre for Biodiversity Genomics (CBG) at the Biodiversity Institute of Ontario (BIO) with the large-scale worldwide deployment of Malaise traps (see https://biodiversitygenomics.net/site/projects/gmp/). Malaise traps are very efficient at capturing flying insects and are, therefore, commonly used in surveys of terrestrial arthropods [23–26]. More than 158 sites in 33 countries were sampled and analyzed via DNA barcoding to provide an overview of the global arthropod biodiversity and provide detailed temporal and spatial information on arthropod communities (see https://biodiversitygenomics.net/site/projects/gmp/). In a joint project with the Andalas University, two Malaise traps were deployed in Padang, Sumatra, Indonesia, and operated for three months each. Insect communities in tropical regions are notorious for being extraordinarily diverse [11, 27, 28] yet severely understudied [29, 30], making the large-scale sequencing of the Malaise traps contents especially interesting. In this study, we present and evaluate the sequencing results recovered for each Malaise trap.

Materials and methods

Collecting

In 2016, we deployed two Malaise traps, installed ca. 360 m apart from each other. We set up the traps at the northern forest edge of the 500-hectare campus area of the University of Andalas at the eastern part of Padang City, West Sumatra Province, Indonesia (Fig 1). The traps were located in a semi-open area dominated by ferns, interspersed with medium-sized and a few large-sized trees. The trap locations were set up in spots with sparse vegetation in such a way that flight paths were open in both directions of the traps. The adjacent tropical forest was dominated by secondary tree vegetation and is connected to the Bukit Barisan mountain range. Both Malaise traps were operated from May 5th to July 30th. The collection bottles were emptied biweekly and topped up with fresh 80% EtOH. Samples were stored in a freezer until further processing. Because both traps were located on the grounds of the university, no collection permit was needed.

Sample processing

All collection bottles were sent to the Centre for Biodiversity Genomics for sorting and further processing. An attempt was made to barcode as many specimens as possible, but due to funding constraints, not all specimens were processed, and for some collection bottles, a maximum of fifteen 96-well microplates were filled (Table 1). A selective strategy was used

Fig 1. Collection sites. Malaise trap sites near Padang, West Sumatra. Created by the authors using QGis.

https://doi.org/10.1371/journal.pone.0290173.g001

to narrow down the number of specimens especially for the samples with only 15 plates processed. Individuals chosen for sequencing were selected to capture as much diversity as possible based on size and morphospecies. Two sizes of sieves were used to subsample from three different size classes (no sieve, 8mm sieve, and 2mm sieve). As most of the diversity was likely hidden in the smaller organisms (particularly the abundant insect orders: Hymenoptera, Coleoptera, and Diptera), more specimens were chosen from the smallest size class. Additionally, more Hymenoptera and Coleoptera were selected as opposed to Diptera because Diptera are often so abundant in Malaise trap samples that there is a higher risk of oversampling the same species.

Table 1. Malaise trap sample information. Collection dates, sequencing capacity, specimens processed, and sequences obtained per sample.

Malaise Trap	Sample Nr.	Collection Interval	Sequencing capacity	Nr. Specimens	Nr. Sequences
Trap 1	1	05–13 May	All specimens	5,969	5,101
Trap 1	2	13–20 May	15 plates	1,475	1,232
Trap 1	3	20–27 May	All specimens	9,070	5,836
Trap 1	4	27 May–03 June	15 plates	1,475	1,082
Trap 1	5	03–10 June	All specimens	9,886	7,760
Trap 1	6	10–24 June	15 plates	1,475	1,325
Trap 1	7	24 June–01 July	All specimens	6,266	5,730
Trap 1	8	01–08 July	15 plates	1,475	1,216
Trap 1	9	08–15 July	All specimens	4,230	3,776
Trap 1	10	15–22 July	15 plates	1,475	1,267
Trap 1	11	22–30 July	All specimens	9,566	8,567
Trap 2	1	05–13 May	All specimens	10,102	8,439
Trap 2	2	13–20 May	15 plates	1,475	1,229
Trap 2	3	20–27 May	15 plates	1,504	1,136
Trap 2	4	27 May–03 June	15 plates	1,475	1,169
Trap 2	5	03–10 June	15 plates	1,529	1,250
Trap 2	6	10–24 June	15 plates	1,475	1,349
Trap 2	7	24 June–01 July	15 plates	1,490	1,241
Trap 2	8	01–08 July	15 plates	1,475	1,280
Trap 2	9	08–15 July	15 plates	1,491	1,285
Trap 2	10	15–22 July	15 plates	1,475	1,272
Trap 2	11	22–30 July	15 plates	1,644	1,258

https://doi.org/10.1371/journal.pone.0290173.t001

Tissue lysis was performed overnight at 56˚C, and DNA extraction was conducted using an automated, silica membrane-based protocol [31]. To reduce costs and the amount of reagents needed for PCR amplification of the COI gene, the DNA extracts from four 96-well plates were consolidated into 386-well PCR plates [23, 32]. The PCR products were diluted, unidirectionally sequenced, then cleaned-up using an automated magnetic bead-based method before being sequences on an ABI 3730 xl DNA Analyzer (Applied Biosystems). For more details on the laboratory protocols, see [23].

All barcoded specimens are currently stored at the Center for Biodiversity Genomics (CBG) natural history archive (collection code BIOUG) at the University of Guelph, Canada. However, this collection, as well as the rest of the unprocessed material, will eventually be repatriated to Museum Zoological Bogoriense in Cibinong, Indonesia.

Data analysis

All specimen metadata and sequence data were uploaded to the Barcode of Life Data System (BOLD), an online workbench and database [32]. All data is publicly available on BOLD in two datasets (doi.org/10.5883/DS-GMTINDO1 and doi.org/10.5883/DS-GMTINDO2). We also uploaded the BOLD data spreadsheet including all metadata of specimens to Figshare.

Sequences were assigned a Barcode Index Number (BIN) by the BOLD system using the RESL-algorithm. BINs represent globally unique identifiers for clusters of sequences as a species proxy [32]. Every sequence $\geq$ 300 base pairs (bp) is automatically assigned to a Barcode Index Number (BIN) that is already in BOLD if sequence similarity based on the RESL-algorithm is fulfilled [32]. Sequences with $\geq$ 500 bp which do not find a match, serve as founders

of new BINs. Family-level identifications were conducted using the BIN taxonomy match tool on BOLD.

All analyses were performed in R version 4.2.1 [33], using the packages *vegan* version 2.5–7 [34], *iNEXT* version 2.0.20 [35], and *SpadeR* version 0.1.1 [36]. To assess our sampling effort, we created accumulation curves of BINs for each Malaise trap (via *iNEXT; iNEXT* package) and estimated the species diversity present at each sampling site (via *ChaoSpecies; SpadeR* package). We created continuous diversity profiles for each trap (via *Diversity; SpadeR* package) to illustrate the variation in the three standard metrics of biodiversity that are quantified by Hill numbers (q): species richness (q = 0), Shannon diversity (q = 1), and Simpson diversity (q = 2). Hill numbers are a mathematically consolidated group of diversity indices that include relative species abundances to quantify biodiversity. To evaluate the faunal similarity between Malaise traps, we performed permutation multivariate analysis of variance (PERMANOVA) (via *adonis2*; *vegan* package; Bray Curtis dissimilarity; 999 permutations). We differentiated between location and dispersion effects by applying a beta dispersion test analogous to Levene's test (via *betadisper*; *vegan* package) and an *F*-test (via *permutest*; *vegan* package). For visualization, we created a non-metric dimensional scaling (NMDS) ordination (via *metamds*; *vegan* package; Bray Curtis dissimilarity). Using the universal insect trait tool (ITT; version 1.0) [37], we categorized all arthropod families into ecological guilds to analyze differences of the functional diversity between the insect communities of the two trap sites in addition to their taxonomic diversity.

Because the BIN concept has been challenged recently [38], we decided to compare the number of OTUs recovered with other clustering algorithms. BINs should not be considered synonymously of "species", but rather as a dynamic tool to presort the global DNA barcode database into MOTUs that taxonomists can further evaluate; BIN definitions might change on BOLD as more sequences are added to the database. Since the assignment of BINs in a dataset is affected by other sequences in BOLD that are not included in the dataset, we analyzed the sequences of our datasets using the "Cluster Sequences" option in BOLD. This way, the resulting OTUs are directly comparable to the results of other species delimitation algorithms. As a consequence, the number BINs found in our project on BOLD are slightly higher than in our analyses because the system assigns BINs to sequences between 300 and 499 bp if the BIN is already present in the database, whereas we limited analyses to sequences displaying a minimum length of 500 bp. In addition to RESL, we analyzed our data using the Assemble Species by Automatic Partitioning program (ASAP) [39] using the web interface, and we analyzed the same data using SpeciesIdentifier version 1.9 [40]. ASAP employs pairwise genetic distances for hierarchical clustering without using the information on intraspecific diversity, and SpeciesIdentifier is an algorithm that allows clustering sequences based on their pairwise genetic distances (p-distances). To visualize the outputs of the different clustering algorithms (RESL, ASAP, SpeciesIdentifier), we created accumulation curves (via *iNEXT; iNEXT* package) depicting the number of clusters obtained for each Malaise trap. Detailed specimen and sequence data are accessible in BOLD as two citable datasets (doi.org/10.5883/DS-GMTINDO1 and doi.org/10.5883/DS-GMTINDO2).

Results

Alpha-diversity assessments

We obtained 39,374 COI-sequences from Malaise trap 1, and 19,394 for Malaise trap 2 which led to the recovery of 6,177 and 5,206 BINs respectively. Together, we obtained a total of 9,212 BINs, with 2,171 being shared between traps. More than two-thirds (6,125) of all BINs were unique to BOLD, meaning that they were added for the first time with the upload of these

sequences. Of the 58,769 specimens that were successfully sequenced, only 961 automatically obtained a species-level identification, providing coverage for 231 species. The majority of sequences provided identification only to the family level (94%), and most of these were associated to families of insects that are reknown for being challenging to study and therefore highly underrepresented in databases (see Discussion). In this study, eight families of dark taxa were largely represented among our data, namely Cecidomyiidae (gall midges), Ceratopogonidae (biting midges), Chironomidae (non-biting midges), Phoridae (scuttle flies), Psychodidae (sand flies), Sciaridae (dark-winged fungus gnats), Platygastridae and Braconidae (parasitoid wasps). These eight families make up 70% of all specimen numbers, and 58% of all BINs. Fig 2, which presents the frequency of rare and common BINs among the merged dataset, shows that the majority of BINs (66%; 6,078 BINs) were represented by one or two specimens only.

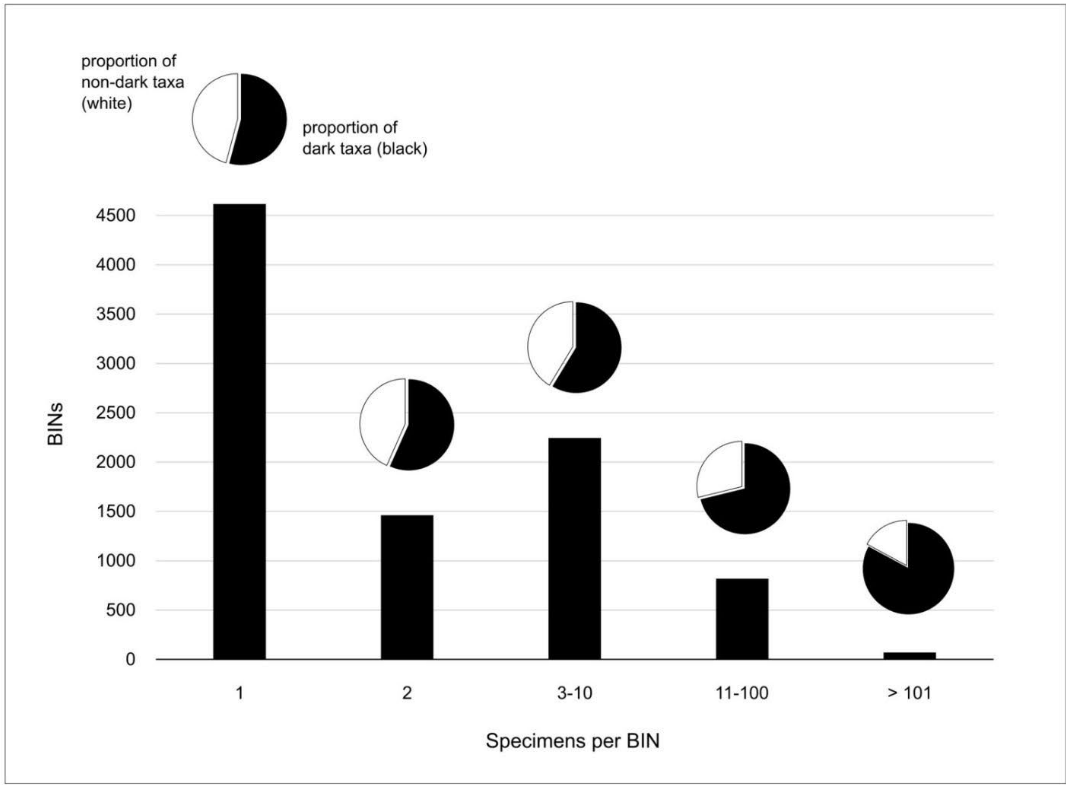


Fig 2. Frequency of BINs and proportion of dark taxa. The majority of BINs are rare and are represented by one (BIN frequency = 1) or two (BIN frequency = 2) specimens only. The pie charts represent the proportion of dark taxa among the BIN diversity (in black). These include members of Cecidomyiidae, Ceratopogonidae, Chironomidae, Phoridae, Psychodidae, Braconidae, and Platygastridae.

https://doi.org/10.1371/journal.pone.0290173.g002

Malaise trap 1. The BINs recovered from Malaise trap 1 provide coverage for 231 families in 21 arthropod orders. The top ten most diverse (from most to least diverse) families are Cecidomyiidae (Diptera; 1,858 BINs), Chironomidae (Diptera; 491 BINs), Ceratopogonidae (Diptera; 470 BINs); Phoridae (Diptera; 439 BINs), Platygastridae (Hymenoptera; 284 BINs); Sciaridae (Diptera; 239 BINs), Psychodidae (Diptera; 145 BINs), Formicidae (Hymenoptera; 125 BINs), Cicadellidae (Hemiptera; 111 BINs), and Braconidae (Hymenoptera; 105 BINs). In total, these families represent 70% of all recovered BINs for this Malaise trap. Chao1 analysis estimated that about 11,000 species may occur at this sampling site, and extrapolation to double the number of captured and processed specimens would have increased the number of recovered BINs to 8,531, which is an increase of 38% (Table 2). In the diversity profile, there is no overlap between the species richness obtained through the analysis of specimens and that estimated to occur at the trap sites (Hill number $q = 0$, Fig 3B).

Malaise trap 2. Although we processed substantially fewer specimens from Malaise trap 2, we obtained almost as many BINs (Table 2 and Fig 3A). The BINs from Malaise trap 2 provide coverage for 254 families in 24 arthropod orders. The ten most diverse families are (from most to least diverse): Cecidomyiidae (Diptera; 1,003 BINs), Phoridae (Diptera; 484 BINs), Platygastridae (Hymenoptera; 305 BINs), Sciaridae (Diptera; 220 BINs), Ceratopogonidae (Diptera; 189 BINs), Chironomidae (Diptera; 186 BINs), Cicadellidae (Hemiptera; 158 BINs), Braconidae (Hymenoptera; 152), Erebidae (Lepidoptera; 128 BINs), and Psychodidae (Diptera; 128 BINs). In total, these families represent 86% of all recovered BINs. Chao1 analysis revealed that about 10,000 species might occur at this trap site. Doubling the number of captured specimens would have increased the obtained BIN diversity to 7,481, an increase of 44% (Table 2). As for Malaise trap 1, there is no overlap between the number of empirical BINs obtained from our analyses and the species richness estimated to be present at the site (Fig 3C)

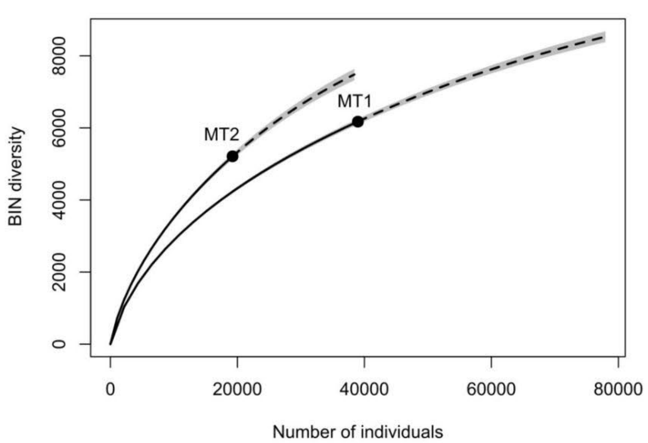
Beta-diversity analysis

Analysis revealed that 2,171 BINs are shared between both traps, and Chao1-shared estimates suggest that up to 4,281 (± 183) BINs are shared between both communities at the trap sites. PERMANOVA analysis of the sample contents uncovered that the arthropod communities from the Malaise traps are significantly distinct from one another (adonis2 $p = 0.001$) and that

Table 2. The number of clusters obtained from the COI sequence data of each Malaise trap when applying different clustering algorithms (RESL; ASAP; SpeciesIdentifier).

Algorithm	Output	Malaise trap 1	Malaise trap 2
RESL OTUs	Number of clusters (n)	6,283	5,253
	Sample coverage	0.918	0.843
	Extrapolation to 2n	8,699 ± 255	7,569 ± 266
ASAP OTUs	Number of clusters (n)	5,185	4,594
	Sample coverage	0.934	0.869
	Extrapolation to 2n	7106 ± 211	6501 ± 213
SpeciesIdentifier OTUs	Number of clusters (n)	5,967	5,054
	Sample coverage	0.921	0.851
	Extrapolation to 2n	8347 ± 274	7260 ± 252
BINs	Number of clusters (n)	6,177	5,206
	Number rare clusters	4,132	4,730
	Chao1 community estimator	11,280 ± 255	10,382 ± 268
	Sample coverage	0.919	0.844
	Extrapolation to 2n	8,531	7,481

https://doi.org/10.1371/journal.pone.0290173.t002



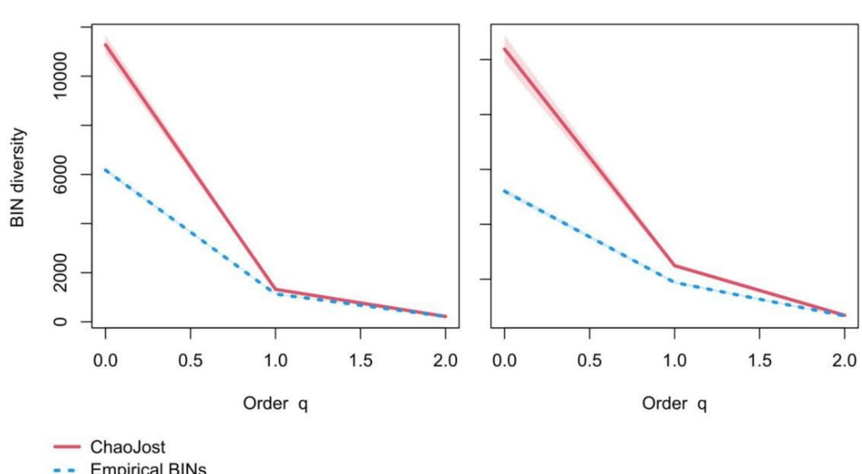


Fig 3. Accumulation curves of the BIN diversity recovered for each Malaise trap. Dotted lines represent extrapolated values (up to double the sampling effort), bold lines represent interpolated values. Shaded areas represent the 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0290173.g003

this significance is driven by location effects only (S1 Table). In the NMDS ordination, collection samples are clearly clustered based on Malaise trap (S1 Fig). Evaluating the data in more detail, we see that despite high species turnover, both traps depict similar compositions at the family level, which in turn has the same effect on the guild composition (Fig 4A and 4B).

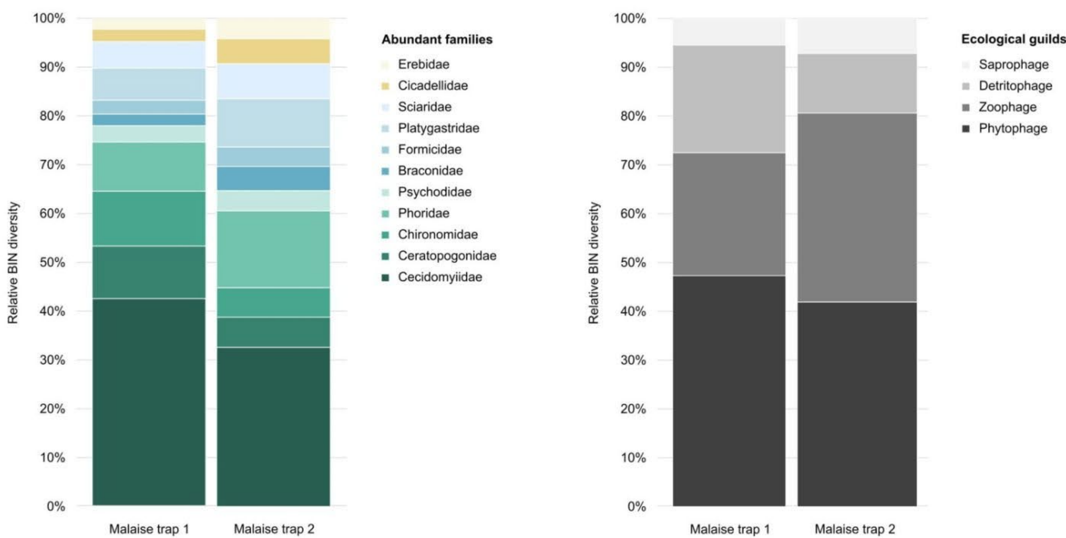


Fig 4. Relative BIN diversity. a. Relative BIN diversity across the top most abundant families in our Malaise traps. b. Relative BIN diversity across ecological guilds.

https://doi.org/10.1371/journal.pone.0290173.g004

COI clusters across methods

In total, 77,497 specimens of insects were processed, 52,362 from Malaise trap 1 and 25,135 from Malaise trap 2. Excluding all flagged sequences from analysis (and retaining only those with at least 500 bp) reduced our numbers to 39,374 and 19,394 COI-sequences for each trap respectively. For comparative analysis of cluster algorithms (in terms of cluster diversity), we reran the RESL-algorithm on these sequences which led to the recovery of 6,283 (MT1) and 5,253 (MT2) OTUs that are unique to our project (Table 2). SpeciesIdentifier (using the 3% threshold) suggested slightly fewer clusters than the RESL-algorithm, while ASAP (1st partition) calculated more conservative values, i.e., a much lower number of putative species (Table 2 and Fig 5).

Discussion

Overwhelming species richness despite drastic undersampling

All accumulation curves (Figs 3A and 5) and diversity profiles (Fig 3B and 3C) indicate that we have drastically undersampled both trap sites. This was expected for several reasons. First, our collection effort was limited in space and time, using two Malaise traps for three months only. Unlike temperate regions, generally speaking, no individual season in the tropics is highly unsuitable in terms of activity for all arthropod species [41], meaning that arthropods are present and mobile all year round [41, 42]. Therefore, sampling only three months provides a limited coverage of temporal species diversity. Second, while Malaise traps are very effective at collecting arthropods [26], we did not use any additional sampling method to incorporate the diverse canopy communities present in many tropical forests [29, 43, 44]. Our sampling

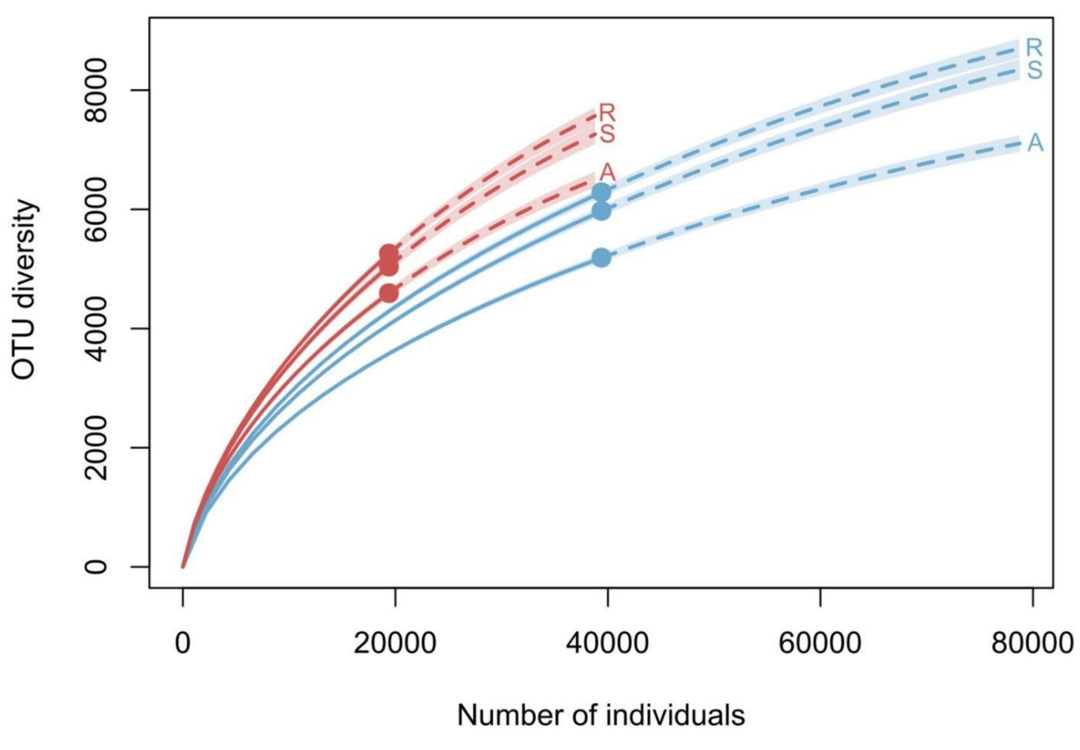


Fig 5. Accumulation curves of OTU diversity. Recovered with each clustering algorithm (R: RESL, A: ASAP, S: SpeciesIdentifier) for each Malaise trap. Dotted lines represent extrapolated values (up to double the sampling effort), bold lines represent interpolated values. Shaded areas represent the 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0290173.g005

techniques targeted arthropods that are found in the litter and understory habitats, whereas [43] and colleagues have demonstrated that the highest species richness is found in the forest canopy. Third, we did not process all collected individuals due to economic constraints. We had a total of eleven collection events per trap. Seven bulk samples that were collected with Malaise trap 1 were processed entirely; however, sequencing of all other samples was limited to 15 (1,475 specimens) plates per sample. Had we doubled our sampling effort, we would have recovered at least 38% and 44% more putative species for Malaise trap 1 and 2, respectively (Fig 3A). Sampling was slightly more comprehensive with Malaise trap 1, which is presumably due to the fact that more individuals were processed from these samples. Nevertheless, we

clearly only recovered a fraction of the actual diversity present at the sites: Chao1 calculations estimated much higher species numbers for each trap site, and we see no overlap between empirical and estimated BIN numbers for all three diversity orders (species richness q = 0; Shannon diversity q = 1; Simpson diversity q = 2).

Patchiness in arthropod diversity

Beta diversity assessments show that the communities from each trap site are significantly distinct and that this difference is driven by location effects only (all samples were dispersed homogeneously) (S1 Table and S1 Fig). Even after pooling all collection events together, we observed only 24% overlap in putative species between traps despite the close proximity (< 400 m). One can argue that due to the limited sequencing of sample contents, we are unknowingly comparing two very different subsets of actual similar communities, which was also mentioned by [45] as a factor contributing to the overestimation of beta-diversity [45]. However, we suggest that this is not the case because we recovered more than 80% sample coverage for each Malaise trap. Instead, we argue that we here witness arthropod diversity patchiness, as described by [46]. Forest floors are highly heterogeneous in the tropics over small spatial scales, resulting in high microhabitat richness in x-dimensions [46–48]. Nutrient availability, habitat heterogeneity, spatial variation of plant communities, degree of exposure to predators, and ecosystem disturbances are just some factors that define these microhabitats and their arthropod communities [47, 49, 50].

Prior studies on tropical rainforests have demonstrated that because the majority of insects are herbivores and host-specific, vegetation has a high impact on the prevailing arthropod species, which can account for up to 60% of insect variation [45, 51]. In our study, almost half of all recovered species per trap were phytophages (Fig 4B), meaning that differing vegetation at each trap site could be a driving factor behind the high species turnover [52]. Moreover, because Malaise traps capture insects that happen to fly through a very limited area, various factors such as trap location, orientation, height based on vegetation, light exposure and surrounding structures also have a direct effect on captured communities [53, 54]. In a recent study [54], examined the effects of Malaise trap spacing on species richness and composition, and found that community-similarity decreased among all major taxa with increasing trap-to-trap distances. Also, they found that 18 m between traps was the cut-off value where the number of shared species dropped significantly [54]. These results reinforce our assumption that we are in fact sampling and comparing two very different insect communities with one another.

Guild structures

Despite recovering a high species turnover between trap sites, community compositions at the family level were very similar (Fig 4A). In consequence, guild structure was also conserved (Fig 4B). However, we highly encourage further research to look into this because we analyzed guild structures only at the family-level. Although it is convenient to place entire families into guilds, it is also a source of error because species of the same families can cover a wide range of feeding behaviors [48]. However, assigning single species to guilds is a major challenge, especially in large-scale surveys. There is too little literature on the feeding activities of single species, and even then, different life stages of the same species can fall into different guild categories (e.g., parasitoid Hymenoptera), and for some taxa, feeding activities of adult species are completely unknown [48]. Also, only a small proportion of our sequences provided identification at the species level, meaning that we cannot apply feeding traits to species proxies. In this study, we did not conduct morphological identifications. Instead, all family-level

identifications were assigned automatically using the identification tool on BOLD. It is therefore important to note that accurate results are only guaranteed provided that high quality reference libraries are being used as a backbone, which include sequences of vouchers that have been accurately identified morphologically. Despite these sources of bias, we still believe that we can rely on these assigned identifications as we are only using them at the family-level, for which extensive information is available on BOLD.

For the family-level guild assignment, we used the Insect Trait Tool that was developed by [37]. Because this tool was developed for the Central European fauna, the extended trait information provided by the tool may not be accurate for tropical fauna. However, because we conducted only a broad guild analysis, we do not think that this is problematic in our study.

Dark taxa: Abundant, diverse, unknown

In our study, the majority of all BINs were rare, being represented by one or two specimens only (Fig 2). Although we did expect to capture a high proportion of singleton species, we recovered a surprisingly higher frequency of rare species than expected for large-scale tropical surveys, which is typically at about 32% [55, 56]. A closer look at the data revealed that the majority of these singletons are associated with "dark taxa", highly diverse groups of arthropods (mostly Diptera and Hymenoptera) for which little taxonomic or life-history information is available [6, 8]. In total, 70% (40,807) of all processed specimens and 58% (5,340 BINs) of all recovered putative species in this study are shared by eight dark taxa families only, namely Cecidomyiidae (gall midges), Ceratopogonidae (biting midges), Chironomidae (non-biting midges), Phoridae (scuttle flies), Psychodidae (sand flies), Sciaridae (dark-winged fungus gnats), Platygastridae and Braconidae (both parasitoid wasps).

As demonstrated in this study, dark taxa can be highly abundant and often make up the bulk of an insect sample not only in the tropics, but also in temperate regions [6, 57]. With this being a global phenomenon, the inability to associate these insects to species names or ecological functions is a large constraint to biodiversity research, conservation priority setting as well as understanding ecosystem functioning. One recent publication [20] highlighted that dark taxa are so abundant that they should be included in any holistic biodiversity assessment, but tackling them with traditional taxonomic techniques is too slow [20, 58]. Specimens of dark taxa are often small-bodied and cryptic diverse, so often (especially for Diptera), specimens need to be dissected and studied microscopically. Moreover, species identifications for these insects is often only possible with the use of multiple approaches in parallel to ensure accurate results. Integrative approaches that combine various methodologies are therefore becoming ever more important in making these groups tangible to science [4, 20, 59].

Since 2020, the third phase of the nationwide German Barcode of Life project (GBOL III: Dark Taxa; https://bolgermany.de/home/gbol3/de/projekte/) is dedicated to tackling difficult groups of taxa and training a new generation of taxonomists. In this initiative, integrative methods are being used in order to speed up the identification of dark taxa and making them more tangible to science. to do this, researchers are using (among others) a reverse and integrative taxonomical approach to effectively target and study their groups of interest. This consists of first applying molecular methods (including MinION technologies) to rapidly distinguish sequences clusters among thousands of preselected specimens, then applying morphological methods to target specimens of specific clusters for species identification. This technique drastically reduces the workload because time-consuming specimen processing and morphological analysis is drastically reduced. However, this approach it still time consuming, because it still requires the processing of thousands of individuals, as in our case [60]. One technology that is currently expediting biomonitoring surveys is metabarcoding, which allows

the analysis of entire bulk samples in one sequencing run [14, 17]. However, this method only provides information on community compositions and not on abundance data, nor it the link between sequence and specimen conserved [60, 61]. This makes it especially difficult to study dark taxa because they consist of many species that are not yet described, so these remain undescribed because specimens cannot be easily pinpointed [60].

Just recently, new technological developments have emerged which can help accelerate biomonitoring studies by speeding up the greatest bottleneck of ecological research–sample sorting. Bulk samples of arthropods often contain hundreds to thousands of specimens, that need to be sorted before conducting species-level analyses. In their study, [60] present a compact insect sorting robot which has the ability to recognize and sort insect specimens based on overview images of bulk samples. Especially interesting is the fact this robot, the DiversityScanner, is able to process very small specimens (<3 mm) [60]. Specimens are automatically selected by the scanner, imaged, assigned to a class or family, then moved to a microplate. Another study, [62], propose a workflow that combines HotSHOT with MinION technologies to conduct fast and accurate species-level sorting of ecological samples [62]. With a modest amount of equipment, manpower, and training, the authors were able to conduct species-level sorting within hours, which came down to 2.5 minutes per specimen. Of course, species identification can only be provided if identified sequences are present in databases, however, coupling this approach with the aforementioned reverse workflow that is applied in the GBOL III project could drastically expedite the work for taxonomists. Because no taxonomic expertise is necessary for the laboratory produces, taxonomists can be first brought on board to analyze vouchers after cluster analysis.

Employing DNA-based delimitation methods: Working with species proxies

BOLD not only provides a variety of analytical- and visualization techniques, its interface is also very user-friendly, making it easy for all researchers (even with little or no bioinformatic knowledge) to use [32]. Due to this, BOLD is commonly used in DNA barcoding research, so consequently, its integrated RESL algorithm and BIN system is also commonly used for sequence data clustering. For our analyses, we used BIN-counts as a proxy for species diversity, as has been done in various studies [6, 23, 63–66]. However, there are varying opinions regarding using BINs for species delimitation [39, 46], especially when assuming that BIN numbers are equal to species numbers in a 1:1 ratio. Therefore, as [67] recommended, we analyzed our sequence data with several species delimitation methods that apply different algorithms to compare the number of clusters recovered with each method. We used SpeciesIdentifier for objective clustering using a preset threshold (3%) for comparative purposes and to increase confidence regarding the relative extent of diversity in our traps. We recovered slightly fewer clusters than with the RESL-algorithm from BOLD [Malaise Trap 1: 5,967 versus 6,283 OTUs; Malaise Trap 2: 5,054 versus 5,253 OTUs]. With ASAP, hierarchical clustering was done using pairwise genetic distances of sequences. The program builds numerous partitions ranked by scores, with the best ones provided in the output to be used for analysis. With ASAP, we obtained much more conservative cluster counts than with the RESL (and SpeciesIdentifier) algorithm, especially for Malaise trap 1 (Table 2 and Fig 5). Analysis across methods displayed similar trends in regard to sample coverage, depicting that the sample contents of Malaise trap 1 were much better sampled than of Malaise trap 2 (Table 2).

Conclusion

Here, processing only a fraction of bulk samples collected during merely three months of Malaise trap sampling recovered more than 9,000 putative species and high species turnover

among two very close sites. Despite processing more than 77,000 specimens, community analysis suggests that we strongly undersampled both collection sites. Community compositions at the family level were conserved between traps, revealing similar ecological guild functions. The majority of specimens collected and processed belong to the so-called dark taxa, for which little taxonomic and life history information is available. Comprehensive specimen sampling, KI-powered sample processing, and highest throughput sequencing coupled with trait analysis will be crucial to address this knowledge gap, for which the technological is being created now [38, 68–70].

Supporting information

S1 Fig. Non-metric dimensional scaling (NMDS) of the community compositions. NMDS plot of the insect community compositions within each collection sample. Ellipses are 95% confidence intervals of centroids for each Malaise trap.
(PDF)

S1 Table. Statistical analysis of the community compositions.
(PDF)

Acknowledgments

We would like to thank the Ontario Genomics Institute conducting sequence analysis and informatics.

Author Contributions

Conceptualization: Stefan Schmidt, Bruno Cancian de Araujo, Thomas von Rintelen, Olga Schmidt, Hasmiandy Hamid, Raden Pramesa Narakusumo, Michael Balke.

Data curation: Bruno Cancian de Araujo.

Formal analysis: Caroline Chimeno, Olga Schmidt.

Funding acquisition: Michael Balke.

Investigation: Stefan Schmidt, Bruno Cancian de Araujo, Thomas von Rintelen, Olga Schmidt.

Methodology: Stefan Schmidt, Kate Perez, Thomas von Rintelen, Olga Schmidt, Michael Balke.

Project administration: Michael Balke.

Software: Caroline Chimeno.

Visualization: Caroline Chimeno.

Writing – original draft: Caroline Chimeno.

Writing – review & editing: Caroline Chimeno, Stefan Schmidt, Hasmiandy Hamid, Raden Pramesa Narakusumo, Michael Balke.

References

1. Borkent A, Brown BV. How to inventory tropical flies (Diptera)—One of the megadiverse orders of insects. Zootaxa. 2015 Apr 28; 3949(3):301–22. https://doi.org/10.11646/zootaxa.3949.3.1 PMID: 25947810

2. 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
3. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, et al. The biodiversity of species and their rates of extinction, distribution, and protection. Science. 2014 May 30; 344(6187):1246752. https://doi.org/10.1126/science.1246752 PMID: 24876501
4. Riedel A, Sagata K, Suhardjono YR, Tänzler R, Balke M. Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. Front Zoology. 2013 Mar 27; 10(1):15. https://doi.org/10.1186/1742-9994-10-15 PMID: 23537182
5. 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
6. Chimeno C, Hausmann A, Schmidt S, Raupach MJ, Doczkal D, Baranov V, et al. Peering into the darkness: DNA barcoding reveals surprisingly high diversity of unknown species of Diptera (Insecta) in Germany. Insects. 2022 Jan; 13(1):82. https://doi.org/10.3390/insects13010082 PMID: 35055925
7. Page RDM. DNA barcoding and taxonomy: dark taxa and dark texts. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016 Sep 5; 371(1702):20150334. https://doi.org/10.1098/rstb.2015.0334 PMID: 27481786
8. Hausmann A, Krogmann L, Peters R, Rduch V, Schmidt S. GBOL III: DARK TAXA. iBOL Barcode Bulletin. 2020 Jul 10; 10.
9. Dufour DL. Insects as Food: A Case Study from the Northwest Amazon. American Anthropologist. 1987; 89(2):383–97.
10. Kremen C, Colwell RK, Erwin TL, Murphy DD, Noss RF, Sanjayan MA. Terrestrial Arthropod Assemblages: Their Use in Conservation Planning. Conservation Biology. 1993; 7(4):796–808.
11. Longino JT. Tropical Insect Diversity -How to Sample It. In: Claro KD, Oliveira PS, Rico-Gray V, editors. Tropical Biology and Conservation Management—Volume XI: Case Studies. EOLSS Publications; 2009. p.59–84.
12. Farwig N, Bendix J, Beck E. Introduction to the Special Issue "Functional monitoring in megadiverse tropical ecosystems." Ecological Indicators. 2017 Dec 1; 83:524–6.
13. Sattler T, Duelli P, Obrist MK, Arlettaz R, Moretti M. Response of arthropod species richness and functional groups to urban habitat structure and management. Landscape Ecol. 2010 Jul 1; 25(6):941–54.
14. Chimeno C, Hübner J, Seifert L, Morinière J, Bozicevic V, Hausmann A, et al. Depicting environmental gradients from Malaise trap samples: Is ethanol-based DNA metabarcoding enough? Insect Conserv Divers. 2023; 16(1):47–64.
15. Aylagas E, Borja Á, Irigoien X, Rodríguez-Ezpeleta N. Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. Front Mar Sci [Internet]. 2016 [cited 2021 Apr 13]. Available from: https://www.frontiersin.org/articles/10.3389/fmars.2016.00096/full
16. Jinbo U, Kato T, Ito M. Current progress in DNA barcoding and future implications for entomology. Entomological Science. 2011; 14(2):107–24.
17. 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
18. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc R Soc Lond B. 2003 Feb 7; 270(1512):313–21. https://doi.org/10.1098/rspb.2002.2218 PMID: 12614582
19. Baloğlu B, Clews E, Meier R. NGS barcoding reveals high resistance of a hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs. Frontiers in Zoology. 2018 Aug 14; 15(1):31. https://doi.org/10.1186/s12983-018-0276-7 PMID: 30127839
20. Hartop E, Srivathsan A, Ronquist F, Meier R. Large-scale Integrative Taxonomy (LIT): resolving the data conundrum for dark taxa. bioRxiv. 2021 May 9;2021.04.13.439467.
21. Hausmann A, Godfray HCJ, Huemer P, Mutanen M, Rougerie R, Nieukerken EJ van, et al. Genetic Patterns in European Geometrid Moths Revealed by the Barcode Index Number (BIN) System. PLOS ONE. 2013 Dec 17; 8(12):e84518. https://doi.org/10.1371/journal.pone.0084518 PMID: 24358363
22. Yeo D, Puniamoorthy J, Ngiam RWJ, Meier R. Towards holomorphology in entomology: rapid and cost-effective adult-larva matching using NGS barcodes: Life-history stage matching with NGS barcodes. Syst Entomol. 2018 Oct; 43(4):678–91.
23. deWaard JR, Ratnasingham S, Zakharov EV, Borisenko AV, Steinke D, Telfer AC, et al. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. Sci Data. 2019 Dec 6; 6(1):308.

24. Geiger MF, Astrin JJ, Borsch T, Burkhardt U, Grobe P, Hand R, et al. How to tackle the molecular species inventory for an industrialized nation-lessons from the first phase of the German Barcode of Life initiative GBOL (2012–2015). Genome. 2016 Sep; 59(9):661–70. https://doi.org/10.1139/gen-2015-0185 PMID: 27314158
25. Karlsson D, Hartop E, Forshage M, Jaschhof M, Ronquist F. The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. Biodiversity Data Journal. 2020 Jan 21; 8: e47255. https://doi.org/10.3897/BDJ.8.e47255 PMID: 32015667
26. Matthews RW, Matthews JR. The Malaise Trap: Its Utility and Potential for Sampling Insect Populations. 2017; 4(4):7.
27. Cancian de Araujo B, Schmidt S, Schmidt O, Rintelen T von, Rintelen K von, Floren A, et al. DNA barcoding data release for Coleoptera from the Gunung Halimun canopy fogging workpackage of the Indonesian Biodiversity Information System (IndoBioSys) project. Biodiversity Data Journal. 2019 Jan 15; 7: e31432. https://doi.org/10.3897/BDJ.7.e31432 PMID: 30686928
28. Cancian de Araujo B, Schmidt S, Schmidt O, Rintelen T von, Ubaidillah R, Balke M. The Mt Halimun-Salak Malaise Trap project—releasing the most species rich DNA Barcode library for Indonesia. Biodiversity Data Journal. 2018 Dec 19; 6:e29927.
29. 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
30. 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
31. Ivanova NV, Dewaard JR, Hebert PDN. An inexpensive, automation-friendly protocol for recovering high-quality DNA. Molecular Ecology Notes. 2006; 6(4):998–1002.
32. Ratnasingham S, Hebert PDN. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes. 2007 Jan 24; 7(3):355–64.
33. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Journal. 2012;
34. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package [Internet]. 2020 [cited 2021 Oct 6]. Available from: https://CRAN.R-project.org/package=vegan
35. Hsieh TC, Chao KHM and A. iNEXT: Interpolation and Extrapolation for Species Diversity [Internet]. 2020 [cited 2021 Oct 6]. Available from: https://CRAN.R-project.org/package=iNEXT
36. Chao A, Ma KH, Chiu TCH and CH. SpadeR: Species-Richness Prediction and Diversity Estimation with R [Internet]. 2016 [cited 2021 Oct 8]. Available from: https://CRAN.R-project.org/package=SpadeR
37. Hörren T, Sorg M, Hallmann CA, Zizka VMA, Ssymank A, Noll NW, et al. A universal insect trait tool (ITT, v1.0) for statistical analysis and evaluation of biodiversity research data [Internet]. Ecology; 2022 Jan [cited 2022 Sep 7]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.01.25.477751
38. Meier R, Blaimer BB, Buenaventura E, Hartop E, Rintelen T, Srivathsan A, et al. A re-analysis of the data in Sharkey et al.'s (2021) minimalist revision reveals that BINs do not deserve names, but BOLD Systems needs a stronger commitment to open science. Cladistics. 2022 Apr; 38(2):264–75.
39. Puillandre N, Brouillet S, Achaz G. ASAP: assemble species by automatic partitioning. Mol Ecol Resour. 2021 Feb; 21(2):609–20. https://doi.org/10.1111/1755-0998.13281 PMID: 33058550
40. Meier R, Shiyang K, Vaidya G, Ng PKL. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst Biol. 2006 Oct; 55(5):715–28. https://doi.org/10.1080/10635150600969864 PMID: 17060194
41. Denlinger DL. Dormancy in Tropical Insects. Denlinger, D L (1986) Dormancy in Tropical Insects Annual Review of Entomology, 31(1), 239–264 https://doi.org/10.1146/annurev.en.31.010186.001323 1986;31(1):239–64. PMID: 3510585
42. Kishimoto-Yamada K, Itioka T. How much have we learned about seasonality in tropical insect abundance since Wolda (1988)? Entomological Science. 2015; 18(4):407–19.
43. Basset Y, Cizek L, Cuénoud P, Didham RK, Novotny V, Ødegaard F, et al. Arthropod Distribution in a Tropical Rainforest: Tackling a Four Dimensional Puzzle. PLOS ONE. 2015 Dec 3; 10(12):e0144110.
44. Ozanne CMP, Anhuf D, Boulter SL, Keller M, Kitching RL, Körner C, et al. Biodiversity Meets the Atmosphere: A Global View of Forest Canopies. Science. 2003 Jul 11; 301(5630):183–6. https://doi.org/10.1126/science.1084507 PMID: 12855799
45. Novotny V, Drozd P, Miller SE, Kulfan M, Janda M, Basset Y, et al. Why Are There So Many Species of Herbivorous Insects in Tropical Rainforests? Science. 2006 Aug 25; 313(5790):1115–8. https://doi.org/10.1126/science.1129237 PMID: 16840659

46. Milton Y, Kaspari M. Bottom-up and top-down regulation of decomposition in a tropical forest. Oecologia. 2007 Aug 1; 153(1):163–72. https://doi.org/10.1007/s00442-007-0710-6 PMID: 17375326
47. Sayer EJ, Sutcliffe LME, Ross RIC, Tanner EVJ. Arthropod Abundance and Diversity in a Lowland Tropical Forest Floor in Panama: The Role of Habitat Space vs. Nutrient Concentrations. Biotropica. 2010; 42(2):194–200.
48. Stork NE. Guild structure of arthropods from Bornean rain forest trees. Ecol Entomol. 1987 Feb; 12 (1):69–80.
49. Farfán-Beltrán ME, Chávez-Pesqueira M, Hernández-Cumplido J, Cano-Santana Z. A quick evaluation of ecological restoration based on arthropod communities and trophic guilds in an urban ecological preserve in Mexico City. Rev Chil de Hist Nat. 2022 Dec; 95(1):4.
50. Krell FT. Parataxonomy vs. taxonomy in biodiversity studies—Pitfalls and applicability of "morphospecies" sorting. Biodiversity and Conservation. 2004 Feb 4; 13(4):795–812.
51. Lewinsohn TM, Roslin T. Four ways towards tropical herbivore megadiversity. Ecol Letters. 2008 Apr; 11(4):398–416. https://doi.org/10.1111/j.1461-0248.2008.01155.x PMID: 18248447
52. Stuntz S, Ziegler C, Simon U, Zotz G. Diversity and structure of the arthropod fauna within three canopy epiphyte species in central Panama. J Trop Ecol. 2002 Mar; 18(2):161–76.
53. Chan-Canché R, Ballina-Gómez H, Leirana-Alcocer J, Bordera S, González-Moreno A. Sampling of parasitoid Hymenoptera: influence of the height on the ground. Journal of Hymenoptera Research. 2020 Aug 31; 78:19–31.
54. Steinke D, Braukmann TW, Manerus L, Woodhouse A, Elbrecht V. Effects of Malaise trap spacing on species richness and composition of terrestrial arthropod bulk samples. Metabarcoding and Metagenomics. 2021 Apr 9; 5:e59201.
55. Coddington JA, Agnarsson I, Miller JA, Kuntner M, Hormiga G. Undersampling Bias: The Null Hypothesis for Singleton Species in Tropical Arthropod Surveys. Journal of Animal Ecology. 2009; 78(3):573–84. https://doi.org/10.1111/j.1365-2656.2009.01525.x PMID: 19245379
56. Lim GS, Balke M, Meier R. Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods. Systematic Biology. 2012 Jan 1; 61(1):165–9. https://doi.org/10.1093/sysbio/syr030 PMID: 21482553
57. Srivathsan A, Ang Y, Heraty JM, Hwang WS, Jusoh WFA, Narayanan S, et al. Convergence of dominance and neglect in flying insect diversity. Nature Ecology & Evolution. 2023; 7: 1012–1021. https://doi.org/10.1038/s41559-023-02066-0 PMID: 37202502
58. Puillandre N, Modica MV, Zhang Y, Sirovich L, Boisselier MC, Cruaud C, et al. Large-scale species delimitation method for hyperdiverse groups. Molecular Ecology. 2012; 21(11):2671–91. https://doi.org/10.1111/j.1365-294X.2012.05559.x PMID: 22494453
59. Chimeno C, Rulik B, Manfrin A, Kalinkat G, Hölker F, Baranov V. Facing the infinity: tackling large samples of challenging Chironomidae (Diptera) with an integrative approach. PeerJ. 2023 May 22; 11: e15336. https://doi.org/10.7717/peerj.15336 PMID: 37250705
60. Wührl L, Pylatiuk C, Giersch M, Lapp F, von Rintelen T, Balke M, et al. DiversityScanner: Robotic handling of small invertebrates with machine learning methods. Molecular Ecology Resources. 2022; 22 (4):1626–38. https://doi.org/10.1111/1755-0998.13567 PMID: 34863029
61. Creedy TJ, Ng WS, Vogler AP. Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. Ecology and Evolution. 2019; 9(6):3105–16. https://doi.org/10.1002/ece3.4839 PMID: 30962884
62. Vasilita C, Feng V, Hansen AK, Hartop E, Srivathsan A, Struijk R, et al. Express barcoding with NextGenPCR and MinION for species-level sorting of ecological samples [Internet]. bioRxiv; 2023 [cited 2023 Jun 13]. p. 2023.04.27.538648. Available from: https://www.biorxiv.org/content/10.1101/2023.04.27.538648v1
63. Hebert PDN, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, et al. Counting animal species with DNA barcodes: Canadian insects. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016 Sep 5; 371(1702):20150333. https://doi.org/10.1098/rstb.2015.0333 PMID: 27481785
64. Morinière J, Balke M, Doczkal D, Geiger MF, Hardulak LA, Haszprunar G, et al. A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring. Molecular Ecology Resources. 2019; 19(4):900–28. https://doi.org/10.1111/1755-0998.13022 PMID: 30977972
65. Pentinsaari M, Blagoev G, Hogg I, Levesque-Beaudin V, Perez K, Sobel C, et al. A DNA Barcoding Survey of an Arctic Arthropod Community: Implications for Future Monitoring. Insects. 2020 Jan 9; 11. https://doi.org/10.3390/insects11010046 PMID: 31936447

66. Seymour M, Roslin T, deWaard J, Perez K, D'Souza M, Ratnasingham S, et al. Arthropod beta-diversity is spatially and temporally structured by latitude [Internet]. In Review; 2022 Oct [cited 2023 Jan 25]. Available from: https://www.researchsquare.com/article/rs-2180975/v1

67. Carstens BC, Pelletier TA, Reid NM, Satler JD. How to fail at species delimitation. Molecular Ecology. 2013; 22(17):4369–83. https://doi.org/10.1111/mec.12413 PMID: 23855767

68. Fujisawa T, Noguerales V, Meramveliotakis E, Papadopoulou A, Vogler AP. Image-based taxonomic classification of bulk insect biodiversity samples using deep learning and domain adaptation. Systematic Entomology. 2023;

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

70. Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, et al. ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. BMC Biol. 2021 Sep 29; 19(1):217. https://doi.org/10.1186/s12915-021-01141-x PMID: 34587965

Publication II:

Gallmücken in Bayern: DNA Barcoding vermittelt neue Einblicke in die Mega-Vielfalt bislang (zu oft) ignorierter Mikrodipteren

Published in Nachrichtenblatt Bayerischen Entomologen. © 2020 by Münchner Entomologische

Gesellschaft

Chimeno, C., Doczkal, D., Haszprunar, G., Hausmann, A., Jaschhof, M., Kotrba, M., Perez, K., Raupach, M. & Schmidt, S. (2021). Gallmücken in Bayern: DNA Barcoding vermittelt neue Einblicke in die Mega-Vielfalt bislang (zu oft) ignorierter Mikrodipteren. *Nachrichtenblatt Bayerischen Entomologen*, 70 (3/4).
Not peer-reviewed.

Gallmücken in Bayern: DNA Barcoding vermittelt neue Einblicke in die Mega-Vielfalt bislang (zu oft) ignorierter Mikrodipteren

(Diptera, Cecidomyiidae)

Caroline CHIMENO, Dieter DOCZKAL, Gerhard HASZPRUNAR, Axel HAUSMANN, Mathias JASCHHOF, Marion KOTRBA, Kate PEREZ, Michael RAUPACH & Stefan SCHMIDT

Abstract

531 genetic clusters (BINs) of gall midges were retrieved in a single Malaise trap that was set up in 2017 within the inner-city premises of the Bavarian State Collection of Zoology in Munich, Germany. Although the sample obtained from this trap is very limited in time and space, the number of detected BINs correspond to 63,5% of the number of Cecidomyiidae species known in Germany up to date. Based on data collected from our utterly limited sample, and taking the variety of yet unexplored habitats into account, we conclude that Germany's fauna of gall midges must comprise considerably more species than previously expected.

Einleitung

„Wie viele Tierarten gibt es?" ist eine Frage, die Biologen schon immer beschäftigt hat und gleichzeitig eine, die stets nur mit mehr oder weniger vagen Schätzungen beantwortet wurde. Obwohl es immer wieder neue Abschätzungen gab und gibt, haben diese bislang wenig Gewissheiten gebracht. Das liegt unter anderem daran, dass vor allem taxonomisch gut bearbeitete Gruppen untersucht wurden, während weniger bekannte, oftmals kryptisch-vielfältige Gruppen unberücksichtigt blieben (z. B. ÖDEGAARD 2000, MORA et al. 2011, CICCONARDI et al. 2013, CALEY et al. 2014, WUEHRL et al., 2021). Aktuellere Studien, welche die Artenzahl anhand von Sequenzvariationen in der COI-DNA Barcode Region des mitochondrialen Genoms erfassen, sind vom taxonomischen Bearbeitungsstand weitgehend unabhängig. Dabei fasst ein Algorithmus ähnliche Sequenzen zu sogenannten „molecular Operational Taxonomic Units" (mOTUs) zusammen. Diesen werden im internationalen Barcode of Life Data System (BOLD) sogenannte „Barcode Index Numbers" (BINs) zugeordnet, welche biologische Arten verhältnismäßig gut abbilden können (RATNASINGHAM & HEBERT, 2013, SCHMIDT et al. 2015, HAUSMANN et al. 2013). Eine entsprechende Studie von HEBERT et al. (2016) kommt für die Insektenfauna zu dem Schluss, dass es global weit über 10 Millionen Insektenarten geben könnte. Diese Zahl überschreitet die kurz zuvor veröffentlichte, auf einer Extrapolation aus Käfer-Pflanzen-Interaktionen beruhenden Schätzung von 5,5 Millionen Arten bei weitem (STORK et al. 2015). Frühere Schätzungen lagen bei 5–10 Millionen (MORA et al. 2011) bis hin zu 30 Millionen Arten (ERWIN 1982).

Innerhalb der Insekten gehören die Zweiflügler (Diptera; Fliegen und Mücken) und Hautflügler (Hymenoptera; Ameisen, Bienen und Wespen) zu den vier artenreichsten Ordnungen, zusammen mit den Coleoptera und Lepidoptera. Beide Ordnungen enthalten Familien, die trotz eines vermuteten extrem hohen Artenreichtums in der taxonomischen Forschungslandschaft bislang nicht gebührend berücksichtigt wurden. Es ist heute offenkundig, dass die weltweiten Schätzungen von Artenzahlen vor allem deshalb so ungenau und widersprüchlich ausfallen, weil man über das unüberschaubare Heer kleiner Fliegen und Wespen bislang nur sehr vage Vorstellungen hat. Einige der in diesem Sinne vernachlässigten Familien – auch bekannt als „Dark Taxa" – stehen jetzt im Mittelpunkt des neuen Großprojekts "German Barcode of Life III: Dark Taxa" (GBOL III), das im Juli 2020 begann. Ziel dieses Projekts ist es, das taxonomische und molekulare Wissen für diese Gruppen mittels eines integrativ-taxonomischen Ansatzes erheblich zu erweitern, um die erfassten Arten weiterführender Forschung (z.B. Nutzpflanzenschutz) zugänglich zu machen. Eine der Zielgruppen von GBOL III sind die Gallmücken (Cecidomyiidae), nach Meinung vieler Dipterologen *das* „dunkle Taxon" schlechthin. Die Vertreter dieser Familie gehören mit einer Körperlänge von etwa 1,5–2,5 mm zu den kleinsten Dipteren („Mikrodipteren"). Die Imagines vieler Arten sind sogar kleiner als 1 mm. HEBERT et al. (2016) kamen

zu dem Schluss, dass allein in Kanada von etwa 16.000 Gallmücken-Arten auszugehen ist, was eine Verzehnfachung bisheriger Erwartungen bedeutet und nahelegt, dass es weltweit 1,8 Millionen Gallmücken-Spezies geben könnte. Somit wäre diese Gruppe die mit Abstand artenreichste Familie im Tierreich (Hebert et al. 2016). Für Deutschland wurden bisher ca. 840 Arten der Familie Cecidomyiidae identifiziert (Meyer & Jaschhof 1999). Eine von Morinière et al. (2019) publizierte DNA Barcode Sammlung – ein Ergebnis von GBOL I und II – umfasst 927 BINs, von denen nur 44 ein Artname zugeordnet werden konnte.

Als Teil von GBOL III hat die Zoologische Staatssammlung München (SNSB-ZSM) den Projektteil „Gallmücken“ übernommen. Schon im Vorfeld der Studie war im Rahmen des internationalen „Global Malaise Trap Program“ ein Jahr lang (2017) eine Malaisefalle auf dem Gelände des Instituts im Münchener Stadtgebiet betrieben worden.

Materialien und Methoden

Die Feldarbeit

Im Jahr 2017 wurde eine Malaisefalle auf dem Gelände der Zoologischen Staatssammlung aufgestellt (48.165° N, 11.485° O; 519 m ü.N.N.). Der im Münchener Stadtgebiet gelegene Fallenstandort liegt in einem Wohnviertel mit kleinräumigen Gartenanlagen, während das Gelände der ZSM auf ca. 25.000 qm Fläche von Wiesen und einigen Gehölzen bedeckt ist. Auf dem Gelände wurden vom Bund Naturschutz in Bayern e.V. seit mehreren Jahren Pflegemaßnahmen zur Erhöhung der Biodiversität durchgeführt. Die Malaisefalle wurde von April bis Dezember betrieben und die Fangflasche zweiwöchentlich ausgewechselt.

Jede zweite Sammelprobe wurde zur nachfolgenden Bearbeitung ausgewählt, mit den folgenden Leerungsterminen: 10. April, 08. Mai, 05. Juni, 03. Juli, 31. Juli, 28. August, 25. September, 23. Oktober, 20. November und 29. Dezember 2017.

Probenbearbeitung

Alle Proben wurden zur Bearbeitung an das Centre for Biodiversity Genomics (CBG) in Guelph, Kanada, geschickt. Dort wurden nach morphologischen Kriterien insgesamt 32.592 Arthropoden einzeln aus den Sammelproben selektiert. Da die Arthropoden im Ganzen für die Sequenzierung verwendet wurden, wurden nicht-invasive Voucher-Recovery-Protokolle für die DNA-Extraktion verwendet, d.h. die äußere Insektenhülle blieb intakt für morphologische Untersuchungen. Die Belegexemplare wurden im Nachhinein zur anschließenden Dokumentation und Archivierung an die ZSM zurückgeschickt. Das CBG verarbeitete die Proben unter Verwendung standardisierter Protokolle zur DNA-Extraktion, PCR-Amplifikation und Sanger-Sequenzierung (http://ccdb.ca/resources/). Alle Sequenzdaten und Trace-Dateien (Sequenz-Chromatogramme) sind in der Barcode of Life Database (BOLD; http://www.boldsystems.org) im Container-Projekt GMTBZ in den Einzel-Projekten GMGMO bis GMGMX hinterlegt. Die in dieser Arbeit analysierten Daten der Familie Cecidomyiidae sind im Datensatz DS-CECIDZSM (doi: dx.doi.org/10.5883/DS-CECIDZSM) öffentlich zugänglich. Zurückgesendete Proben wurden für neue BINs z. T. fotografiert und dokumentiert.

Datenanalyse

Zur Analyse der Daten wurden die in BOLD verfügbaren Analyse-Tools verwendet. Die mittlere und maximale intraspezifische Variation und die minimale genetische Distanz zur nächsten Art wurden unter Verwendung des Kimura-2-Parameter-Modells berechnet (Kimura 1980, Puillandre et al. 2012). Anhand der Distanzen wurden für die Sequenzen mittels RESL-Algorithmus mOTUs ermittelt und BINs zugewiesen (Ratnasingham & Hebert 2013).

Resultate

Von 32.593 Sequenzen wurden 26.534 (81,4%) der Sequenzen BINs zugeteilt, die insgesamt 3.870 BIN-Clustern zuzuordnen sind. Voraussetzung für die Zuordnung einer Sequenz zu einer bereits existierenden BIN ist eine Mindestlänge von 300 bp. Für die Generierung einer neuen BIN ist eine Sequenzlänge von mindestens 500 bp erforderlich. Die 26.534 Sequenzen erstrecken sich über drei Arthropoden-Klassen und decken 21 Ordnungen ab, von denen Diptera (14.850 Exemplare) und Hymenoptera (7.859 Exemplare) bei weitem den Großteil ausmachen. Von den insgesamt 2.643 Sequenzen von Vertretern der Familie Cecidomyiidae konnten 325 Sequenzen (12,3%) insgesamt 30 bekannten Arten zugeordnet werden, da diese in BOLD bereits erfasst waren. Die restlichen 2.318 Sequenzen konnten bisher keiner bekannten Art zugeordnet werden **(Tab. 1).** Dabei ließen sich 2.157 Sequenzen nur bis zur Familie, und 161 nur bis zur Gattung bestimmen. Den 2.643 Cecidomyiidae-Sequenzen wurden insgesamt 531 BINs zugewiesen. In vier Fällen ist eine bislang vermeintlich gute „Art" jeweils durch zwei BINs repräsentiert: *Dasineura spadicea* Rübsaamen, 1917; *Campylomyza flavipes* Meigen, 1818; *Lestodiplosis juniperina* Felt, 1907; *Peromyia caricis* Kieffer 1901. 175 der insgesamt 531 BINs (33%) haben durch dieses Projekt erstmalig Eingang in BOLD gefunden; diesen neuen BINs wurden insgesamt 453 Individuen zugeordnet.

Tabelle 1: Übersicht der 2.643 Sequenzen im Cecidomyiidae-Datensatz: Anzahl der Sequenzen für die entsprechende Gattung sowie Anzahl der Sequenzen mit Artname und prozentualem Anteil (in Klammer) und Auflistung der beschriebenen Arten.

Gattung/Art Anzahlen:	Sequenzen	Sequenzen mit Artname (%)	BINs	Arten
Asphondylia	11	0 (0%)	1	0
Asteromyia	62	0 (0%)	2	0
Campylomyza *Campylomyza dilatata* (Felt, 1907) *Campylomyza flavipes* (Meigen, 1818)	220	220 (100%)	3	2
*CecidIntGen**	24	0 (0%)	2	0
Clinodiplosis *Clinodiplosis cilicrus* (Kieffer, 1889)	4	4 (100%)	1	1
Contarinia *Contarinia asclepiadis* (Giraud, 1863)	6	2 (33,3%)	3	1
Dasineura *Dasineura leguminicola* (Lintner, 1879) *Dasineura spadicea* (Rübsaamen, 1917)	15	11 (73.3%)	6	2
Feltiella	7	0 (0%)	2	0
Leptosyna *Leptosyna nervosa* (Winnertz, 1852)	2	2 (100%)	1	1
Lestodiplosis *Lestodiplosis juniperina* (Felt, 1907)	6	6 (100%)	2	1
Micromya *Micromya lucorum* (Rondani, 1840)	2	1 (50%)	2	1

Gattung/Art Anzahlen:	**Sequenzen**	**Sequenzen mit Artname (%)**	**BINs**	**Arten**
Mikiola	3	0 (0%)	1	0
Monardia *Monardia toxicodendri* (FELT, 1907)	3	3 (100%)	1	1
Monobremia *Monobremia subterranea* (KIEFFER, 1898)	2	2 (100%)	1	1
Nikandria *Nikandria brevitarsis* (MAMAEV, 1964)	1	1 (100%)	1	1
Obolodiplosis *Obolodiplosis robiniae* (HALDEMAN, 1847)	1	1 (100%)	1	1
Ozirhincus *Ozirhincus millefolli* (WACHTL, 1884)	22	1 (4,5%)	3	1
Peromyia *Peromyia aeratipennis* (SKUSE, 1888) *Peromyia borealis* (HACKMAN, 1970) *Peromyia boreophila* (JASCHHOF, 2001) *Peromyia caricis* (KIEFFER, 1901) *Peromyia cornuta* (EDWARDS, 1938) *Peromyia fungicola* (KIEFFER, 1901) *Peromyia impexa* (SKUSE, 1888) *Peromyia monilis* (MAMAEV, 1965) *Peromyia ovalis* (EDWARDS, 1938) *Peromyia perpusilla* (WINNERTZ, 1870) *Peromyia ramose* (EDWARD, 1938) *Peromyia scutellata* (MAMAEV, 1990) *Peromyia upupoides* (JASCHHOF, 1997)	46	44 (95,7%)	16	13
Porricondyla *Porricondyla colpodioides* (MAMAEV, 1963)	25	25 (100%)	1	1
Resseliella *Resseliella theobaldi* (BARNES, 1927)	1	1 (100%)	1	1
Rhopalomyia	14	0 (0%)	5	0

Gattung/Art Anzahlen:	Sequenzen	Sequenzen mit Artname (%)	BINs	Arten
Scizomyia	6	0 (0%)	5	0
Sitodiplosis	2	0 (0%)	1	0
Winnertzia *Winnertzia tridens* (Panelius, 1965)	1	1 (100%)	1	1
nicht identifiziert	2.157	0 (0%)	472	0
Summe	**2.643**	**325 (12,3%)**	**535****	**30**

* Temporär-Name
** Vier Sequenzen wurden zwar jeweils einer BIN zugeordnet, blieben jedoch unbestimmt. Diese BINs erscheinen doppelt in der Tabelle: einmal unter der dazugehörige Art, und nochmal unter „nicht identifizierte Sequenzen". Aus diesem Grund sind in der Summe 535 BINs aufgelistet, obwohl es sich um 531 eindeutige BINs handelt.

Die 531 BINs aus dieser Studie entsprechen 63,5% der Anzahl der aktuell für ganz Deutschland bekannten Arten und 57,3% der von Morinière et al. (2019) gefundenen Anzahl von BINs, wobei nur 153 BINs dieser Studie auch von Morinière et al. (l.c.) nachgewiesen wurden. Abbildung 1 stellt die Häufigkeitsverteilung der Sequenzzahlen pro BIN im Datensatz dar. Einige wenige BINs waren in sehr großen Individuenzahlen vertreten, während einem Großteil der BINs sehr wenige Individuen zuzuordnen waren. 251 BINs, fast die Hälfte aller BINs, wurden durch ein einziges Individuum repräsentiert **(Abb. 1).**

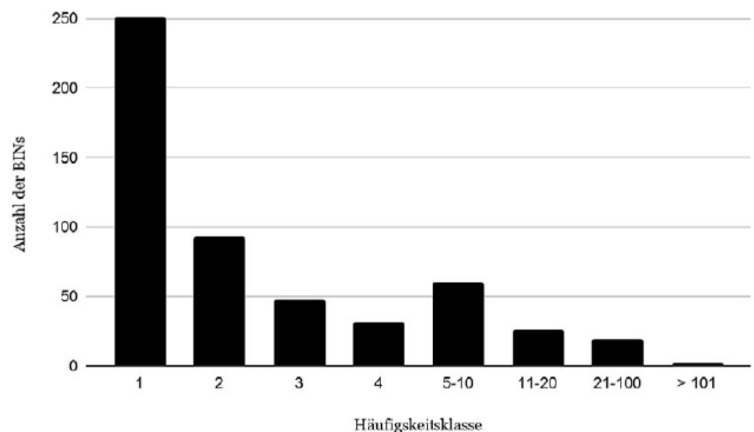


Abb. 1: Anzahl der BINs pro Häufigkeitsklasse (abhängig von der Individuenzahl pro BIN). Fast zwei Drittel aller BINs sind nur durch ein (251) oder zwei (93) Individuen vertreten.

Diskussion

Mit einer einzigen Falle im Stadtgebiet von München konnte auf dem Gelände der Zoologischen Staatssammlung München eine Diversität von 531 Gallmücken-BINs nachgewiesen werden. Auch wenn BINS nicht Arten gleichzusetzen sind, so besteht doch eine hohe Übereinstimmung zwischen Anzahl BINs und Anzahl Arten (Hausmann et al. 2013, Pentinsaari et al. 2014, Schmidt et al. 2015, Ratnasingham & Hebert 2013), so dass mit einer Artenzahl von über 500 Arten von Gallmücken zu rechnen ist. Diese hohe Artenvielfalt an Gallmücken ist auch deshalb bemerkenswert, weil mit hoher

Wahrscheinlichkeit bei weitem nicht alle am Standort vorkommenden Arten erfasst wurden. Zum einen beschränkt sich die bisherige Auswertung auf lediglich 10 Wochen. Zum anderen ist die Erfassung des Artenspektrums eines Standortes von zahlreichen Faktoren abhängig. Im bisher ausgewerteten Material sind fast zwei Drittel aller BINs nur durch ein oder zwei Individuen vertreten, also als selten einzustufen. Es ist davon auszugehen, dass es noch viel mehr seltene Arten im Untersuchungsgebiet gibt, deren Erfassung noch aussteht. Hinzu kommt die Erfahrung, dass bei der Vorauswahl von Morphospezies am Stereomikroskop für das nachfolgende Sequenzieren 20% oder sogar deutlich mehr der Arten nicht erkannt werden. Die tatsächliche Diversität dürfte also deutlich höher liegen, so dass es prinzipiell denkbar ist, dass allein an dem von uns untersuchten Standort mehr Arten vorkommen, als bisher in ganz Deutschland nachgewiesen wurden. Die tatsächliche Diversität der Gallmücken in Deutschland dürfte alle bisherigen Schätzungen weit übersteigen.

Von den bislang ermittelten 531 BINs konnten nur 30 (<6%) per Sequenzvergleich einer bereits beschriebenen und in BOLD identifizierten Art zugeordnet werden. Die übrigen 501 BINs blieben unbestimmt. Zudem waren 175 Sequenzen neu für die internationale Barcode of Life-Datenbank (BOLD), d.h. sie wurden bisher ausschließlich auf dem Gelände der ZSM nachgewiesen. Diese Befunde verdeutlichen zweierlei: in BOLD ist der Anteil bestimmter Arten bei taxonomisch vernachlässigten „Dark Taxa“ nach wie vor sehr gering, und es ist mit der Entdeckung einer großen Anzahl neuer Arten zu rechnen, und das selbst im Stadtgebiet von München.

Ausblick

Mit der Erforschung der Gallmücken im Rahmen des Projektes „GBOL III - Dark Taxa“ verfolgt die Zoologische Staatssammlung München das Ziel, die Diversität der deutschen Gallmückenfauna mit einem integrativ-taxonomischen Ansatz in einer Weise zu erfassen, dass (1) die Gesamtartenzahl realistischer als bisher abgeschätzt werden kann, (2) das Verhältnis von beschriebenen zu unbeschriebenen Arten deutlich wird und (3) für eine möglichst große Zahl von Arten ein deutscher Nachweis per DNA-Barcode erbracht wird. Zudem erwarten wir, dass die Taxonomie der Gallmücken um eine Reihe von Art-Neubeschreibungen aus Deutschland bereichert werden kann. Im Rahmen von „GBOL III: Dark Taxa“ werden an der SNSB-ZSM in den Jahren 2020 bis 2023 weitere wenig bekannte Insektengruppen bearbeitet: die Chironomidae (Diptera), Microgastrinae (Hymenoptera, Braconidae), Diapriidae (Hymenoptera) und einige weitere Gruppen. Die SNSB-ZSM reiht sich damit in das internationale Netzwerk von Forschern ein, die in den Jahren 2019–2026 die ehrgeizigen Ziele des globalen BIOSCAN Programms (https://ibol.org/programs/bioscan/) zu verwirklichen suchen.

Literatur

CALEY, J.M., FISHER, R. & K. MENGERSEN 2014: Global species richness estimates have not converged. – Trends in Ecology and Evolution **29**, 187–188.

CICCONARDI, F., FANCIULLI, P.P. & B. C. EMERSON 2013: Collembola, the biological species concept and the underestimation of global species richness. – Molecular Ecology **22**, 5382–5396.

ERWIN, T. L. 1982. Tropical forests: their richness in Coleoptera and other arthropod species. – The Coleopterists Bulletin **6**, 74–75.

GAGNÉ, R.J. & M. JASCHHOF 2017: A Catalog of the Cecidomyiidae (Diptera) of the World. – Digital, 762. Verfügbar unter: http://www.ars.usda.gov/ARSUserFiles/80420580/Gagne_2017_World_Cat_4th_ed.pdf.

HAUSMANN, A., GODFRAY, H. C. J., HUEMER, P., MUTANEN, M., ROUGERIE, R., VAN NIEUKERKEN, E. J., RATNASINGHAM, S. & P. D. HEBERT 2013: Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. – PloS ONE **8** (12), e84518 (11 pp).

HEBERT, P. D. N., RATNASINGHAM, S., ZAKHAROV, E. V., TELFER, A. C., LEVESQUE-BEAUDIN, V., MILTON, M. A., PEDERSON, S., JANNETTA, P. & J. R. DEWAARD 2016: Counting animal species with DNA barcodes: Canadian insects. – Philosophical Transactions of the Royal Society B: Biological Sciences, **371** (1702), 20150333 (10 pp).

KIMURA, M. 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. – Journal of Molecular Evolution **16** (2), 111–120.

MEYER, H. & M. JASCHHOF 1999: Cecidomyiidae, 18-32. In: SCHUMANN, H., BÄHRMANN, R. & A. STARK (eds): Entomofauna Germanica 2. Checkliste der Dipteren Deutschlands. – Studia dipterologica, Supplement **2**, 1–354.

MORA, C., TITTENSOR, D.P., ADI, S., SIMPSON, A.G.B. & B. WORM 2011: How many species are there on earth and in the ocean? – PLOS Biology **9**, e1001127 (8 pp).

MORINIÈRE, J., BALKE, M., DOCZKAL, D., GEIGER, M. F., HARDULAK, L. A., HASZPRUNAR, G., HAUSMANN, A., HENDRICH, L., REGALADO, L., RULIK, B., SCHMIDT, S., WÄGELE, J. W. & P. D. N. HEBERT, 2019: A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring. – Molecular Ecology Resources **19**, 900–928.

ÖDEGAARD, F. 2000: How many species of arthropods? Erwin's estimate revisited. – Biological Journal of the Linnean Society **71**, 583–597.

PENTINSAARI, M., HEBERT, P. D. N. & M. MUTANEN 2014: Barcoding beetles: a regional survey of 1872 species reveals high identification success and unusally deep interspecific divergences. – PLOS ONE **9**, e108651 (10 pp).

PUILLANDRE, N., LAMBERT, A., BROUILLET, S. & G. ACHAZ 2012: ABGD, Automatic Barcode Gap Discovery for primary species delimitation. – Molecular Ecology **21** (8), 1864–1877.

RATNASINGHAM, S. & P.D.N. HEBERT 2013: A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. – PLOS ONE **8** (7), e66213 (16 pp).

SCHMIDT, S., SCHMID-EGGER, C., MORINIÈRE, J., HASZPRUNAR, G. & P. D. N. HEBERT 2015: DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim). – Molecular Ecology Resources **15** (4), 985–1000.

STORK, N.E., MCBROOM, J., GELY, C. & HAMILTON, A.J. 2015: New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. – PNAS June 16, 2015 112 (24) 7519–7523; first published June 1, 2015; https://doi.org/10.1073/pnas.1502408112.

WUEHRL, L., PYLATIUK, C., GIERSCH, M., LAPP, F., VON RINTELEN, T., BALKE, M., SCHMIDT, S., CERRETTI, P. & R. MEIER 2021: DiversityScanner: Robotic discovery of small invertebrates with machine learning methods. bioRxiv, 2021.05.17.444523.

Anschriften der Verfasser:

M.Sc. Caroline CHIMENO
Münchhausenstraße 21
81247 München, Deutschland
E-Mail: chimeno@snsb.de

Dieter DOCZKAL
Münchhausenstraße 2
81247 München, Deutschland
E-Mail: doczkal@snsb.de

Prof. Dr. Gerhard HASZPRUNAR
Münchhausenstraße 21
81247 München, Deutschland
E-Mail: haszprunar@snsb.de

Dr. Axel HAUSMANN
Münchhausenstraße 21
81247 München, Deutschland
E-Mail: hausmann.a@snsb.de

Dr. Mathias JASCHHOF
Ölands Skogsby 161
38693 Färjestaden, Schweden
E-Mail: mjaschhof@yahoo.de

Dr. Marion KOTRBA
Münchhausenstraße 21
81247 München, Deutschland
E-Mail: kotrba@snsb.de

Kate PEREZ
50 Stone Road East
Guelph, Ontario, N1G2W1, Kanada
E-Mail: kperez@uoguelph.ca

Dr. habil. Michael RAUPACH
Münchhausenstraße 21
81247 München, Deutschland
E-Mail: raupach@snsb.de

Dr. Stefan SCHMIDT
Münchhausenstraße 21
81247 München, Deutschland
E-Mail: schmidt.s@snsb.de

Publication III:

Peering into the Darkness: DNA Barcoding Reveals Surprisingly High Diversity of Unknown Species of Diptera (Insecta) in Germany

Published in Insects

Chimeno, C., Hausmann, A., Schmidt, S., Raupach, M. J., Doczkal, D., Baranov, V., Hübner, J., Höcherl, A., Albrecht, R., Jaschhof, M., Haszprunar, G. & Hebert, P. D. (2022). Peering into the Darkness: DNA Barcoding Reveals Surprisingly High Diversity of Unknown Species of Diptera (Insecta) in Germany. *Insects*, 13(1), 82. 10.3390/insects13010082

Article

Peering into the Darkness: DNA Barcoding Reveals Surprisingly High Diversity of Unknown Species of Diptera (Insecta) in Germany

Caroline Chimeno [1,*], Axel Hausmann [1], Stefan Schmidt [1], Michael J. Raupach [1], Dieter Doczkal [1], Viktor Baranov [2], Jeremy Hübner [1], Amelie Höcherl [1], Rosa Albrecht [1], Mathias Jaschhof [3], Gerhard Haszprunar [1,2] and Paul D. N. Hebert [4]

[1] SNSB-Zoologische Staatssammlung München, Münchhausenstr. 21, 81247 München, Germany; hausmann.a@snsb.de (A.H.); schmidt.s@snsb.de (S.S.); raupach@snsb.de (M.J.R.); doczkal@snsb.de (D.D.); huebner@snsb.de (J.H.); hoecherl@snsb.de (A.H.); albrecht@snsb.de (R.A.); haszprunar@snsb.de (G.H.)
[2] Department Biology II, Ludwig-Maximilians-University of Munich (LMU), Großhaderner Str. 2, Martinsried, 82152 Planegg, Germany; baranowiktor@gmail.com
[3] Station Linné, Ölands Skogsby 161, 38693 Färjestaden, Sweden; mjaschhof@yahoo.de
[4] Centre for Biodiversity Genomics, University of Guelph, Guelph, ON N1G 2W1, Canada; phebert@uoguelph.ca
* Correspondence: chimeno@snsb.de

Simple Summary: Roughly two-thirds of the insect species described from Germany belong to the orders Diptera (flies) or Hymenoptera (wasps, bees, ants and sawflies). However, both orders contain several species-rich families that have received little taxonomic attention until now. This study takes the first step in assessing these "dark taxa" families and provides species estimates for four challenging groups of Diptera (Cecidomyiidae, Chironomidae, Phoridae and Sciaridae). The estimates given in this paper are based on the sequencing results of over 48,000 fly specimens that have been collected in southern Germany via Malaise traps that were operated for one season each. We evaluated the fraction of species in our samples belonging to well-known fly families in order to estimate the species richness of the challenging "dark taxa" (DT families hereafter). Our results suggest a surprisingly high proportion of undetected biodiversity in a supposedly well-investigated country: at least 1800–2200 species await discovery and description in Germany in these four families.

Citation: Chimeno, C.; Hausmann, A.; Schmidt, S.; Raupach, M.J.; Doczkal, D.; Baranov, V.; Hübner, J.; Höcherl, A.; Albrecht, R.; Jaschhof, M.; et al. Peering into the Darkness: DNA Barcoding Reveals Surprisingly High Diversity of Unknown Species of Diptera (Insecta) in Germany. *Insects* **2022**, *13*, 82. https://doi.org/10.3390/insects13010082

Academic Editor: Ding Yang

Received: 16 November 2021
Accepted: 5 January 2022
Published: 12 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Determining the size of the German insect fauna requires better knowledge of several megadiverse families of Diptera and Hymenoptera that are taxonomically challenging. This study takes the first step in assessing these "dark taxa" families and provides species estimates for four challenging groups of Diptera (Cecidomyiidae, Chironomidae, Phoridae, and Sciaridae). These estimates are based on more than 48,000 DNA barcodes (COI) from Diptera collected by Malaise traps that were deployed in southern Germany. We assessed the fraction of German species belonging to 11 fly families with well-studied taxonomy in these samples. The resultant ratios were then used to estimate the species richness of the four "dark taxa" families (DT families hereafter). Our results suggest a surprisingly high proportion of undetected biodiversity in a supposedly well-investigated country: at least 1800–2200 species await discovery in Germany in these four families. As this estimate is based on collections from one region of Germany, the species count will likely increase with expanded geographic sampling.

Keywords: Diptera; insects; dark taxa; taxonomic impediment; species estimates; DNA barcoding; biodiversity; German insect fauna

1. Introduction

Although the Central European insect fauna is considered to be well studied, gaps in knowledge of its taxonomy and biodiversity remain [1]. About 33,300 species of insects

are documented from Germany, of which roughly two-thirds of these taxa belong to one of the two orders: Diptera (flies) and Hymenoptera (wasps, bees, ants, and sawflies) [1–8]. However, both orders contain several species-rich families which have received less attention than others in Germany's long history of taxonomic research [1]. This reflects the confluence of several factors, such as extreme species richness combined with a high rate of cryptic diversity and, most importantly, the limited taxonomic attention directed to small specimens (<2 mm) whose morphological characteristics are difficult to evaluate. Successful identification of species in these groups using morphology is time-consuming and requires taxonomic expertise, the availability of which is decreasing [9–14]. This imbalance of few researchers but high species numbers still awaiting documentation is commonly referred to as the taxonomic impediment [9,15,16]. Against the backdrop of a worldwide decline in insect abundance, the taxonomic impediment is an alarming constraint to biodiversity surveys [17–21]. One such constraint is noticeable in the framework of DNA barcoding applications, where species proxies (Barcode Index Numbers, BINs) often lack a linkage to a known species [22]. Page [22] coined the term "dark taxa" for these nameless BINs, and in 2020, Hausmann et al. [1] used it to address species-rich, taxonomically challenging groups of insect families whose diversity remains mostly undescribed. These include certain families of non-brachyceran Diptera (mosquitoes, gnats, midges), some families of Brachycera (flies), and nearly all families of parasitoid Hymenoptera (wasps) which often make up the majority of the insect biodiversity present in environmental and bulk samples [23]. With the shortage of taxonomic specialists, the functional role of "dark taxa" in ecosystems is far too understudied, meaning that they cannot be included in biomonitoring or conservation surveys.

The most recent project in the German Barcode of Life initiative, GBOL III: Dark Taxa, was launched in mid-2020 to tackle these challenging groups. Its two main goals are: (1) to study various DT families using an integrative taxonomic approach which combines morphological and sequence data [1,24], and (2) to expand the DNA barcode reference library established by three earlier initiatives (Barcoding Fauna Bavarica, GBOL I, GBOL II) [24–26]. Work conducted by GBOL II generated a reference library for the order Diptera based on 50,963 COI sequences, data that provided barcodes for 5200 BINs [13]. A recent commentary on this study presented a classical dipterist's perspective on the situation for the better-known families of Diptera [27]. It explored ways to extend the involvement of expert taxonomists in assigning Linnean names to BINs. However, the challenge in implementing similar work on DT families was not addressed, highlighting the need to seek new approaches so these taxa can finally become more accessible to research.

This study begins this effort by considering the German fauna of four DT families of Diptera which lack estimates of their species numbers: Cecidomyiidae (gall midges), Chironomidae (non-biting midges), Phoridae (scuttle flies), and Sciaridae (dark-winged fungus gnats) (Figure 1). To address this goal, we examine the diversity of these DT families in our Malaise trap collections. We employ BIN data resulting from the sequence analysis of samples from southern Germany and use these results to estimate the extent of undocumented biodiversity in these families in Bavaria and Germany. An important backbone to our calculations is species numbers inferred from essential contributions of Germany's over 200-year-long history of taxonomy [5–8,28–38].

Figure 1. Selected representatives of the DT families analyzed in our study: Cecidomyiidae (**top left**); Phoridae (**top right**); Sciaridae (**bottom left**) and Chironomidae (**bottom right**). Scale bars represent 1 mm.

2. Materials and Methods

2.1. Malaise Tap Sites

In 2012, the Global Malaise Trap Program was launched by the Centre for Biodiversity Genomics (CBG) at the University of Guelph to provide a global overview of arthropod diversity [39]. As part of this project, 14 Malaise traps were deployed at various sites in Germany (Figure 2 and Table 1). In 2012, one trap was operated from May to September in the Bavarian Forest National Park (BFNP), a conifer-dominated montane forest. In 2014, 12 Malaise traps were placed along an altitudinal transect (1036–2160 m) in the Allgäu Alps, ranging from the Oytal to the Schochen and Nebelhorn Mountains. Traps in lower altitudes (Oytal) were deployed in May, whereas those in higher altitudes (Schochen and Koblat) were deployed in June. All traps in the Allgäu Alps were operated until October. Finally, in 2017, one trap was deployed at the Bavarian State Collection of Zoology (ZSM) in Munich, which is situated in a residential neighborhood rich in backyard gardens. This trap was operated from April to December. Altogether, the sampled sites represent a heterogeneous array of habitats typical of southern Germany. The specifics of trap deployment (habitat type, site, orientation, height) strongly influence its catch [40]. Collection dates varied among sites but are detailed in Table A1. Denatured ethanol (80%) was used to preserve specimens.

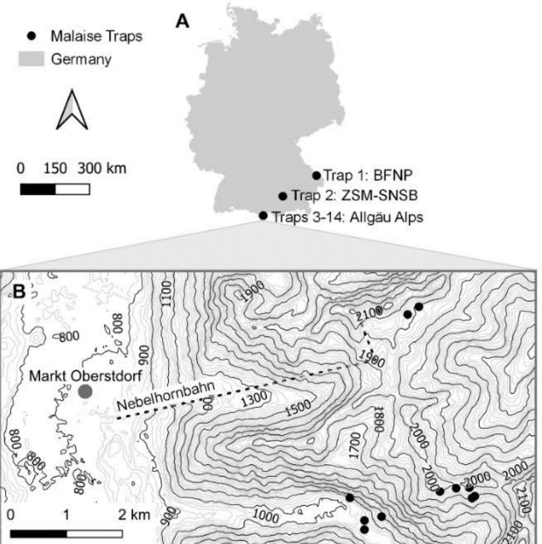


Figure 2. Malaise trap sites. Locations where the 14 Malaise traps were deployed in 2012, 2014, and 2017 ((**A**,**B**) shows enlarged map of Allgäu Alps) as Germany's contribution to the Global Malaise Trap Program.

Table 1. Malaise trap information. Trap site, exact location, elevation, and habitat type.

Site	Trap	Coordinates	Elevation	Habitat
BFNP	Trap 1	48.9509° N 13.422° E	842 m	Natural forest
ZSM	Trap 2	48.1648° N 11.4849° E	519 m	Urban, pre-alpine meadow
Allgäu Alps: Oytal	Trap 3	47.39205° N 10.34093° E	1122 m	Lake rock face
Allgäu Alps: Oytal	Trap 4	47.38903° N 10.34846° E	1200 m	Cone of scree
Allgäu Alps: Oytal	Trap 5	47.38842° N 10.34440° E	1056 m	Rough pasture
Allgäu Alps: Oytal	Trap 6	47.38695° N 10.34438° E	1036 m	River
Allgäu Alps: Schochen	Trap 7	47.39202° N 10.36991° E	1930 m	Alpine grassland
Allgäu Alps: Schochen	Trap 8	47.39232° N 10.37057° E	1908 m	Spring
Allgäu Alps: Schochen	Trap 9	47.39368° N 10.36926° E	2032 m	South-exposed ridge with Blaugras-Horstseggenrasen
Allgäu Alps: Schochen	Trap 10	47.39307° N 10.36229° E	2010 m	South-exposed rock
Allgäu Alps: Schochen	Trap 11	47.39360° N 10.36615° E	1980 m	Snow bed
Allgäu Alps: Koblat	Trap 12	47.42223° N 10.34783° E	2160 m	South-exposed rock face
Allgäu Alps: Koblat	Trap 13	47.42147° N 10.35465° E	2033 m	Snow bed
Allgäu Alps: Koblat	Trap 14	47.42272° N 10.35730° E	2005 m	Mountain pine bush

2.2. Processing of Specimens

Samples from two sites (BFNP, ZSM) were sent directly to the CBG for analysis. Due to funding constraints, roughly every second weekly sample from the BFNP and every fourth weekly sample from the ZSM were selected for DNA barcode analysis. Based on the number of specimens in the samples that were processed, the full year of collecting at these sites yielded about 52,000 and 130,000 specimens, respectively. Using morphology, specimens from these locales were sorted to an order prior to sequence analysis and to a family after analysis. In total, tissue samples or whole individuals of 62,073 specimens (29,481 from BFNP; 32,592 from ZSM) were transferred to 96-well microplates for DNA

extraction. Samples from the Allgäu Alps were sorted by a dipterist at the ZSM before being dispatched in 96-well microplates to the CBG for sequence analysis. Rough estimates suggest the Allgäu samples included well over a million specimens, but funding was only available to process about 2% of them (20,250 specimens).

At the CBG, specimens were processed using standard protocols for DNA extraction, PCR amplification of the barcode region of COI, and sequencing. Specimens from the BFNP and the Allgäu Alps were Sanger sequenced on an ABI 3730XL [41], while specimens from the ZSM were sequenced on Sequel [42].

2.3. Data Analysis

All specimen metadata and sequence data were uploaded to the Barcode of Life Data System (BOLD), an online workbench and database [32]. These data are publicly available in three datasets: DS-BFNP, DS-ZSMTRAP and DS-ALGALPS. Each sequence $\geq$ 300 base pairs (bp) was automatically assigned to a Barcode Index Number (BIN) already in BOLD if sequence similarity based on the (RESL-) BIN algorithm was fulfilled [43]. Sequences $\geq$ 500 bp which did not find a match served as founders of new BINs. All data were downloaded on 8 February 2021 for further analysis. Therefore, the present results correspond to BINs assigned at that time (BIN assignments can change as new sequences are added to BOLD). Employing BINs as a proxy for species, we employed Chao1 [44] to estimate species counts for the dipteran families selected for analysis. We then calculated the ratio between the observed number of BINs in our samples to the estimate of species richness generated by Chao1 to ascertain the proportion of species at the sampling sites that have not been captured by our Malaise traps and that await analysis. We also generated continuous diversity profiles that illustrated variation in three standard metrics of biodiversity, which are quantified by Hill numbers (q): species richness (q = 0), Shannon diversity (q = 1), and Simpson diversity (q = 2) [34]. Hill numbers are a mathematically consolidated group of diversity indices which include relative species abundances in order to quantify biodiversity [45]. All calculations were performed in R version 3.3.6 with the Chao1 estimates calculated using the *SpadeR* package [46].

2.4. Extrapolating Species Numbers

We selected, more or less randomly, 11 dipteran families whose taxonomy and fauna have been intensively studied to date in order to assess the fractions of the Bavarian and German faunas represented in our samples (Table 2). By comparing the known species counts for these 11 families with the species recovered from our Malaise traps, we could estimate the percentage of these taxa that were recovered, providing a basis for estimating the completeness of our sampling. These values could then be used to estimate species diversity for our four DT families: Cecidomyiidae—gall midges; Chironomidae—non-biting midges; Phoridae—scuttle flies, and Sciaridae—dark-winged fungus gnats.

Species numbers for Germany and for Bavaria were obtained from extensive literature (Table 2). For each family where a species count for Bavaria was unavailable, we adopted a count equal to 0.80 of the species number for Germany. This value was conservative because where species lists were available for both Bavaria and Germany, the ratio often exceeded 0.80 (Table 2). Moreover, this proportion corresponds to past evidence that Bavaria hosts 80–85% of the German fauna in well-studied invertebrate groups, both terrestrial and limnic [2,47].

Table 2. Species numbers for 15 families of Diptera. Species numbers for the Bavarian and German faunas are shown for 11 families of Diptera with well-established taxonomy and for four families with limited knowledge (Cecidomyiidae, Chironomidae, Phoridae, Sciaridae). *—estimated at 80% of German fauna.

Taxon	Bavarian Species Count	German Species Count	Species Count Bavaria/Germany
Asilidae	68 [28]	85 [29]	0.80
Calliphoridae	50 *	62 [35]	0.80 *
Drosophilidae	64 [28]	81 [37]	0.79
Ephydridae	140 *	174 [38]	0.80 *
Muscidae	267 *	334 [48]	0.80 *
Sarcophagidae	107 *	134 [35]	0.80 *
Stratiomyidae	59 [28]	71 [30,48]	0.83
Syrphidae	389 [28]	458 [31]	0.85
Tabanidae	47 [28]	58 [8,48]	0.81
Tachinidae	361 [28]	501 [48]	0.72
Tipulidae	120 [33]	142 [32]	0.85
Cecidomyiidae	328 [38]	859 [5–8]	0.38
Chironomidae	576 [28]	781 [5–8]	0.74
Phoridae	302 *	378 [5–8]	0.80 *
Sciaridae	231 [28]	343 [43]	0.67
All Diptera	7635 *	9544 [8]	0.80 *

We estimated species numbers for the DT families through the following steps:

1. We calculated a Recovery Ratio by dividing the number of BINs detected through sequencing by the species count for each of the 15 families and for all Diptera (BIN/species ratio). This approach generated a ratio for each well-known family, for each DT family, and for all Diptera.
2. We estimated the maximum number of species for each "dark taxon" for both Germany and Bavaria by dividing its BIN count by the average BIN/species ratio of all 11 well-known families.
3. We estimated the minimum species number for each "dark taxon" by dividing all Diptera BINs by all Diptera species (i.e., 9544). Because this calculation includes numerous families with cryptic diversity, the resultant values underestimate the diversity of the DT families.

In the same fashion, we extrapolated species numbers employing the Chao1 values for the four DT families.

3. Results

3.1. Sequencing Results

COI sequences were recovered from 85.4% of the insects (70,293/82,323) that were analyzed (Table 3) and success was even higher for Diptera (91%). Diptera comprised nearly two thirds of the specimens that were analyzed and more than half of the resultant BINs. When results for Diptera from the three collection sites were pooled, the resulting 48,230 COI sequences were assigned to 4863 BINs and included species from 85 families. Across all sites, roughly 20% of the BINs were new to BOLD and almost 70% of them were Diptera with representatives from 56 families. Almost half of all dipteran BINs (2146; 44.1%) and 55% of the new dipteran BINs belonged to the four DT families.

Table 3. Sequence results for the three sampling sites. Total sample size, number of processed specimens, sequences recovered, BINs, BINs new to BOLD, Diptera specimens, and Diptera BINs.

	BFNP	ZSM	Allgäu Alps	Total
Samples (trap × collection events)	1 × 9 = 9	1 × 10 = 10	8 × 7 + 4 × 10 = 96	100
All				
Specimens	29,481	32,592	20,250	82,323
COI sequences (% success)	25,217 (85.6%)	28,923 (88.7%)	16,152 (79.8%)	70,293 (85.4%)
BINs (% new to BOLD)	2565 (19.4%)	3870 (15.8%)	4043 (23.0%)	8790 (23.8%)
Diptera				
Specimens (% of all specimens)	23,114 (78%)	15,448 (47%)	14,238 (70%)	52,800 (64%)
COI sequences (% success)	20,909 (91%)	14,983 (97%)	12,338 (87%)	48,230 (91%)
BINs (in % of all BINs)	1571 (61%)	1676 (43%)	2632 (65%)	4863 (55%)
Diptera BINs new to BOLD	375	260	736	1413
DT BINs new to BOLD (% of all new Diptera BINs)	337 (90%)	215 (83%)	215 (29%)	780 (55%)

3.2. Estimation of Taxon Diversity Using BIN/Species Ratios

The 11 well-known families of Diptera displayed BIN/species ratios that ranged from 0.19–0.60 (ø 0.33 ± 0.9) for Bavaria and from 0.15–0.48 (ø 0.27 ± 0. 7) for Germany (Table 4, Figure A1a). Dividing all Diptera BINs by all known Diptera species produced a ratio of 0.64 for Bavaria and 0.51 for Germany. While one DT family (Chironomidae) possessed a ratio (0.38, Germany) that overlapped the upper end of the values for the 11 well-known families, the other three had far higher ratios. In fact, the BIN count for Phoridae and Sciaridae nearly matched the known species count for Germany, while the count for Cecidomyiidae exceeded it.

Table 4. Fifteen families of Diptera, 11 with well-developed taxonomy and four that are less well known. The number of BINs recovered in this study is followed by the known species count for Bavaria and Germany, the ratio of species counts for Bavaria and Germany, and BIN/Species ratios for Bavaria and Germany.

Taxa	BINs	Bavarian Species	German Species	Bavarian/German Species	BINs/Bavarian Species	BINs/German Species
Asilidae	13	68	85	0.80	0.19	0.15
Calliphoridae	22	50	62	0.80	0.44	0.35
Drosophilidae	27	64	81	0.79	0.42	0.34
Ephydridae	32	140	174	0.80	0.23	0.18
Muscidae	160	267	334	0.80	0.60	0.48
Sarcophagidae	35	107	134	0.80	0.33	0.26
Stratiomyidae	14	59	71	0.83	0.24	0.20
Syrphidae	131	389	458	0.85	0.34	0.29
Tabanidae	9	47	58	0.81	0.19	0.16
Tachinidae	126	361	501	0.72	0.35	0.25
Tipulidae	43	120	142	0.85	0.36	0.30
Average values					0.33 ± 0.9	0.27 ± 0.7
Cecidomyiidae	1163	328	859	0.38	3.55	1.35
Chironomidae	296	576	781	0.74	0.51	0.38
Phoridae	348	302	378	0.80	1.15	0.92
Sciaridae	339	231	343	0.72	1.47	0.99
Average values					1.67 ± 0.9	0.91 ± 0.3
All Diptera	4863	7635	9544	0.80	0.64	0.51

3.3. Estimation of Taxon Diversity Using Chao1/Species Ratios

Chao1 estimates of species richness were obtained for the 15 families of Diptera (Table 5). BIN/Chao1 ratios averaged 0.76 for the 11 well-known families. The diversity profiles for 10 of these families showed overlap between the species richness in our samples

and that estimated to occur at the sites sampled by our Malaise traps (Hill number q = 0, Figure 3). Muscidae was the sole exception as its predicted diversity was considerably higher than currently recognized. Chao1/species ratios ranged from 0.21–0.82 (0.46 ± 0.2) for Bavaria and from 0.16–0.66 (0.37 ± 0.2) for Germany (Table 5).

Table 5. Proportion of undocumented Diptera biodiversity for Bavaria and Germany based on Chao1 estimates for 15 families.

Taxon	BINs	Chao1	BIN/Chao1	Bavarian Species	German Species	Chao1/ Bavarian Species	Chao1/ German Species
Asilidae	13	16	0.81	68	85	0.24	0.16
Calliphoridae	22	28	0.79	50	62	0.56	0.45
Drosophilidae	27	38	0.71	64	81	0.59	0.47
Ephydridae	32	88	0.36	140	174	0.63	0.51
Muscidae	160	220	0.73	267	334	0.82	0.66
Sarcophagidae	35	41	0.85	107	134	0.38	0.31
Stratiomyidae	14	16	0.88	59	71	0.27	0.23
Syrphidae	131	158	0.83	389	458	0.41	0.34
Tabanidae	9	10	0.90	47	58	0.21	0.17
Tachinidae	126	153	0.82	361	501	0.42	0.31
Tipulidae	43	59	0.73	120	142	0.49	0.42
Average values						0.46 ± 0.2	0.37 ± 0.2
Cecidomyiidae	1163	1937	0.60	328	859	5.91	2.25
Chironomidae	296	479	0.62	576	781	0.83	0.61
Phoridae	348	432	0.81	302	378	1.43	1.14
Sciaridae	339	468	0.72	231	343	2.03	1.36
Average values						2.55 ± 1.7	1.34 ± 0.5
All Diptera	4863	6927	0.70	7635	9544	0.91	0.73

The BIN/Chao1 ratios for the DT families were similar to those for the well-known families, ranging from 0.60–0.81 (ø 0.69 ± 0.8). The diversity profiles for all four families (Figure 4) showed no overlap between observed and estimated species richness (i.e., Hill number q = 0). Chao1/species ratios indicated coverages of 0.83–5.91 for Bavaria and 0.61–2.25 for Germany (Table 5). Excluding Chironomidae, all DT families possessed ratios well above 1. Considering all Diptera, our samples recovered about 70% of the species estimated to occur at the study sites, meaning that as many as 6927 BINs of Diptera could have been collected during sampling. Chao1/species ratios were 0.91 for Bavaria and 0.73 for Germany.

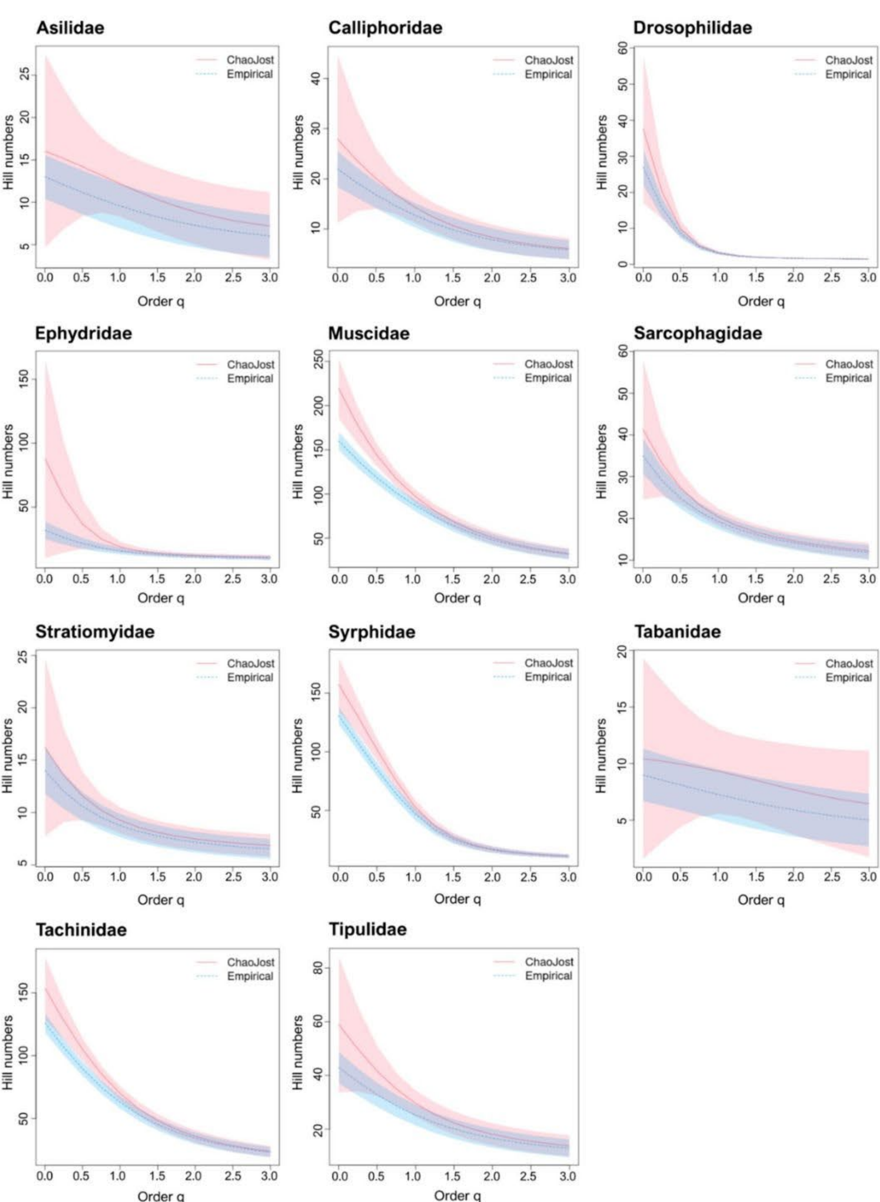


Figure 3. Diversity profiles for 11 well-known taxa. The empirical (BIN counts; dotted blue) and estimated (Chao1; red) diversity profiles for communities where Malaise traps were deployed, as quantified by Hill numbers for each of the 11 well-known families for values of the diversity order (q) from 0–3 with 95% confidence intervals (shaded areas based on bootstrap analysis of 100 permutations). Species richness is depicted by q = 0; Shannon diversity by q = 1; and Simpson diversity by q = 2.

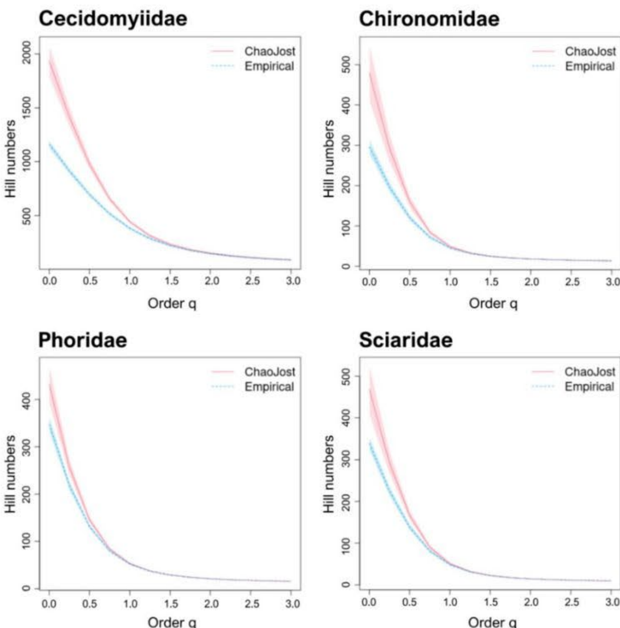


Figure 4. Diversity profiles for the four DT families. The empirical (BIN counts; dotted blue) and estimated (Chao1; red) diversity profiles for communities where Malaise traps were deployed, as quantified by Hill numbers for each of the four "dark taxa" families for values of the diversity order (q) from 0–3 with 95% confidence intervals (shaded areas based on bootstrap analysis of 100 permutations). Species richness is depicted by q = 0; Shannon diversity by q = 1; and Simpson diversity by q = 2.

3.4. Extrapolating Species Numbers

We employed the two ratios to estimate the number of species in the DT families. First, we used BIN/species ratios to extrapolate species numbers based on the number of observed BINs. Second, we used the Chao1/species ratios to estimate species numbers based on the estimated BIN diversity. The first approach generates more conservative values than the second. We divided the number of observed BINs by the (BIN or Chao1)/species ratio for all Diptera to calculate minimum species numbers. To obtain an upper limit, we divided the number of observed BINs for each family by the average (BIN or Chao1)/species ratio for all well-known families. The following calculation is presented below (e.g., Sciaridae).

As 339 Sciaridae BINs were recovered, the minimum species estimate for Bavaria was 530 (339/0.64), while the upper estimate was 1027 (339/0.33). Similarly, the number of species in Germany could be estimated as ranging from 665 (339/0.51) to 1255 (339/0.27) species. By making similar calculations for each DT family, an overall estimate for total species numbers in Bavaria and Germany was obtained (Table 6). The number of species that await discovery in each region can then be obtained by subtracting the number of known species from these estimates.

Table 6. BINs and calculated estimates. Total number of BINs recovered for each family from all traps, our calculated estimates, number of recorded species, and potential amplitude of new records for Bavaria and Germany.

Dark Taxa	BINs	Estimates Bavaria	Bavarian Species	New Records Bavaria	Estimates Germany	German Species	New Records Germany
BIN/species ratio							
Cecidomyiidae	1163	1817–3524	328	1489–3196	2280–4307	859	1421–3448
Chironomidae	296	463–897	576	0–321	580–1096	781	0–315
Phoridae	348	544–1055	302	242–753	682–1289	378	304–911
Sciaridae	339	530–1027	231	299–796	665–1256	343	322–913
Chao1/species ratio							
Cecidomyiidae	1937	2129–4211	328	1801–3883	2653–5235	859	1794–4376
Chironomidae	479	526–1041	576	0–465	656–1295	781	0–514
Phoridae	432	475–939	302	173–637	592–1168	378	214–790
Sciaridae	468	514–1017	231	283–786	641–1265	343	298–922

In total, we recovered 2146 BINs for the DT families which is 22% of the total count of dipteran species known from Germany. Our conservative estimate suggested that just the DT families comprise about 3300–6500 species in Bavaria versus 4200–7900 in Germany. Based on the current species count for Diptera in Bavaria (7635) and Germany (9544), and our estimate of new record, this implies an increase of 25–66% and by 19–59% respectively.

By comparison, the Chao1 analysis suggested that 3316 BINs of the DT families occurred at our sampling sites, a 54% increase from current estimates. Based on this approach, there about 2200–5800 species in Bavaria and 2200–6600 in Germany that may still await documentation. Hence, this approach raises the species count for Diptera by 29–75% for Bavaria and by 22–69% for Germany.

4. Discussion

Although members of the order Diptera comprise almost a third of Germany's insect fauna, the true diversity of the four highly diverse families [1] examined in this study is likely much higher than previously assumed [13,38]. By assessing the number of BINs sequenced from our collections and extrapolating species numbers, we obtained an initial estimate of their species numbers. Our results suggest that at least 1900–2200 dipteran species await discovery in Bavaria versus 1800–2200 in Germany. Although our species estimates were only based on sequencing Bavarian specimens, they are likely a good approximation of diversity in Germany as 80–85% of the invertebrate species found in Germany occur in Bavaria [2,36]. While Bavaria does have some habitats (e.g., alpine) that are not found in other regions of Germany, other habitats (e.g., coastal marshes) are absent [2], meaning that species specialized in the latter habitats will not occur in the state.

4.1. DNA Barcoding: Using BIN Numbers as Proxies for Species Numbers

Prior studies [49] have demonstrated that DNA barcoding is not only effective for specimen identification, but is also valuable for estimating species numbers [50–53]. Although there is strong correspondence between BIN counts and species numbers [49,54], several factors can lead to differences [54]. For example, COI numts can lead to the overestimation of species numbers if they are preferentially amplified in some specimens [55–58]. Conversely, the introgression of mitochondrial DNA (mtDNA), incomplete lineage sorting, and recent speciation can lead to underestimation of species numbers [59–61]. Other factors that challenge COI-based species identifications include heteroplasmy [62] and the homogenization of mtDNA haplotypes due to the maternally inherited endosymbiont *Wolbachia* [63,64]. These underlying molecular factors can lead the BIN algorithm on BOLD to assign members of a single species to several BINs or to assign several species to a single BIN. In groups

with well-developed taxonomic systems, the BIN algorithm typically underestimates the true species count by about 10% as it was designed to deliver a conservative value for species diversity [65]. In addition to this internal constraint, two operational factors may have led our study to substantially underestimate actual species numbers:

1. Limited geographic sampling as our data originates from few sites in Bavaria only, covering a tiny fraction of habitat types otherwise present.
2. Limited funding constrained analysis to just 5% of the 1.2 million specimens that were collected.

4.2. BIN & Chao1/Species Ratios: Well-Known Families versus DT Families

We assessed the completeness of the species coverage provided by our Malaise trap samples in two ways. First, we calculated the ratio of the BINs recovered for each family and its known species count for Bavaria and Germany. We then made the same calculation employing Chao1 estimates, which, in contrast to the first approach, includes species that were present at our sampling sites but not caught nor sequenced. Thus, it is important to note that our first approach generates more conservative values than the second. By calculating the BIN/Chao1 ratios for each taxon, we were able to make the proportion of diversity that was not captured tangible.

Overall, the resulting (BIN or Chao1)/species ratios were much higher for the DT families than for the well-known ones (Tables 4 and 5). Average ratios among the well-known families were well under 1 (ranging from 0.33–0.46 for Bavaria and 0.27–0.37 for Germany), indicating that our collections only included a fraction of the known diversity from Bavaria and Germany. This was expected because we only sampled few sites and only processed a fraction of our dipteran specimens. The much higher ratios for the DT families (average ranging from 1.67–2.55 for Bavaria and 0.91–1.34 for Germany) strongly suggest the presence of undescribed, unknown species. The Cecidomyiidae were the most dramatic case as we detected 1163 BINs, a value 35% higher than the species count for this family in Germany [8]. In fact, a quarter of all Diptera BINs belonged to this family, reinforcing conclusions from earlier studies indicating that this is the most diverse family of flies [13,49]. For example, extensive sampling at sites across Canada [49] revealed more than 10,000 BINs, a result which suggested that the Cecidomyiidae may include two million species worldwide. The Bavarian fauna has received little taxonomic attention as only 328 species are recorded versus a likely count of 687 species based on the presumption that 80% of the German fauna occurs there. By contrast, our analysis of 7148 specimens revealed 1163 BINs, a count for Bavaria which is threefold higher than the number of recorded species. Chironomidae was an exception among our DT families, as we obtained ratios that were consistent with those of the well-known families (Table 5). Although Chironomidae is a dark taxon, extensive research concerning the systematics, taxonomy, and nomenclature of European and Neotropical species has and is being conducted at the Bavarian State Collection of Zoology (ZSM) by the late Ernst Fittkau (former director of the ZSM) and his students including Martin Spies, the current editor of the Chironomid Home Page [66]. We therefore expect that the chironomid fauna of Bavaria and Germany is well documented and that, in contrast to the other DT families, a much lower amplitude of new species will be discovered in the following years of GBOL III. Among the well-known families, the Muscidae displayed the highest BIN/species indicating that the current species count considerably underrepresents its actual diversity. As a result, the Muscidae should also be recognized as a DT family.

4.3. Discrepancies in Taxa Coverage in Our Malaise Traps

Our estimated species counts for the DT families are based on the presumption that recovery success for the 11 families with strong taxonomy is a useful predictor of recovery success for the DT families. Our results did reveal threefold differences in recovery success among the well-known families, being lowest for Asilidae and Tabanidae and highest for the Muscidae. In our study, we used Malaise traps as a source of insect material, because

they enable sampling of high numbers of flying insects, especially Diptera [67–69]. However, a bias favoring the sampling of some taxa over others is always present, meaning that the community captured with such traps does not depict the true insect community of a sampled site [67]. Furthermore, the setup of a Malaise trap in terms of site choice, orientation, and above-ground-level is another source of bias, and these factors strongly influence sampling results [40]. To incorporate such variations, we used different approaches for extrapolating species numbers including Chao1 estimate calculations, which consider the unsampled taxa present at the sampling sites. The resulting Chao1 values indicated that we only recovered about 70% of the dipteran species present at the sites. In this manner, we obtained BIN estimates for each family that consider recovery success and unsampled taxa. Our results indicate that more than 3316 more BINs await detection, a total that would raise the number of Dipteran species in Germany by a third.

5. Conclusions

In this study, we aimed at estimating the number of species in the Bavarian and German faunas for four families of Diptera that are prime examples of “dark taxa”. Our estimates were inferred from the analysis of sequence data, reproducible genetic patterns, rather than on speculations. The confidence intervals on these estimates are broad (Table 5), reflecting the various factors that influence any effort to gauge species diversity. Despite our limited geographic sampling effort, our results strongly suggest that a surprisingly high proportion of Germany’s biodiversity is yet to be discovered.

Author Contributions: Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work: A.H. (Axel Hausmann), C.C., V.B., D.D., S.S.; drafting the work or revising it critically for important intellectual content: A.H. (Axel Hausmann), C.C., V.B., P.D.N.H., S.S., G.H., M.J., A.H. (Amelie Höcherl), J.H., M.J.R., R.A.; final approval of the version to be published: A.H. (Axel Hausmann), C.C., S.S.; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: A.H. (Axel Hausmann), C.C., V.B., P.D.N.H., S.S., G.H., M.J., A.H. (Amelie Höcherl), J.H., M.J.R., R.A. All authors have read and agreed to the published version of the manuscript.

Funding: GBOL projects were supported by grants from the German Federal Ministry of Education and Research (GBOL I: 01LI1101; GBOL II: 01LI1501, GBOL III: 16LI1901B). The Barcoding Fauna Bavarica project was funded by the Bayerisches Staatsministerium für Wissenschaft und Kunst. Sequence analyses were partially financed by funding from the government of Canada through Genome Canada and the Ontario Genomics Institute in support of the International Barcode of Life project.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The datasets containing all sequence data are publicly available in three datasets on the Barcode of Life Data System: DS-BFNP, DS-ZSMTRAP and DS-ALGALPS.

Acknowledgments: The GBOL III project is generously supported by a grant from the German Federal Ministry of Education and Research (FKZ 16LI1901B). We express our sincere thanks to the team at the Zoological Research Museum Alexander Koenig (ZMFK) for leading the GBOL III consortium and extensive project coordination. In particular, we would like to thank Marion Kotrba for valuable scientific advice and support.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Collection events for each Malaise trap.

Site	Trap	Processed Collection Events
BFNP 2012	1	8 May; 22 May; 8 June; 20 June; 4 July; 25 July; 12 August; 3 September; 22 September 2012.
ZSM-SNSB	2	10 April; 8 May; 5 June; 3 July; 31 July; 28 August; 25 September; 23 October; 20 November; 29 December 2017.
Allgäu Alps: Oytal	3	4 May; 17 May; 1 June; 16 June; 5 July: 20 July; 7 August; 29 August; 2 October; 27 October 2014.
Allgäu Alps: Oytal	4	4 May; 17 May; 1 June; 16 June; 5 July: 20 July; 7 August; 29 August; 2 October; 27 October 2014.
Allgäu Alps: Oytal	5	4 May; 17 May; 1 June; 16 June; 5 July: 20 July; 7 August; 29 August; 2 October; 27 October 2014.
Allgäu Alps: Oytal	6	4 May; 17 May; 1 June; 16 June; 5 July: 20 July; 7 August; 29 August; 2 October; 27 October 2014.
Allgäu Alps: Schochen	7	21 June; 4 July; 17 July; 6 August; 4 September; 29 September; 19 October 2014.
Allgäu Alps: Schochen	8	21 June; 4 July; 17 July; 6 August; 4 September; 29 September; 19 October 2014.
Allgäu Alps: Schochen	9	21 June; 4 July; 17 July; 6 August; 4 September; 29 September; 19 October 2014.
Allgäu Alps: Schochen	10	21 June; 4 July; 17 July; 6 August; 4 September; 29 September; 19 October 2014.
Allgäu Alps: Schochen	11	21 June; 4 July; 17 July; 6 August; 4 September; 29 September; 19 October 2014.
Allgäu Alps: Koblat	12	23 June, 4 July, 17 July; 8 August; 8 September; 5 September, 27 September; 20 October 2014.
Allgäu Alps: Koblat	13	23 June, 4 July, 17 July; 8 August; 8 September; 5 September, 27 September; 20 October 2014.
Allgäu Alps: Koblat	14	23 June, 4 July, 17 July; 8 August; 8 September; 5 September, 27 September; 20 October 2014.

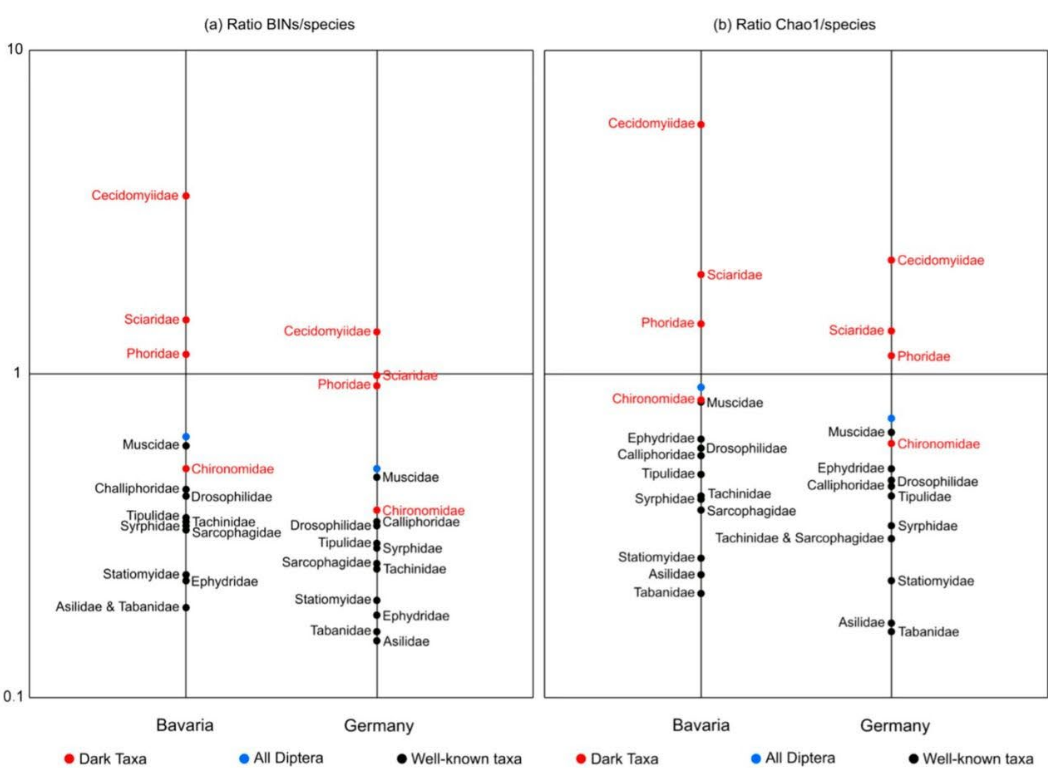


Figure A1. Ratio of BIN or Chao1 counts versus recorded species counts ratios for each family on a logarithmic scale. (**a**) BINs/species and (**b**) Chao1/species for well-known families, problematic families, and for all Diptera for Bavaria and Germany.

References

1. Hausmann, A.; Krogmann, L.; Peters, R.; Rduch, V.; Schmidt, S. GBOL III: DARK TAXA. *iBOL Barcode Bull.* **2020**, *10*, 4. [CrossRef]
2. Dathe, H.H.; Taeger, A.; Blank, S.M. Entomofauna Germanica Band 4 (Hymenoptera). Verzeichnis der Hautflügler Deutschlands. *Entomol. Nachr. Ber.* **2001**, *Beiheft 7*, 1–180.
3. Dathe, H.H.; Blank, S.M. Nachträge zum Verzeichnis der Hautflügler Deutschlands, Entomofauna Germanica Band 4 (Hymenoptera) (1). *Entomol. Nachr. Ber.* **2004**, *48*, 179–183.
4. Völkl, W.; Blick, T.; Kornacher, P.M.; Martens, H. Quantitativer Überblick über die rezente Fauna von Deutschland. *Nat. Landsch.* **2004**, *79*, 293–295.
5. Schumann, H. Checkliste der Dipteren Deutschlands. In *Entomofauna Germanica 2. Studia Dipterologica Supplement*; Ampyx-Verlag: Halle/Saale, Germany, 1999; Volume 2, p. 354.
6. Schumann, H. Erster Nachtrag zur "Checkliste der Dipteren Deutschlands". *Studia Dipterol.* **2002**, *9*, 437–445.
7. Schumann, H. Zweiter Nachtrag zur "Checkliste der Dipteren Deutschlands". *Studia Dipterol.* **2004**, *11*, 619–630.
8. Schumann, H. Dritter Nachtrag zur "Checkliste der Dipteren Deutschlands". *Studia Dipterol.* **2010**, *16*, 17–27.
9. de Carvalho, M.R. Revisiting the taxonomic impediment. *Science* **2005**, *307*, 353b. [CrossRef]
10. Mandelik, Y.; Roll, U.; Fleischer, A. Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *J. Appl. Ecol.* **2010**, *47*, 1179–1188. [CrossRef]
11. Chan, A.; Chiang, L.-P.; Hapuarachchi, H.C.; Tan, C.-H.; Pang, S.-C.; Lee, R.; Lee, K.-S.; Ng, L.-C.; Lam-Phua, S.-G. DNA Barcoding: Complementing morphological identification of mosquito species in Singapore. *Parasites Vectors* **2014**, *7*, 569. [CrossRef]
12. Porter, T.M.; Gibson, J.F.; Shokralla, S.; Baird, D.J.; Golding, G.B.; Hajibabaei, M. Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome *c* oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Mol. Ecol. Resour.* **2014**, *14*, 929–942. [CrossRef]

13. Morinière, J.; Balke, M.; Doczkal, D.; Geiger, M.F.; Hardulak, L.A.; Haszprunar, G.; Hausmann, A.; Hendrich, L.; Regalado, L.; Rulik, B.; et al. A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring. *Mol. Ecol. Resour.* **2019**, *19*, 900–928. [CrossRef]
14. Brown, B.V. Sampling methods for adult flies (Diptera). In *Measuring Arthropod Biodiversity: A Handbook of Sampling Methods*; Santos, J.C., Fernandes, G.W., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 187–204. [CrossRef]
15. Giangrande, A. Biodiversity, conservation, and the 'taxonomic impediment'. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **2003**, *13*, 451–459. [CrossRef]
16. de Carvalho, M.R.; Bockmann, F.A.; Amorim, D.S.; Brandão, C.R.F.; de Vivo, M.; de Figueiredo, J.L.; Britski, H.A.; de Pinna, M.C.C.; Menezes, N.A.; Marques, F.P.L.; et al. Taxonomic impediment or impediment to taxonomy? A commentary on systematics and the cybertaxonomic-automation paradigm. *Evol. Biol.* **2007**, *34*, 140–143. [CrossRef]
17. Wenzel, M.; Schmitt, T.; Weitzel, M.; Seitz, A. The severe decline of butterflies on western German calcareous grasslands during the last 30 years: A conservation problem. *Biol. Conserv.* **2006**, *4*, 542–552. [CrossRef]
18. Sodhi, N.S.; Brook, B.W.; Bradshaw, C.J.A. Causes and consequences of species extinctions. In *The Princeton Guide to Ecology*; Levin, S.A., Ed.; Princton University Press: Princeton, NJ, USA, 2009; pp. 514–520.
19. Coleman, C.O. Taxonomy in times of the taxonomic impediment—Examples from the community of experts on amphipod crustaceans. *J. Crustacean Biol.* **2015**, *35*, 729–740. [CrossRef]
20. Hallmann, C.A.; Sorg, M.; Jongejans, E.; Siepel, H.; Hofland, N.; Schwan, H.; Stenmans, W.; Müller, A.; Sumser, H.; Hörren, T.; et al. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE* **2017**, *12*, e0185809. [CrossRef]
21. Goulson, D. The insect apocalypse, and why it matters. *Curr. Biol.* **2019**, *29*, R967–R971. [CrossRef]
22. Page, R.D.M. DNA barcoding and taxonomy: Dark taxa and dark texts. *Philos. Trans. R. Soc. B Biol. Sci.* **2016**, *371*, 20150334. [CrossRef]
23. GBOL III: Dark Taxa—Dark Taxa. Available online: https://bolgermany.de/home/gbol3/ (accessed on 16 December 2021).
24. German Barcode of Life. Available online: https://bolgermany.de (accessed on 8 October 2021).
25. Hausmann, A.; Haszprunar, G.; Hebert, P.D.N. DNA barcoding the geometrid fauna of Bavaria (Lepidoptera): Successes, surprises, and questions. *PLoS ONE* **2011**, *6*, e17134. [CrossRef]
26. Hendrich, L.; Morinière, J.; Haszprunar, G.; Hebert, P.D.N.; Hausmann, A.; Köhler, F.; Balke, M. A comprehensive DNA barcode database for Central European beetles with a focus on Germany: Adding more than 3500 identified species to BOLD. *Mol. Ecol. Resour.* **2015**, *15*, 795–818. [CrossRef] [PubMed]
27. Kotrba, M. The DNA barcoding project on German Diptera: An appreciative and critical analysis with four suggestions for improving the development and reliability of DNA-based identification. *Eur. J. Entomol.* **2020**, *117*, 315–327. [CrossRef]
28. Schacht, W. Katalog der Zweiflügler (Mücken und Fliegen) Bayerns Familienliste—Artenliste—Literaturverzeichnis—Bestimmungsliteratur (Insecta: Diptera). Available online: http://www.zsm.mwn.de/docs_zsm/htdocs/dip/DiptBayKat.pdf (accessed on 8 October 2021).
29. Wolff, D.; Gebel, M.; Geller-Grimm, F. *Die Raubfliegen Deutschlands*; Quelle & Meyer: Wiebelsheim, Germany, 2018.
30. Jentzsch, M. Erster Entwurf einer Bibliographie zur Waffenfliegen-Fauna Deutschlands (Diptera: Stratiomyidae). *Studia Dipterol.* **2015**, *22*, 187–197.
31. Doczkal, D.; Claußen, C.; Ssymank, A. Erster Nachtrag und Korrekturen zur Checkliste der Schwebfliegen Deutschlands (Diptera, Syrphidae). *Volucella* **2002**, *6*, 167–173.
32. Heiss, R.; Fahldieck, M.; Stuke, J.-H. Kommentierte Checkliste der Schnaken der Bundesländer Niedersachsen und Bremen (Diptera, Tipulidae). *Entomol. Nachr. Ber.* **2019**, *63*, 245–259.
33. Heiss, R.; Merkel-Wallner, G. Ein Beitrag zur Schnaken-Fauna Bayerns—Neue und wenig bekannte Arten aus Malaisefallen-Fängen 2007 bis 2009. *Beiträge Bayer. Entomofaunist.* **2013**, *12*, 17–30.
34. Menzel, F. Die Trauermücken—Fauna der Bundesrepublik Deutschland (Diptera: Sciaridae). *Beiträge Entomol.* **2000**, *50*, 317–355. [CrossRef]
35. Drees, M. Die Fleischfliegen des Hagener Raumes (Diptera: Sarcophagidae). *Entomol. Z.* **2014**, *124*, 3.
36. Stuke, J.-H. Eine Kritische Liste der aus Deutschland nachgewiesenen Ephydridae mit der Beschreibung einer neuen Art (Diptera). *Entomol. Z.* **2011**, *121*, 115–126.
37. Stuke, J.-H. Die Fruchtfliegen Niedersachsens und Bremens (Diptera, Drosophilidae). *Linz. Biol. Beiträge* **2020**, *52*, 475–499.
38. Skuhravá, M.; Skuhravý, V.; Meyer, H. Gall midges (Diptera: Cecidomyiidae: Cecidomyiinae) of Germany—Faunistics, ecology and zoogeography. *Faun.-Ökol. Mitt.* **2014**, Suppl. 38, 1–200.
39. Global Malaise Program. Available online: https://biodiversitygenomics.net/projects/gmp/ (accessed on 8 October 2021).
40. Chan-Canché, R.; Ballina-Gómez, H.; Leirana-Alcocer, J.; Bordera, S.; González-Moreno, A. Sampling of parasitoid Hymenoptera: Influence of the height on the ground. *J. Hymenopt. Res.* **2020**, *78*, 19–31. [CrossRef]
41. Resources of CCDB. Available online: https://ccdb.ca/resources/ (accessed on 8 October 2021).
42. Hebert, P.D.N.; Braukmann, T.W.A.; Prosser, S.W.J.; Ratnasingham, S.; deWaard, J.R.; Ivanova, N.V.; Janzen, D.H.; Hallwachs, W.; Naik, S.; Sones, J.E.; et al. A sequel to Sanger: Amplicon sequencing that scales. *BMC Genom.* **2018**, *19*, 219. [CrossRef]
43. Ratnasingham, S.; Hebert, P.D.N. A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. *PLoS ONE* **2013**, *8*, e66213. [CrossRef]

44. Chao, A.; Chiu, C.-H. Estimation of species richness and shared species richness. In *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*; Balakrishnan, N., Ed.; Wiley: New York, NY, USA, 2012; pp. 76–111.
45. Chao, A.; Gotelli, N.J.; Hsieh, T.C.; Sander, E.L.; Ma, K.H.; Colwell, R.K.; Ellison, A.M. Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **2014**, *84*, 45–67. [CrossRef]
46. Chao, A.; Ma, K.H.; Chiu, T.C.H. Spade R: Species-richness prediction and diversity estimation with R. Available online: https://cran.r-project.org/web/packages/SpadeR/SpadeR.pdf (accessed on 16 December 2021).
47. Haszprunar, G. Barcoding Fauna Bavarica—Eine Chance für die Entomologie. *Nachr. Bayer. Entomol.* **2009**, *58*, 4.
48. Büro für Freilandökologie Dr. Jürgen Esser. Available online: https://freilandoekologie-esser.de/ (accessed on 16 December 2021).
49. Hebert, P.D.N.; Ratnasingham, S.; Zakharov, E.V.; Telfer, A.C.; Levesque-Beaudin, V.; Milton, M.A.; Pedersen, S.; Jannetta, P.; deWaard, J.R. Counting animal species with DNA barcodes: Canadian insects. *Philos. Trans. R. Soc. B Biol. Sci.* **2016**, *371*, 20150333. [CrossRef]
50. Ratnasingham, S.; Hebert, P.D.N. BOLD: The Barcode of Life Data System. *Mol. Ecol. Notes* **2007**, *7*, 355–364. Available online: http://www.barcodinglife.org (accessed on 16 December 2021). [CrossRef]
51. Hebert, P.D.N.; deWaard, J.R.; Landry, J.-F. DNA barcodes for 1/1000 of the animal kingdom. *Biol. Lett.* **2010**, *6*, 359–362. [CrossRef]
52. Reid, B.N.; Le, M.; McCord, W.P.; Iverson, J.B.; Georges, A.; Bergmann, T.; Amato, G.; Desalle, R.; Naro-Maciel, E. Comparing and combining distance-based and character-based approaches for barcoding turtles. *Mol. Ecol. Resour.* **2011**, *11*, 956–967. [CrossRef]
53. Chimeno, C.; Morinière, J.; Podhorna, J.; Hardulak, L.; Hausmann, A.; Reckel, F.; Grunwald, J.E.; Penning, R.; Haszprunar, G. DNA barcoding in forensic entomology—Establishing a DNA reference library of potentially forensic relevant arthropod species. *J. Forensic Sci.* **2019**, *64*, 593–601. [CrossRef] [PubMed]
54. Raupach, M.J.; Astrin, J.J.; Hannig, K.; Peters, M.K.; Stoeckle, M.Y.; Wägele, J.-W. Molecular species identification of Central European ground beetles (Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Front. Zool.* **2010**, *7*, 26. [CrossRef] [PubMed]
55. Pamilo, P.; Viljakainen, L.; Vihavainen, A. Exceptionally high density of NUMTs in the honeybee genome. *Mol. Biol. Evol.* **2007**, *24*, 1340–1346. [CrossRef] [PubMed]
56. Song, H.; Buhay, J.E.; Whiting, M.F.; Crandall, K.A. Many Species in One: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13486–13491. [CrossRef]
57. Buhay, J.E. "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. *J. Crustacean Biol.* **2009**, *29*, 96–110. [CrossRef]
58. Hazkani-Covo, E.; Zeller, R.M.; Martin, W. Molecular poltergeists: Mitochondrial DNA copies (Numts) in sequenced nuclear genomes. *PLoS Genet.* **2010**, *6*, e1000834. [CrossRef]
59. Nesi, N.; Nakouné, E.; Cruaud, C.; Hassanin, A. DNA barcoding of African fruit bats (Mammalia, Pteropodidae). the mitochondrial genome does not provide a reliable discrimination between *Epomophorus gambianus* and *Micropteropus pusillus*. *Comptes Rendus Biol.* **2011**, *334*, 544–554. [CrossRef]
60. Ermakov, O.A.; Simonov, E.; Surin, V.L.; Titov, S.V.; Brandler, O.V.; Ivanova, N.V.; Borisenko, A.V. Implications of hybridization, Numts, and overlooked diversity for DNA barcoding of Eurasian ground squirrels. *PLoS ONE* **2015**, *10*, e0117201. [CrossRef]
61. Dufresnes, C.; Berroneau, M.; Dubey, S.; Litvinchuk, S.N.; Perrin, N. The effect of phylogeographic history on species boundaries: A comparative framework in Hyla tree frogs. *Sci. Rep.* **2020**, *10*, 5502. [CrossRef]
62. Kmiec, B.; Woloszynska, M.; Janska, H. Heteroplasmy as a Common State of Mitochondrial Genetic Information in Plants and Animals. *Curr. Genet.* **2006**, *50*, 149–159. [CrossRef]
63. Dobson, S.L. Evolution of Wolbachia cytoplasmic incompatibility. *Evolution* **2007**, *58*, 2156–2166. [CrossRef] [PubMed]
64. Duron, O.; Bouchon, D.; Boutin, S.; Bellamy, L.; Zhou, L.; Engelstädter, J.; Hurst, G.D. the diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. *BMC Biol.* **2008**, *6*, 27. [CrossRef]
65. Hausmann, A.; Godfray, H.C.J.; Huemer, P.; Mutanen, M.; Rougerie, R.; van Nieukerken, E.J.; Ratnasingham, S.; Hebert, P.D.N. Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) System. *PLoS ONE* **2013**, *8*, e84518. [CrossRef]
66. The Chironomid Home Page. Available online: www.chironomidae.net (accessed on 16 December 2021).
67. Karlsson, D.; Forshage, M.; Holston, K.; Ronquist, F. The data of the Swedish Malaise Trap Project, a countrywide inventory of Sweden's insect fauna. *Biodivers. Data J.* **2020**, *8*, e56286. [CrossRef] [PubMed]
68. Moeed, A.; Meads, M.J. Seasonality of arthropods caught in a malaise trap in mixed lowland forest of the Orongorongo Valley, New Zealand. *N. Z. J. Zool.* **1987**, *14*, 197–208. [CrossRef]
69. Matthews, R.; Matthews, J. The Malaise trap: Its utility and potential for sampling insect populations. *Great Lakes Entomol.* **2017**, *4*, 4.

Publication IV:

Facing the Infinity: Tackling large samples of challenging Chironomidae (Diptera) with an integrative approach

Published in PeerJ

Chimeno, C., Björn, R., Manfrin, A., Kalinkat, G., Hölker, F. & Baranov, V. (2023). Facing the Infinity: Tackling large samples of challenging Chironomidae (Diptera) with an integrative approach. *PeerJ*, 11: e15336. 10.7717/peerj.15336

Facing the infinity: tackling large samples of challenging Chironomidae (Diptera) with an integrative approach

Caroline Chimeno[1], Björn Rulik[2], Alessandro Manfrin[3], Gregor Kalinkat[4], Franz Hölker[4] and Viktor Baranov[5]

[1] Bavarian State Collection of Zoology (SNSB-ZSM), Munich, Germany
[2] Zoological Research Museum Alexander Koenig, Leibniz Institute for the Analysis of Biodiversity Change (LIB), Bonn, Germany
[3] Institute for Environmental Sciences, iES Landau, RPTU University of Kaiserslautern-Landau, Landau, Germany
[4] Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany
[5] Estación Biológica de Doñana-CSIC/Doñana Biological Station-CSIC, Seville, Spain

ABSTRACT

Background: Integrative taxonomy is becoming ever more significant in biodiversity research as scientists are tackling increasingly taxonomically challenging groups. Implementing a combined approach not only guarantees more accurate species identification, but also helps overcome limitations that each method presents when applied on its own. In this study, we present one application of integrative taxonomy for the highly abundant and particularly diverse fly taxon Chironomidae (Diptera). Although non-biting midges are key organisms in merolimnic systems, they are often cast aside in ecological surveys because they are very challenging to identify and extremely abundant.
Methods: Here, we demonstrate one way of applying integrative methods to tackle this highly diverse taxon. We present a three-level subsampling method to drastically reduce the workload of bulk sample processing, then apply morphological and molecular identification methods in parallel to evaluate species diversity and to examine inconsistencies across methods.
Results: Our results suggest that using our subsampling approach, identifying less than 10% of a sample's contents can reliably detect >90% of its diversity. However, despite reducing the processing workload drastically, the performance of our taxonomist was affected by mistakes, caused by large amounts of material.
We conducted misidentifications for 9% of vouchers, which may not have been recovered had we not applied a second identification method. On the other hand, we were able to provide species information in cases where molecular methods could not, which was the case for 14% of vouchers. Therefore, we conclude that when wanting to implement non-biting midges into ecological frameworks, it is imperative to use an integrative approach.

Subjects Biodiversity, Entomology, Molecular Biology, Taxonomy, Freshwater Biology
Keywords Integrative taxonomy, Chironomidae, Biodiversity, DNA barcoding, Megadiverse, Bulk sample processing, Non biting midges, Molecular species identification, Morphological species identification, Freshwater

Submitted 10 January 2023
Accepted 11 April 2023
Published 22 May 2023

Corresponding author
Caroline Chimeno,
chimeno@snsb.de

Academic editor
Paul Ayayee

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj.15336**

Copyright
2023 Chimeno et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Chimeno C, Rulik B, Manfrin A, Kalinkat G, Hölker F, Baranov V. 2023. Facing the infinity: tackling large samples of challenging Chironomidae (Diptera) with an integrative approach. **PeerJ 11:e15336 DOI 10.7717/peerj.15336**

INTRODUCTION

Chironomidae (non-biting midges) is by far the most ecomorphologically diverse and widely distributed ingroup of aquatic insects (*Hilsenhoff, Thorp & Covich, 2001*; *Armitage, Pinder & Cranston, 2012*). Occurring in every zoogeographic region, including Antarctica, non-biting midges inhabit nearly all aquatic and semiaquatic, marine and terrestrial habitats (*Armitage, Pinder & Cranston, 2012*). Characteristic behavioral and physiological adaptations have enabled these flies to colonize extreme environments such as caves up to 1,000 m deep, hot springs, high-altitude waters, glacial streams, and even highly polluted waters or sewage systems (*Andersen et al., 2016*; *Gadawski et al., 2022*). In aquatic systems, their abundance can be higher than that of all other macroinvertebrates combined, making them a keystone taxon in freshwater ecology (*Gratton & Zanden, 2009*; *Marziali et al., 2010*; *Karima, 2021*). The bottom-dwelling larvae not only represent almost every feeding group but, being ecosystem engineers, they also contribute enormously to sediment- and water-mixing, and to the global oxygen- and carbon-cycle (*Hölker et al., 2015*; *Baranov, Lewandowski & Krause, 2016*; *Antczak-Orlewska et al., 2021*). As ecosystem engineers, the Chironomidae are involved in modifying the availability of nutrients (chiefly phosphorous, but also nitrogen), as well as oxygen and carbon availability for other aquatic organisms (*Hölker et al., 2015*; *Baranov, Lewandowski & Krause, 2016*). All life stages (even the short-lived adults) play a vital role in aquatic and terrestrial food webs, serving as an important food source for fish, birds, bats and other arthropods (*Gratton & Zanden, 2009*; *Raunio, Heino & Paasivirta, 2011*; *Armitage, Pinder & Cranston, 2012*; *Wirta et al., 2015*; *Herren et al., 2017*). This combination of high ecosystem functionality, high abundance, and habitat specificity of the Chironomidae to their environment makes them suitable biological indicators for ecological assessments (*e.g.*, water quality control) (*Sæther, 1977*; *Lencioni, Marziali & Rossaro, 2012*; *Dorić et al., 2021*).

Despite this, only a limited subset of biodiversity studies or biomonitoring surveys of aquatic habitats incorporate species- or genus-level information of the Chironomidae and oftentimes, they are neglected altogether (*Raunio, Heino & Paasivirta, 2011*; *Dorić et al., 2021*). This is due to several factors: (i) non-biting midges are relatively difficult to identify (*Cranston, 2008*; *Proulx et al., 2013*), (ii) only few taxonomists with the required expertise are available for species-level identification (*Cranston et al., 2013*; *Chan et al., 2014*), (iii) traditional morphological-based species delimitations often require laborious dissection and mounting of specimens on microscope slides (*Ekrem, Stur & Hebert, 2010*; *Gadawski et al., 2022*), and (iv) they can be extremely species rich even in relative low-diversity temperate and boreal ecosystems (*Lundström et al., 2010*). The workload associated with the processing of non-biting midges from large bulk samples, common in ecological surveys, is immense when applying traditional identification methods (*Rosenberg, 1992*; *Brodin et al., 2012*). In humid climates, or during wetter years, the number of specimens to be processed can increase from hundreds of thousands to sometimes millions of specimens.

There are few methods that can help overcome the pitfall of processing an "infinite" number of specimens, with the most obvious one (and most resource-demanding) being

the employment of more taxonomists or parataxonomists (*Engel et al., 2021*) to help accelerate specimen processing and identification. The availability of expert taxonomists, however, is in decline and even then, financing such manpower at a large scale is often not feasible and remains time-consuming (*Hausmann et al., 2020*; *Chimeno et al., 2022*). Therefore, researchers often subsample bulk samples to reduce the sorting effort, or limit sample processing to a few key families or species (*Mandelik, Roll & Fleischer, 2010*; *Porter et al., 2014*; *Keck et al., 2017*; *Bohan et al., 2017*; *Chimeno et al., 2023*). One promising alternative that is currently in development is the use of automatic machine-based identification approaches for species identification (see *Milošević et al., 2020*). As demonstrated by Milošević and authors, after vigorously training their artificial neural network on 1,836 specimens belonging to ten similar-looking species of Chironomidae, they recovered 99% identification success when presenting their network new images. Despite these promising results, this technology is not yet applicable at a large scale because it requires laborious sample preparation and a vigorous training-phase of the target taxa (*Milošević et al., 2020*).

Currently, one of the most common and promising methodologies for large-scale species identification is DNA barcoding, a molecular-based identification method (*Brodin et al., 2012*; *Morinière et al., 2016*). It uses a short DNA fragment to differentiate species from one another, and does so at a lower cost and faster pace than traditional morphological methods (*Hebert et al., 2003*; *Ekrem, Willassen & Stur, 2007*; *Porter et al., 2014*; *Morinière et al., 2016*). With the rise of DNA barcoding, high quality species-level information of Chironomidae is increasingly becoming more accessible to research (*Ekrem, Stur & Hebert, 2010*; *Baloğlu, Clews & Meier, 2018*), and studies examining the efficiency of this method in research of these insects reveal an overall congruence of 80–90%, making it a great complement to taxonomic methodologies (*Carew, Pettigrove & Hoffmann, 2005*; *Pfenninger et al., 2007*; *Ekrem, Willassen & Stur, 2007*; *Carew et al., 2007*; *Carew, Marshall & Hoffmann, 2011*; *Lin, Stur & Ekrem, 2015*). However, just as any identification method, DNA barcoding has its own limitations (*Dayrat, 2005*; *Will, Mishler & Wheeler, 2005*; *Schlick-Steiner et al., 2010*) and therefore, numerous studies resort to applying a combined methodological approach for species identifications (*Pires & Marinoni, 2010*; *Sheth & Thaker, 2017*).

With many studies highlighting the need for a smart and efficient integration of both morphological and molecular species identification methods (*Hausmann et al., 2020*; *Hartop et al., 2022*), our study aims to present and evaluate one way to do so for a particularly diverse and complicated group of insects: the Chironomidae. To tackle the large amounts of insect material, we apply a three-level subsampling technique that we present in the Methods section. We also compare our DNA- and morphology-based species identifications in terms of accuracy, to demonstrate how the use of each method on its own can provide discrepant results. We are processing bulk samples of Diptera that have been collected in the framework of the federal-funded field experiment "Verlust der Nacht" (https://www.igb-berlin.de/projekt/verlust-der-nacht) and the follow-up project "Artenschutz durch umweltfreundliche Beleuchtung" (https://www.igb-berlin.de/projekt/artenschutz-durch-umweltvertraegliche-beleuchtung-aube) located in the Westhavelland

Nature Park in northeast Germany. The project was launched in 2012 with the goal of studying the effects that artificial lighting at night has on species communities.

MATERIALS AND METHODS

Study area and experimental design

The "Verlust der Nacht" experiment was conducted by the Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB) in a large-scale facility established in 2012 (see *Holzhauer et al. (2015)*, *Manfrin et al. (2017)* for details). The facility is located in a 750-km^2 Dark-Sky Reserve within the Westhavelland Nature Park in the Berlin-Brandenburg Metropolitan Region (https://www.darksky.org/our-work/conservation/idsp/reserves/westhavelland/). The landscape is characterized by a system of drainage ditches (approximately 5 m wide, average annual water depth 50 ± 26 cm). In the grassland adjacent to the drainage ditch, we installed three parallel rows (3, 23 and 43 m away from the drainage ditch) of four conventional 4.75 m high streetlights located 20 m apart. Each lamp post in the lit site was equipped with one 70-W high-pressure sodium lamp (VIALOX NAV-T Super 4Y, yellow 2,000 K, Osram, Munich, Germany). In the control (dark) site only the lamp posts were installed (*i.e.*, without bulbs) providing identical physical structure yet remaining dark. The lamps used in the lit site had a maximum illuminance of approximately 50 lx directly under the lamp, with the minimum illuminance between two adjacent streetlamps of the same row being approximately 10 lx, and a minimum illuminance between rows of streetlamps of ca. 1 lx (see *Holzhauer et al. (2015)* for further details about light distribution and spectral composition). From spring 2012 onward, the lit site was illuminated at night, *i.e.*, between civil twilight at dusk and dawn. The lit and control sites are very similar in their environmental characteristics (*e.g.*, water physico-chemistry, hydromorphology, riparian vegetation) and ~600 m (800 m along the drainage ditch) apart, separated by a row of trees.

Insect collection

We collected insects emerging from the drainage ditch from both lit and dark sites from May to October 2014. Emerging insects were sampled using four floating pyramidal emergence traps (0.85 m × 0.85 m, 300-µm mesh), placed in the drainage ditch *ca*. 1 m from the bank and directly in front of each streetlamp. Sampling duration ranged from seven (one night samplings) to approximately 185 h (1 week samplings) and occurred monthly except in July when the sampling was conducted twice. Flying adult non-biting midges were collected from the grassland adjacent to the drainage ditch using 24 flight interception traps, 12 at each site. Flight intercepting traps were placed 0.5 m below each lamp and consisted of two perpendicular acrylic panels (each 204 mm × 500 mm × 3 mm) mounted above a collecting funnel. The flight intercepting traps were collecting insects for one 24-h sampling period every month except in July when sampling was conducted twice. Based on astronomical sunset and sunrise, the 24-h sampling periods were always split into a night-sampling (8–14 h, depending on the season) followed by a day-sampling (10–16 h), replacing the collecting jars after each of them. Sampling always occurred on rainless days/nights within 24 h of either first- or third-quarter moon. Both emergence and flight

intercepting traps were equipped with collecting jars containing 70% ethanol as a preservative medium (see *Manfrin et al. (2017)* for further details).

Morphotype sorting and subsampling for processing

We obtained bulk samples of pre-sorted adult "Nematoceran" flies (crane flies, midges, gnats, mosquitoes *etc.*) stored in 90% ethanol that were collected in the sampling year 2014 (see "Insect collection"). From these samples, our senior author, who is a trained expert of non-biting midges, sorted specimens using a stereo microscope and grouped them into different morphotypes. To do this, we used three different approaches based on the "difficulty" of specimen sorting (Fig. 1). Large and/or conspicuous species that are easy to recognize, such as *Prodiamesa olivacea* (*Meigen, 1818*) or *Ablabesmyia phatta* (*Egger, 1863*), were quickly sorted into their own distinct morphotypes and assigned a preliminary species name. Specimens that were more difficult to group (because they belong to genera that have similar-looking representatives when viewed under the stereo microscope) were sorted at the genus-level, hence, grouped into genera-morphotypes if possible. Hence, if several genera have similar-looking representatives under the stereo microscope, we sorted representatives of several genera into one morphotype. Lastly, for specimens that our expert taxonomist found difficult to address, subsets were mounted on temporary glycerol slides to be examined at ×400 magnification in a first step, so that similar specimens can be assigned to the same morphotype in a second step. From every morphotype group, we selected a representative number of morphotype voucher specimens (about 10%). For very abundant morphotypes where 10% of specimens is still too much, we sampled fewer individuals. Selected specimens were used for molecular and morphological species identifications.

Sequencing of selected specimens

For specimens larger than 2 mm, we used a single leg or leg segment as a tissue sample that was transferred to a 96-well plate. For smaller individuals, we extracted DNA non-destructively (*i.e.*, subsequent voucher recovery) from the whole body. After lysis, we extracted genomic DNA using the BioSprint96 magnetic bead extractor and the respective kits by Qiagen (Hilden, Germany). We carried out a polymerase chain reaction (PCR) in a total reaction volume of 20 µl, including 2 µl of undiluted DNA template, 0.8 µl of each primer (10 pmol/µl), 2 µl of 'Q-Solution' and 10 µl of 'Multiplex PCR Master Mix', containing hot start Taq DNA polymerase and buffers. The latter components are available in the Multiplex PCR kit by Qiagen (Hilden, Germany).

Thermal cycling was performed on GeneAmp PCR System 2,700 machines (Life Technologies, Carlsbad, CA, USA) as follows: hot start *Taq* activation: 15 min at 95 °C; first cycle set (15 repeats): 35 s denaturation at 94 °C, 90 s annealing at 55 °C (−1 °C/cycle) and 90 s extension at 72 °C. Second cycle set (25 repeats): 35 s denaturation at 94 °C, 90 s annealing at 40 °C and 90 s extension at 72 °C; final elongation 10 min at 72 °C. As established within the German Barcode of Life (GBOL) project at the ZFMK, we used the standard degenerate barcoding primers LCO1490-JJ: 5′-CHACWAAYCATAAAGATATYGG- 3′ and HCO2198-JJ: 5′-AWACTTCVGGRTGVCCAAARAATCA- 3′ (*Astrin & Stüben, 2008*).

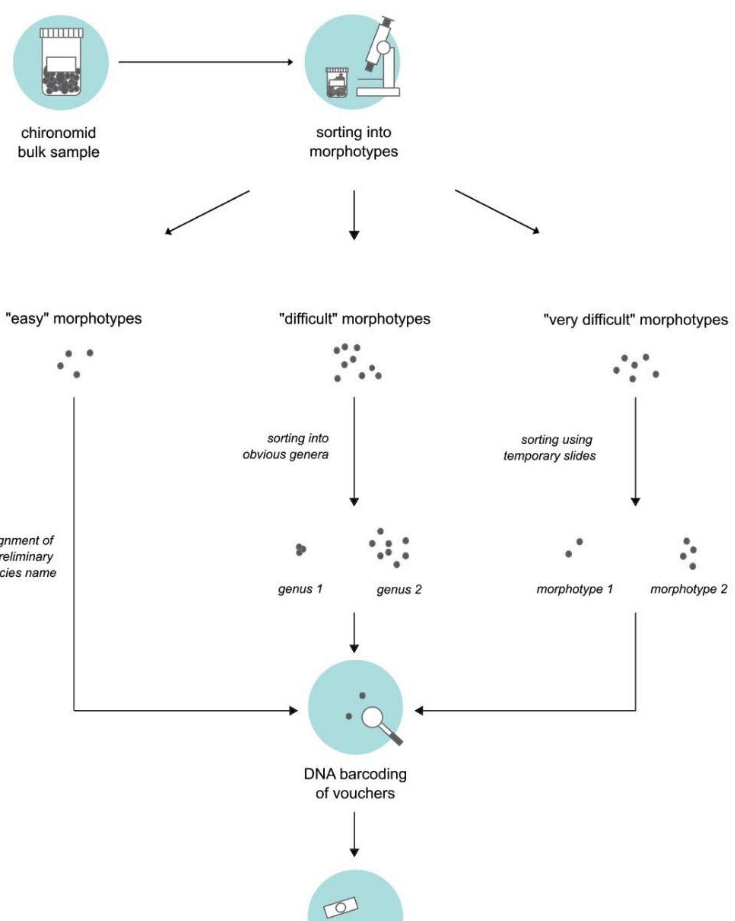


Figure 1 **Three-level sorting workflow that was used in this study for bulk sample processing.** For each morphotype distinguished in a bulk sample, we conducted morphological & molecular identifications of selected vouchers. The procedure was different based on the difficulty of the specimens involved in sorting. Full-size DOI: 10.7717/peerj.15336/fig-1

Purification and sequencing were conducted by the BGI Group (Hong Kong, China) using the amplification primers.

Traces were semi-automatically edited, then assembled sequences using the MUSCLE alignment approach (*Edgar, 2004*), and checked for the occurrence of stop-codons or hints

of nuclear mitochondrial DNA segments (NUMTs) in Geneious version 7.1.9 (http://www.geneious.com; *Kearse et al., 2012*). Further details such as voucher information, primer pairs, sequence data and trace files were deposited to BOLD and GenBank. These can be found under the following information (http://doi.org/10.5883/DS-ALANCHIR; GenBank accession numbers OP927392–OP927685).

Morphological identifications

After DNA barcoding (or in parallel, depending on whether whole specimens were transferred to plates or just tissue samples), we mounted the specimens (or their empty shells) on permanent slides in Euparal and Hydromatrix following standard procedure (*Kirk-Spriggs & Sinclair, 2017*). Morphological identifications were conducted with aid of numerous identification keys and papers covering palaearctic Chironomidae (see *Lehmann (1970)*, *Saether (1971)*, *Hirvenoja (1973)*, *Wiederholm (1989)*, *Ekrem (2002a)*, *Langton & Pinder (2007)*, *Pillot (2008)*, *Giłka (2011)*). These identifications were conducted by our senior author which has conducted various research on the taxonomy of Chironomidae (see *Baranov (2011a*, *2011b*, *2013)*, *Baranov & Perkovsky (2013)*, *Baranov & Przhiboro (2014)*, *Baranov, Andersen & Hagenlund (2015)*, *Baranov, Andersen & Perkovsky (2015)*, *Baranov, Góral & Ross (2017)*, *Baranov et al. (2019)*).

DATA ANALYSIS

All sequence records including metadata were uploaded to the online database Barcode of Life Data System (BOLD; *Ratnasingham & Hebert, 2007*). Sequences ≥300 base pairs (bp) were automatically assigned a Barcode Index Number (BIN) on BOLD if sequence similarity based on the (RESL-) BIN algorithm was fulfilled. Sequences ≥500 bp which did not find a match served as founders of new BINs. The dataset was downloaded on April 11, 2022, for analysis and can be viewed on Figshare (https://doi.org/10.6084/m9.figshare.21803013). Therefore, the present results correspond to BINs assigned at that time (BIN assignments can change as new sequences are added to BOLD). In addition to using the RESL-algorithm that is implemented into BOLD, we also applied Assemble Species by Automatic Partitioning (ASAP; *Puillandre, Brouillet & Achaz, 2021*) and SpeciesIdentifier version 1.9 (*Meier et al., 2006*) to cluster our sequences at 3%. ASAP uses pairwise genetic distances for hierarchical clustering without using information on intraspecific diversity, and SpeciesIdentifier is an algorithm that allows to cluster sequences based on their pairwise intra- and interspecific genetic distances. The outputs of all three algorithms were used to compare the number of Operational Taxonomic Units (OTUs) obtained with each and comparing diversity assessments. To compare all methodologies, we created a Neighbor-Joining in MEGA11 (version 11.0.13) of all sequence data and added morphological species-, ASAP-, RESL-, and SpeciesIdentifier labels (Data S1). Because all depict similar performance (see results), subsequent taxonomic analyses were conducted only using the RESL outputs.

To assess our sampling effort, we calculated Chao1 and Chao2 estimates using the *ChaoSpecies* function of the *SpadeR* package (version 0.1.1; *Chao et al., 2016*) in R (version 4.2.1) on abundance and incidence data, respectively (Data S2). We did this to estimate the

species diversity at the sampling site and to compare it to that which was empirically observed in our samples. Then, we used the *iNEXT* function from the *iNEXT* package (version 3.0.0; *Hsieh, Ma & Chao, 2016*; *Hsieh, Ma & Chao, 2020*) to extrapolate the species diversity obtained with each methodology (morphology, RESL, ASAP, and SpeciesIdentifier) to double the sampling effort. To depict the species diversity recovered per morphotype, we created accumulation curves using the *iNEXT* function on results derived from each identification method (morphological and molecular).

To double-check our identifications and to recover possible misidentifications, we created a dataset from BOLD containing 19,525 public COI-sequences of 1,035 species of non-biting midges collected throughout Europe (Data S3). We applied the following selection criteria to build a neighbor-joining tree: Kimura 2 Parameter distance model, sequences ≥200 bp, and excluding contaminants, records flagged with stop codons, and records flagged as misidentifications. To facilitate review, we colored the tree based on barcode clusters (BINs). We added the names of identifiers along with the identification method to each entry to discriminate high-level taxonomists that used morphological methods to vouchers from parataxonomists relying on the BOLD engine for sequence identification. We considered expert identifications as those conducted by researchers with taxonomic experience of Chironomidae, such as Elisabeth Stur (Norwegian University of Science and Technology; Norway; see *Stur & Ekrem (2000*, *2006*, *2011*, *2015)*, *Stur & Wiedenbrug (2005)*, *Stur & Spies (2011)*), Torbjørn Ekrem (Norwegian University of Science and Technology, Norway; see *Ekrem (2002a*, *2002b*, *2007)*, *Ekrem & Stur (2009)*, *Ekrem, Stur & Hebert (2010)*), Yngve Brodin (Swedish Museum of Natural History, Sweden; see *Brodin, Lundström & Paasivirta (2008)*, *Siri & Brodin (2014)*), Piotr Gadawski (University of Lodz; Poland; see *Gadawski et al. (2022)*, *Giłka & Gadawski (2022)*,), and Sophie Wiedenbrug (University of São Paulo, Brazil) (see *Wiedenbrug, Lamas & Trivinho-Strixino (2012*, *2013)*, *Silva & Wiedenbrug (2015)*, *Wiedenbrug & Silva (2016)*).

RESULTS

Identification of specimens

Overall, we sorted through 4,549 specimens of non-biting midges which made up the bulk (99.6%) of "Nematoceran" specimens in our samples. We recovered 48 morphotype groups, and in total selected 331 specimen-vouchers, of which more than half were females (Data S2).

Molecular identifications

We applied DNA barcoding to all 331 specimens and obtained 315 COI-barcodes (95%) that we uploaded to BOLD. Five sequences contained cross contaminations, and another 16 were identified as not being non-biting midges, but species of the taxa Anisopodidae, Chaoboridae, Culicidae, Hybotidae, Psychodidae, Sciaridae, and Trichoceridae.

The remaining COI-sequences were clustered into 77 BINs which provided coverage for 55 species and four interim species (essentially being morphotype analogs that are widely used in ecological studies) (*Ablabesmyia sp. 2ES*, *Smittia sp. 8ES*, *Smittia sp. 14ES*, and *Thienemanniella sp. 3TE*). Interim species names are assigned on BOLD when molecular

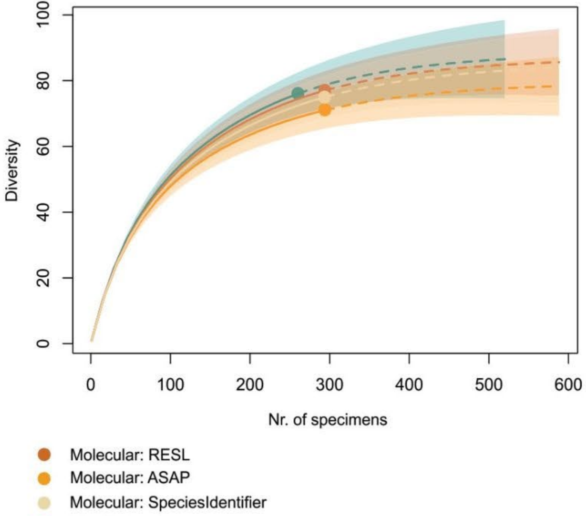


Figure 2 Accumulation curves of species and clusters recovered across methods. Dotted lines represent extrapolated values (for up to double the sampling effort), bold lines represent interpolated values. Accumulation curves show the number of morphologically identified species and that of clusters recovered with RESL, ASAP, and SpeciesIdentifier. Full-size DOI: 10.7717/peerj.15336/fig-2

analysis detects genetic differences, but no species name can be provided due to the lack of a taxonomic revision or of formal species description (*Stur & Ekrem, 2011*; *Morinière et al., 2016*). Seven BINs did not provide conclusive species-level identification and five BINs did not match to public data, providing no molecular identification. In five cases, two BINs were assigned to the same species (*Cladopelma viridulum*—BOLD:AAD7363 and BOLD: AAV3586; *Polypedilum cultellatum*—BOLD:AAH7761 and BOLD:ACX5929; *Polypedilum sordens*—BOLD:ACY3855 and BOLD:ADF3485; *Smittia stercoraria*—BOLD:AAN5358 and BOLD:AAN5355; *Smittia sp. 14ES* —BOLD:AAM7064 and BOLD:ACW5117). Data S2 provides an overview of the entire dataset.

We applied two other clustering algorithms (SpeciesIdentifier and ASAP) to our COI data. Although both SpeciesIdentifier (using 3% threshold) and ASAP (1st partition) did suggest slightly fewer clusters than the RESL-algorithm, all derived species diversities fall into the 95% confidence interval (Fig. 2), and the results were largely consistent across methods (Tables 1–3, Figs. 3B–3C).

Morphological identifications

Using morphological methods, we identified a total of 76 species. A total of 34 specimens were left unidentified at a higher taxonomic level: 22 at the genus-, and 12 at the family-level.

Table 1 Chao1/2 estimates and iNEXT extrapolation values across methods.		
Method/Algorithm	**Output**	**Values**
Morphology	Sample size (n)	260
	Number of tax. entities	76
	Number of rare entities	44
	Sample coverage	0.91
	Chao1 estimate	89 ± 7 SE
	iNEXT extrapolation (2n)	87 ± 12 SE
	Chao2 estimate	109 ± 15 SE
	Jackknife SE/bias	0.0036/0
Molecular: RESL	Sample size (n)	294
	Number of clusters	77
	Number of rare clusters	40
	Sample coverage	0.93
	Chao1 estimate	87 ± 6 SE
	iNEXT extrapolation (2n)	86 ± 10 SE
	Chao2 estimate	100 ± 11 SE
	Jackknife SE/bias	$0.0039/-2.3502e^{-14}$
Molecular: ASAP	Sample size (n)	294
	Number of clusters	71
	Number of rare clusters	34
	Sample coverage	0.94
	Chao1 estimate	79 ± 5 SE
	iNEXT extrapolation (2n)	78 ± 9 SE
	Chao2 estimate	92 ± 11
	Jackknife SE/bias	0.0042/0
Molecular: SpeciesIdentifier	Sample size (n)	294
	Number of clusters	75
	Number of rare clusters	39
	Sample coverage	0.93
	Chao1 estimate	85 ± 6 SE
	iNEXT extrapolation (2n)	84 ± 11 SE
	Chao2 estimate	98 ± 11
	Jackknife SE/bias	0.0040/0

Note:
Results after applying Chao1 and Chao2 biodiversity calculations to each datatype (morphological; molecular: RESL, ASAP, SpeciesIdentifier), including sample sizes (Nr. of specimens), taxonomical entities (Nr. of species for morphological data; clusters for molecular data), sample coverage, Chao1 and Chao2 estimates, jackknife validations, and extrapolations to double the sample size.

Assessing our sampling effort

Chao1 species richness estimates suggest that 79 ± 5 to 89 ± 7 species may have been present in the community that we sampled (Table 1). Sample-based Chao2 estimates were slightly higher, suggesting 92 ± 11 to 109 ± 15 species. Extrapolation to double the sampling effort would have increased the number of recovered entities by 11–17% (Fig. 2). Sample coverage was above 90% for all data (morphology, RESL, ASAP, SpeciesIdentifier).

Table 2 Cases of discrepancies between morphological and molecular-based identifications.

Discrepancy	Morphotype	Nr. of sequences	Morphological ID of specimen	BIN	Molecular ID linked to BIN
Type 1	*"Acricotopus lucens"*	2	*Acricotopus lucens*	BOLD:AAG5487	*Procladius crassinervis*
	"Chironomus"	1	*Chironomus plumosus*	BOLD:ACT6966	*Chironomus obtusidens*
	"Chironomus"	1	*Chironomus prasinatus*	BOLD:AAU4046	*Chironomus annularius*
	"Chironomus"	1	*Chironomus sp.*	BOLD:ADF1214	*Benthalia carbonaria*
	"Dicrotendipes"	1	*Dicrotendipes tritomus*	BOLD:AAU1021	*Dicrotendipes nervosus*
	"Endochironomus"	2	*Endochironomus albipennis*	BOLD:AAW5643	*Endochironomus tendens*
	"Endochironomus"	1	*Endochironomus stackelbergi*	BOLD:AAW5643	*Endochironomus tendens*
	"Glyptotendipes"	1	*Glyptotendipes cauliginellus*	BOLD:ACD4470	*Glyptotendipes pallens*
	"Glyptotendipes"	1	*Glyptotendipes glaucus*	BOLD:ACD4470	*Glyptotendipes pallens*
	"Glyptotendipes"	1	*Glyptotendipes glaucus*	BOLD:AAC0597	*Glyptotendipes paripes*
	"Parachironomus"	3	*Parachironomus gracilior*	BOLD:ACY5073	*Parachironomus monochromus*
	"Paratanytarsus/Rheotanytarsus"	1	*Paratanytarsus laetipes*	BOLD:AAI6018	*Cricotopus bicinctus*
	"Procladius ferrugineus"	2	*Procladius ferrugineus*	BOLD:AAG5487	*Procladius crassinervis*
	"Procladius pectinatus"	1	*Procladius pectinatus*	BOLD:ACW5385	*Procladius culiciformis*
	"Pseudosmittia obtusa"	1	*Pseudosmittia obtusa*	BOLD:ACP4407	*Pseudosmittia trilobata*
	"Smittia aterrima"	2	*Smittia aterrima*	BOLD:AAN5358	*Smittia stercoraria*
	"Tanypus punctipennis"	1	*Tanypus punctipennis*	BOLD:ADJ7832	*Tanypus kraatzi*
	"Tanytarsus"	1	*Tanytarsus reei*	BOLD:ACF7553	*Tanytarsus heusdensis*
	"Tanytarsus"	2	*Tanytarsus dispar*	BOLD:ACG9929	*Tanytarsus medius*
	"Xenopelopia nigricans"	1	*Xenopelopia nigricans*	BOLD:ADJ7832	*Tanypus kraatzi*
Type 2	*"Ablabesmyia phatta"*	1	*Ablabesmyia phatta*	BOLD:ACK3818	*Ablabesmyia sp. 2ES*
	"Chironomidae"	12	*Chironomidae sp.*	BOLD:AAC0597	*Glyptotendipes paripes*
	"Cladopelma/Cryptochironomus/ Harnischia"	1	*Cladopelma sp.*	BOLD:AAV3586	*Cladopelma viridulum*
	"Cladopelma/Cryptochironomus/ Harnischia"	1	*Cladopelma sp.*	BOLD:AAV8096	*Cladopelma virescens*
	"Endochironomus"	9	*Endochironomus sp.*	BOLD:AAW5643	*Endochironomus tendens*
	"Glyptotendipes"	1	*Glyptotendipes sp.*	BOLD:ACD4470	*Glyptotendipes pallenses*
	"Psectrocladius"	1	*Psectrocladius sp.*	BOLD:AAU0273	*Psectrocladius limbatellus*
	"Smittia terrestris"	2	*Smittia terrestris*	BOLD:ACP4736	Interim species *Smittia sp. 8ES*
	"Smittia terrestris"	7	*Smittia terrestris*	BOLD:ACW5117	Interim species *Smittia sp. 14ES*
	"Thienemanniella"	1	*Thienemanniella vittata*	BOLD:AAV3048	Interim species *Thienemanniella sp. 3TE*
Type 3	*"Acricotopus lucens"*	1	*Acricotopus lucens*	BOLD:AEO5089	No public data
	"Chironomus"	6	*Chironomus curabilis*	BOLD:ACD8415	*Chironomus curabilis/nuditarsis*
	"Cricotopus"	2	*Cricotopus sp.*	BOLD:AEO5089	No public data
	"Cricotopus"	3	*Cricotopus sylvestris*	BOLD:AAA5299	*Cricotopus sylvestris/glacialis*
	"Cricotopus"	1	*Cricotopus tricinctus*	BOLD:AEG4456	*Cricotopus tricinctus/sylvestris/ trifasciatus*

(Continued)

Table 2 (continued)

Discrepancy	Morphotype	Nr. of sequences	Morphological ID of specimen	BIN	Molecular ID linked to BIN
	"Cricotopus"	9	*Cricotopus sylvestris*	BOLD:AAA5299	*Cricotopus sylvestris/glacialis*
	"Glyptotendipes"	4	*Glyptotendipes cauliginellus*	BOLD:AAF8348	*Glyptotendipes cauliginellus/ lobiferus*
	"Metriocnemus"	1	*Metriocnemus sp.*	BOLD:ADV3586	No public data
	"Microtendipes chloris"	5	*Microtendipes chloris*	BOLD:ACY5270	*Microtendipes pedellus/chloris*
	"Parachironomus"	1	*Parachironomus sp.*	BOLD:ADV3586	No public data
	"Procladius crassinervis"	4	*Procladius crassinervis*	BOLD:ACB6320	*Procladius sp.*
	"Psectrocladius oxyura"	1	*Psectrocladius oxyura*	BOLD:AEO4348	No public data
	"Tanytarsus usmaensis"	2	*Tanytarsus usmaensis*	BOLD:AEO0788	No public data

Note:
Morphotypes, number of sequences, and identifications that were involved in discrepant results, namely complete incongruences in identification across methods (type 1), molecular methods provided more species-level information than morphology (type 2), and Morphology provided more species-level information while molecular methods provided inconclusive or no identification at all (type 3).

Discrepancies between morphology- and DNA-based identifications

Overall, we recovered discrepant identifications among 103 specimens (Table 2), and categorized them as follows:

Type 1: Cases with complete incongruence in identifications across methods (27 specimens).

Type 2: Molecular methods provided higher taxonomic resolution than morphology (36 specimens).

Type 3: Morphology provided higher taxonomic resolution while molecular methods provided inconclusive or no identification at all (40 specimens).

Meticulous revision of our molecular and morphological data revealed that all type-1 discrepancies were caused by misidentifications that were performed by the senior author (Viktor Baranov), which involves 9% of all voucher specimens. For another 9% of vouchers, morphological identifications could not provide identifications at the species-level (type-2), meaning that for a total of 18% of vouchers, morphology did not provide accurate or comprehensive species-level identifications.

On the other hand, morphological identification methods did provide more comprehensive species information for a total of 40 specimens (14%). Here, we were able to provide species-level IDs for five BINs that did not provide public data on BOLD, and for six BINs that were linked to discrepant identifications by taxonomists.

Uncovering species diversity from morphotypes

Of the 48 morphotypes that we distinguished during sorting, we identified 77 species (including misidentifications) using morphology and 78 BINs using molecular methods (Table 3). The most abundant (and thus higher sampled) morphotypes within our samples were "MT Glyptotendipes", "MT Parachironomus", "MT Paratanytarsus/Rheotanytarsus", "MT Cladopelma/Cryptochironomus/Harnischia", and "MT Cricotopus". These morphotypes encompass 42% (125) of all analyzed specimens. Species identification,

Table 3 Overview of all analysed specimens of Chironomidae. Number of specimens, morphologically identified species, BINs, ASAP- and SpeciesIdentifier OTUs recovered per morphotype.

Morphotype	Specimens	Morph. identified species	BINs	ASAP	SP-ID
"Ablabesmyia longistyla"	1	1	1	1	1
"Ablabesmyia monilis"	2	1	1	1	1
"Ablabesmyia phatta"	2	1	1	1	1
"Acricotopus lucens"	**3**	**1**	**2**	**2**	**2**
"Benthalia"	1	1	1	1	1
"Chironomidae"	12	0	1	1	1
"Chironomus"	**14**	**5**	**6**	**5**	**6**
"Cladopelma/Cryptochironomus/Harnischia"	22	5	6	5	6
"Coryneura"	6	2	2	2	2
"Cricotopus"	20	4	5	5	5
"Dicrotendipes"	**5**	**2**	**1**	**1**	**1**
"Endochironomus"	**12**	**2**	**1**	**1**	**1**
"Glyptotendipes"	**32**	**5**	**4**	**4**	**4**
"Guttipelopia guttipennis"	11	1	1	1	1
"Kiefferulus tendipediformis"	4	1	1	1	1
"Metriocnemus atriclava"	1	1	1	1	1
"Metriocnemus"	2	1	1	1	1
"Microchironomus"	5	1	1	1	1
"Microtendipes chloris"	5	1	1	1	1
"Microtendipes pedellus"	1	1	1	1	1
"Nanocladius dichromus"	1	1	1	1	1
"Orthocladius oblidens"	2	1	1	1	1
"Parachironomus"	**26**	**4**	**5**	**5**	**5**
"Paraphaenocladius impensus"	2	1	1	1	1
"Paratanytarsus/Rheotanytarsus"	**25**	**6**	**7**	**6**	**7**
"Polypedilum sordens"	6	1	2	2	2
"Polypedilum"	10	2	3	2	3
"Procladius crassinervis"	10	1	2	2	2
"Procladius culiciformis"	5	1	1	1	1
"Procladius ferrugineus"	**2**	**1**	**1**	**1**	**1**
"Procladius nigriventris"	2	1	1	1	1
"Procladius pectinatus"	**1**	**1**	**1**	**1**	**1**
"Procladius"	3	0	1	1	1
"Psectrocladius limbatellus"	5	1	2*	1	1
"Psectrocladius oxyura"	2	1	2*	2	2
"Psectrocladius"	1	0	1	1	1
"Pseudosmittia albipennis"	1	1	1	1	1
"Pseudosmittia obtusa"	**1**	**1**	**1**	**1**	**1**
"Smittia aterrima"	**2**	**1**	**1**	**1**	**1**
"Smittia edwardsi"	2	1	1	1	1
"Smittia stercoraria"	1	1	1	1	1

(Continued)

Table 3 (continued)

Morphotype	Specimens	Morph. identified species	BINs	ASAP	SP-ID
"Smittia terrestris"	9	1	3*	2	2
"Tanypus punctipennis"	**1**	**1**	**1**	**1**	**1**
"Tanypus vilipennis"	1	1	1	1	1
"Tanytarsus usmaensis"	2	1	1	1	1
"Tanytarsus"	**8**	**6**	**4**	**4**	**4**
"Thienemanniella"	1	1	1	1	1
"Xenopelopia nigricans"	**1**	**1**	**1**	**1**	**1**
Total	**294**	**76**	**88**	**71**	**75**

Notes:
Morphotype-names are in quotation marks, and those that include morphological misidentifications are in bold.
* Includes multiple BINs.

revealed that each of these morphotypes comprise 4–7 different taxonomic entities. In 15 cases, more BINs than morphologically identified species were recovered per morphotype. Morphotypes that include morphological misidentifications are in bold.

We created accumulation curves based on our morphological (Fig. 3A) and molecular data (Fig. 3B), depicting the number of recovered taxonomic entities for the most diverse morphotypes (with at least four taxonomic entities), and extrapolating to double the sampling effort. Most morphotypes that depict an accumulation curve, reach an asymptote. Comparing graphs, we see that in some cases, too many species were identified morphologically per morphotype (see "MT Tanytarsus" and "MT Glyptotendipes") and too few in others (see "MT Paratanytarsus/Rheotanytarsus").

DISCUSSION

In this study, we applied an integrative approach to facilitate sample processing of highly diverse non-biting midges. We applied a three-level subsampling technique and compared species recovered with each identification method (molecular and morphological) with the goal of assessing how an integrative approach can increase the incorporation of the Chironomidae into monitoring programs and biodiversity studies using a simplified approach (but without losing too much species information).

Morphotype sorting

Our results suggest that our morphotype sorting method was successful: We obtained a coverage of over 90% in species and cluster counts (Table 1), and the plateauing accumulation curves in Fig. 2 indicate that we would not have captured substantially more species by increasing our sampling effort. This is interesting, because after sorting non-biting midges into morphotype groups, we ultimately processed and identified only 7% of all specimens. Considering this, we believe that the task of grouping them into morphotypes, then selecting specimens for subsequent analysis can be easily delegated to parataxonomists. Overall, in-depth knowledge of Chironomidae morphology is not essential for this stage of sample processing, because sorting is based on phenotypic traits

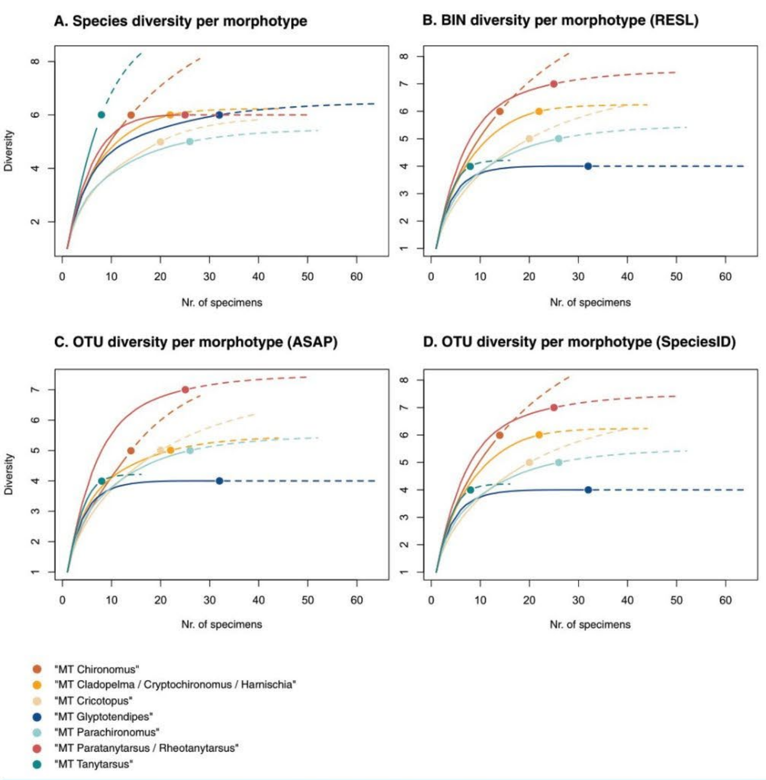


Figure 3 Accumulation curves of the diversity of (A) species, (B) BINs, (C) ASAP-OTUs, and (D) SpeciesIdentifier-OTUs recovered for each chironomid morphotype. Dotted lines represent extrapolated values (up to double the sampling effort), bold lines represent interpolated values. Accumulation curve of number of morphologically identified species (A) and BINs (B) recovered per morphotype based on the number of sampled specimens. Multiple BIN cases have been accounted for and removed. Full-size DOI: 10.7717/peerj.15336/fig-3

such as size, coloration, venation, setation, and shapes of antennae which simply require having a good "eye" and patience (*Krell, 2004*; *Ekrem, Stur & Hebert, 2010*). This approach was also applied by *Ekrem, Stur & Hebert (2010)* and authors to subsample non-biting midges for analysis in their study. We are aware that in our case, sorting was not conducted by a parataxonomist, but by an experienced scientist (*Ekrem, Stur & Hebert, 2010*). However, our taxonomist sorted these directly from the ethanol fluid using a stereo microscope, which does not provide a high-enough resolution for distinguishing genus- or species-level morphological features, especially not in ethanol. When confronted with large numbers of especially challenging specimens, our taxonomist resorted to either mounting representatives on temporary slides for guidance, or grouping specimens in the very few

genera that have distinct features even at low resolutions (*e.g.*, *Cricotopus*, *Ablabesmyia* and *Tanypus*). Our identifications of voucher specimens recovered up to seven taxonomic entities per single morphotype, indicating that when in doubt, it is simply easier to merge more specimens into one larger morphotype and compensate by increasing the number of vouchers.

Applying Chao statistics, we estimated that about 80 putative species may be present at the sampling sites. However, it is important to mention that to a certain degree, we are still underestimating the actual diversity of the Chironomidae that are present at the sampling sites. We applied our Chao statistics to a subset of the data, meaning that we are unintentionally inflating the probability of encountering a "new" and/or rare species, which in turn results in lower species estimates. To counteract this, we additionally applied a sample-based Chao2 estimator on the incidence data, which, resulted in much higher species estimates (Table 1). Needless to say, we may still be underestimating species numbers.

Using DNA barcoding: working with species proxies

In our study, we clustered our COI sequences using three delimitation algorithms, namely RESL, ASAP, and SpeciesIdentifier. Because the RESL algorithm and its BIN system is directly integrated into BOLD's interface, it is commonly used in DNA barcoding applications. However, there are varying opinions regarding the sole use of BINs for species delimitation (see *Cranston et al., 2013*; *Meier et al., 2022*), especially when assuming that BIN numbers are equal to species numbers in a 1:1 ratio. Therefore, as recommended by *Cranston et al. (2013)*, we analyzed our sequence data with several delimitation methods that apply different clustering algorithms. It is important to note that regardless which method one chooses for analysis, clustering algorithms remain arbitrary. Our results indicate that all three algorithms performed well, with molecular operational taxonomic unit (MOTU) diversities derived from each depicting overlapping 95% confidence intervals. Overall, we obtained very comparable results for all three clustering methods. In fact, using the NJ-tree to depict the assignment of specimens into clusters depicted almost identical results (see Data S1).

Using the RESL-algorithm led to the assignment to 77 BINs. Although BINs are a strong proxy for species boundaries (*Zahiri et al., 2014*; *Hebert et al., 2016*), it is important to keep in mind that they do not always reflect existing taxonomic systems (*Raupach et al., 2010*; *Hausmann et al., 2013*; *Zahiri et al., 2014*; *Hawlitschek et al., 2017*). Incongruences between BINs and traditional species names include multiple BIN assignments (more than one BIN is detected in a traditionally recognized species) and BIN sharing (the same BIN is detected across more than one recognized species) (*Hawlitschek et al., 2017*; *Chimeno et al., 2022*). Ideally, multiple BIN assignments would imply the presence of cryptic diversity whereas BIN sharing, which is commonly found among taxa with uncertain taxonomy or challenging species groups, is an indication for the need of species synonymization (*Hausmann et al., 2013*). However, ideal conditions are not the rule and there are various molecular factors (such as heteroplasmy, numts sequencing, introgression or homogenization of mtDNA haplotypes) that can challenge COI-based species

identifications (*Kmiec, Woloszynska & Janska, 2006*; *Dobson, 2004*; *Pamilo, Viljakainen & Vihavainen, 2007*; *Duron et al., 2008*; *Buhay, 2009*; *Hazkani-Covo, Zeller & Martin, 2010*), making it important to incorporate morphological information whenever possible. Additionally, accurate species identification is only guaranteed provided that high quality reference libraries are being used as a backbone to analysis (*Ekrem, Willassen & Stur, 2007*; *Chimeno et al., 2019*). These, in turn, rely on the accuracy of morphological identifications conducted on voucher specimens (*Ekrem, Willassen & Stur, 2007*). Mistakes in reference databases are challenging to uncover, especially if one is working with molecular data only. Yet requesting taxonomists to meticulously revise identifications of vouchers is not feasible. Instead, we suggest that it is mandatory that all records uploaded to BOLD are provided with an identifier and identification method, so that others can rely on the data when no expert is available. As suggested by *Brodin et al. (2012)* and authors, reference databases need to be expanded as best as possible in order to provide a better taxonomic coverage of species and their intraspecific variation. Quantity, however, should not come at a cost of quality. In our case, we double-checked every molecular-based identification using a neighbor-joining tree of public sequence data of vouchers that were morphologically identified by a taxonomist and uploaded to BOLD. Sequence records that were either identified using the "BIN taxonomy match" tool on BOLD, or that did not provide any information on the method of voucher identification whatsoever, were disregarded completely.

Discordances in our molecular dataset include multiple BINs assignments for a total of seven species, and the assignment of four interim species names. Although multiple BIN-assignments are an indication for cryptic diversity, extensive analysis is required to uncover the driving factors in the recovered genetic differences. On the other hand, interim species names are assigned to BINs when a genetic difference is detected, yet no species name can be provided. This can be an indication for the need of a taxonomic revision or a formal species description (*Morinière et al., 2019*; *Ekrem et al., 2019*). In other words: Interim species names provide species with an "intermediate name" until they obtain a formal species name. Because of this, such species can still be implemented into analyses, as in our study, because their BIN assignments act as "taxonomic handles" (see *Morinière et al. (2016)*, *Geiger et al. (2016)*).

The seven species involved in multiple-BIN cases are *Cladopelma viridulum*, *Polypedilum cultellatum*, *Polypedilum sordens*, *Psectrocladius oxyura*, *Psectrocladius limbatellus*, *Smittia stercoraria*, and *Smittia terrestris*. Research has shown that these genera (especially *Cladopelma*, *Polypedilum*, *Pscetrocladius* and *Smittia*) display much higher intraspecific variations in the COI barcode region across species, making it hard to identify a barcode gap that is needed for species discrimination (*Pillot, 2008*; *Cranston, Hardy & Morse, 2012*; *Tang et al., 2022*). These genera include species complexes whose taxonomic position is yet unsolved, and many traditional species are suspected to comprise more than one cryptic diverse species that are awaiting formal description (*Lehmann, 1970*; *Saether, 1971*; *Carew, Pettigrove & Hoffmann, 2005*; *Song et al., 2018*; *Chimeno et al., 2022*). *Song et al. (2018)*, for example, recovered a total of five BINs for *P. cultelatum* without finding any morphological discrepancies between adult specimens, and therefore

concluded that they may be dealing with potential cryptic species within a species complex. However, when *Carew, Pettigrove & Hoffmann (2005)* did not find DNA marker-associated morphological variations among individuals of the genus *Cladopelma*, they realized that this was due to the fact that these variations are only present among immature stages.

With the increase in barcoding campaigns, more COI-data of the Chironomidae is being made publicly available. One valuable asset of DNA barcoding is the fact that different life stages of the same species can be easily linked to one another without having to undergo larvae rearing which can be time-consuming, expensive, and for some species very challenging (*Stoeckle, 2003*; *Blaxter, 2004*; *Ekrem, Willassen & Stur, 2007*; *Stur & Ekrem, 2011*). With increased sequencing of larval stages, the COI sequences can be matched with those inferred from adult species and thus help enormously in resolving at least some taxonomic uncertainties (*Carew, Pettigrove & Hoffmann, 2005*; *Sinclair & Gresens, 2008*; *Montagna et al., 2016*).

Using morphology for species delimitation

In contrast to molecular identification methods, which use an algorithm for unbiased taxonomic clustering, accurate morphological identifications rely highly on (1) the availability and accuracy of species determination keys and (2) the identifier's ability to conduct identifications from an objective perspective (*Ekrem et al., 2019*). Chironomid identification requires extensive knowledge (which can generally only be provided by an expert) and ideally, as demonstrated by *Carew, Pettigrove & Hoffmann (2005)*, more than one single life-stage (*e.g.*, adults) of a single species should be assessed. Unfortunately, taxonomic expertise is overall in steady decline especially for those working on small-bodied and less conspicuous taxa (*Engel et al., 2021*; *Chimeno et al., 2022*). Still, the availability of a taxonomist does not automatically guarantee error-free species identifications, as demonstrated in this and other studies (*Failla et al., 2016*). Not only did we have a 9% error rate among morphological identifications, six of the "single species morphotypes" that were said to be distinguishable enough under the stereo microscope for direct species assignment were incorrectly identified. For another 9% of specimens, we could only provide identification to the family or to the genus-level.

False identifications were almost always within a given genus, hence, between closely related species whose morphological differences are often very subtle and therefore require specimen mounting and meticulous analysis (*Ekrem, Stur & Hebert, 2010*). For diverse morphotypes, the number of taxonomic entities recovered using morphology was often over- or under-estimated. This reflects the fact that on one hand, these taxa can display high levels of intraspecific morphological variation (*Carew et al., 2007*; *Carew, Marshall & Hoffmann, 2011*), and on the other hand, closely related species exhibit strong similarities, leading to the erroneous synonymization of species (*Anderson, Stur & Ekrem, 2013*). Despite having drastically reduced our taxonomist's workload by analyzing only a small portion of collected individuals, our taxonomist still spent about 500 active working hours processing, mounting, and identifying specimens, which was prone to errors over time (person. comment Baranov). This is a stark contrast to the 63 working hours for our

molecular approach. Although females are known to be even more difficult to identify than males, misidentifications were much more frequent among male individuals (70% of all type-1 discrepancies).

Overall, despite applying a three-level subsampling approach, which reduced the processing workload drastically, the performance of our taxonomist was affected by mistakes, caused by large amounts of material. These large amounts of material, however, represent the everyday life conditions in ecological surveys. For almost 20% of selected vouchers, no species-level information was provided, and we therefore conclude that it is difficult to meet the requirements of ecological studies using morphology alone.

CONCLUSION

Our current contribution shows that while both morphological identification and DNA barcoding have their own limitations, they are highly complementary in tackling large insect samples. While DNA barcoding does not require difficult-to-acquire taxonomic knowledge and drastically fast-forwards the process of identification of non-biting midges, barcode registries are only as valuable as the quality of their vouchers. Hence, without morphological identifications, there is no DNA barcoding. We presented one way to apply an integrative approach on Chironomidae, and presented a three-level sorting method for large samples. We were able to demonstrate that DNA barcoding less than 10% of a sample's contents can reliably detect >90% of its diversity, bringing us one step closer towards optimizing processing workflows for very large insect samples.

ACKNOWLEDGEMENTS

We are very grateful to the German Barcode of Life initiative (GBOL II) for providing the opportunity of specimen sequencing. We thank Martin Spies for the help with voucher verification, and Jörg Lewandowski (IGB Berlin) for supporting Viktor Baranov during his PhD, which allowed for side projects (such as this one) to take place.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Verlust der Nacht has been financially supported by the German Federal Ministry of Education and Research (BMBF-033L038A). Sequencing was done in the framework of the German Barcode of Life initiative (GBOL II), which was also funded by the German Federal Ministry of Education and Research (BMBF-01LI1101, BMBF-01LI1501). Alessandro Manfrin was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–326210499/GRK2360. Viktor Baranov received funds supporting the payment of the PeerJ publication fees through the CSIC Open Access Publication Support Initiative from the CSIC Unit of Information Resources for Research (URICI). Viktor Baranov's work is sponsored by the Spanish State Agency for Innovation's Ramon y Cajal fellowship (RyC2021-032144-I), project title "Climate change in the past and present & Insect decline". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
German Federal Ministry of Education and Research: BMBF-033L038A, BMBF-01LI1101, BMBF-01LI1501.
German Barcode of Life initiative (GBOL II).
Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): 326210499/GRK2360.
CSIC Unit of Information Resources for Research (URICI).
Spanish State Agency for Innovation's Ramon y Cajal Fellowship: RyC2021-032144-I.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Caroline Chimeno analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Björn Rulik performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Alessandro Manfrin conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Gregor Kalinkat conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Franz Hölker conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Viktor Baranov performed the experiments, authored or reviewed drafts of the article, and approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

All 294 sequences are available at GenBank: OP927392–OP927685 and BOLD https://dx.doi.org/10.5883/DS-ALANCHIR.

The data spreadsheets downloaded in April 2022 from BOLD (including sequencing and metadata) are available at Figshare: Chimeno, Caroline (2023): BOLD Spreadsheets. figshare. Dataset. https://doi.org/10.6084/m9.figshare.21803013.v1.

Data Availability

The following information was supplied regarding data availability:

The data spreadsheets downloaded in April 2022 from BOLD (including sequencing and metadata) are available at Figshare: Chimeno, Caroline (2023): BOLD Spreadsheets. figshare. Dataset. https://doi.org/10.6084/m9.figshare.21803013.v1.

The raw code for R analysis is available at Figshare: Chimeno, Caroline (2023): R Code for "Facing the Infinity". figshare. Software. https://doi.org/10.6084/m9.figshare.21787259.v3.

The input dataset for the R code is available at Figshare: Chimeno, Caroline (2023): Input Dataset for R code. figshare. Dataset. https://doi.org/10.6084/m9.figshare.21787262.v2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.15336#supplemental-information.

REFERENCES

Andersen T, Baranov V, Hagenlund LK, Ivković M, Kvifte GM, Pavlek M. 2016. Blind flight? A new Troglobiotic Orthoclad (*Diptera*, *Chironomidae*) from the Lukina Jama—Trojama Cave in Croatia. *PLOS ONE* **11(4)**:e0152884 DOI 10.1371/journal.pone.0152884.

Anderson AM, Stur E, Ekrem T. 2013. Molecular and morphological methods reveal cryptic diversity and three new species of Nearctic Micropsectra (*Diptera*: *Chironomidae*). *Freshwater Science* **32(3)**:892–921 DOI 10.1899/12-026.1.

Antczak-Orlewska O, Płóciennik M, Sobczyk R, Okupny D, Stachowicz-Rybka R, Rzodkiewicz M, Siciński J, Mroczkowska A, Krąpiec M, Słowiński M, Kittel P. 2021. *Chironomidae* morphological types and functional feeding groups as a habitat complexity vestige. *Frontiers in Ecology and Evolution* **8**:583831 DOI 10.3389/fevo.2020.583831.

Armitage PD, Pinder LC, Cranston PS. 2012. *The Chironomidae: biology and ecology of non-biting midges*. Cham: Springer Science & Business Media.

Astrin JJ, Stüben PE. 2008. Phylogeny in cryptic weevils: molecules, morphology and new genera of western Palaearctic *Cryptorhynchinae* (*Coleoptera*: *Curculionidae*). *Invertebrate Systematics* **22(5)**:503 DOI 10.1071/IS07057.

Baloğlu B, Clews E, Meier R. 2018. NGS barcoding reveals high resistance of a hyperdiverse chironomid (*Diptera*) swamp fauna against invasion from adjacent freshwater reservoirs. *Frontiers in Zoology* **15(1)**:31 DOI 10.1186/s12983-018-0276-7.

Baranov VA. 2011a. A preliminary annotated checklist of non-biting midges (*Diptera*, *Chironomidae*) of Ukraine. *Ukrainska Entomofaunistyka* **2**:7–24 DOI 10.5281/zenodo.5565483.

Baranov VA. 2011b. New and rare species of *Orthocladiinae* (*Diptera*, *Chironomidae*) from the Crimea, Ukraine. *Vestnik Zoologii* **45(5)**:405–410 *Вестник зоологии* DOI 10.2478/v10058-011-0026-1.

Baranov VA. 2013. First records of several *Tanypodinae* species (*Diptera*, *Chironomidae*) from Ukraine. *Vestnik Zoologii* **47(3)**:59–62 *Вестник зоологии*. DOI 10.2478/vzoo-2013-0027.

Baranov V, Andersen T, Hagenlund LK. 2015. A new species of Bryophaenocladius Thienemann, 1934 (*Diptera*, *Chironomidae*, *Orthocladiinae*) from Baltic amber. *Norwegian Journal of Entomology* **62**:53–56.

Baranov V, Andersen T, Perkovsky EE. 2015. Orthoclads from Eocene Amber from Sakhalin (*Diptera*: *Chironomidae*, *Orthocladiinae*). *Insect Systematics & Evolution* **46(4)**:359–378 DOI 10.1163/1876312X-45032122.

Baranov V, Giłka W, Zakrzewska M, Jarzembowski E. 2019. New non-biting midges (*Diptera*: *Chironomidae*) from Lower Cretaceous Wealden amber of the Isle of Wight (UK). *Cretaceous Research* **95(1)**:138–145 DOI 10.1016/j.cretres.2018.11.012.

Baranov V, Góral T, Ross A. 2017. A new genus of *Buchonomyiinae* (*Diptera*, *Chironomidae*) from Upper Cretaceous Burmese amber, with the phylogeny of the subfamily revisited. *Cretaceous Research* **79(1)**:146–152 DOI 10.1016/j.cretres.2017.07.007.

Baranov V, Lewandowski J, Krause S. 2016. Bioturbation enhances the aerobic respiration of lake sediments in warming lakes. *Biology Letters* **12(8)**:20160448 DOI 10.1098/rsbl.2016.0448.

Baranov V, Perkovsky EE. 2013. New chironomids from Eocene Sakhalinian amber (*Diptera*; *Chironomidae*; *Orthocladiinae*). *Terrestrial Arthropod Reviews* **6(1–2)**:61–69 DOI 10.1163/18749836-06021058.

Baranov VA, Przhiboro AA. 2014. New records of non-biting midges (*Diptera*: *Chironomidae*) from springs and streams of the Ukrainian Carpathians (Gorgany Massif). *Zoosystematica Rossica* **23**:150–157 DOI 10.31610/zsr/2014.23.1.150.

Blaxter ML. 2004. The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359(1444)**:669–679 DOI 10.1098/rstb.2003.1447.

Bohan DA, Vacher C, Tamaddoni-Nezhad A, Raybould A, Dumbrell AJ, Woodward G. 2017. Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in Ecology & Evolution* **32(7)**:477–487 DOI 10.1016/j.tree.2017.03.001.

Brodin Y, Ejdung G, Strandberg J, Lyrholm T. 2012. Improving environmental and biodiversity monitoring in the Baltic Sea using DNA barcoding of *Chironomidae* (*Diptera*). *Molecular Ecology Resources* **13(6)**:996–1004 DOI 10.1111/1755-0998.12053.

Brodin Y, Lundström JO, Paasivirta L. 2008. Tavastia yggdrasilia, a new orthoclad midge (*Diptera*: *Chironomidae*) from Europe. *Aquatic Insects* **30(4)**:261–267 DOI 10.1080/01650420802448515.

Buhay JE. 2009. COI-like sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology* **29(1)**:96–110 DOI 10.1651/08-3020.1.

Carew ME, Marshall SE, Hoffmann AA. 2011. A combination of molecular and morphological approaches resolves species in the taxonomically difficult genus Procladius Skuse (*Diptera*: *Chironomidae*) despite high intra-specific morphological variation. *Bulletin of Entomological Research* **101(5)**:505–519 DOI 10.1017/S000748531100006X.

Carew ME, Pettigrove V, Cox RL, Hoffmann AA. 2007. DNA identification of urban Tanytarsini chironomids (*Diptera*: *Chironomidae*). *Journal of the North American Benthological Society* **26(4)**:587–600 DOI 10.1899/06-120.1.

Carew ME, Pettigrove V, Hoffmann AA. 2005. The utility of DNA markers in classical taxonomy: using cytochrome oxidase I markers to differentiate Australian Cladopelma (*Diptera*: *Chironomidae*) Midges. *Annals of the Entomological Society of America* **98(4)**:587–594 DOI 10.1603/0013-8746(2005)098[0587:TUODMI]2.0.CO;2.

Chan A, Chiang L-P, Hapuarachchi HC, Tan C-H, Pang S-C, Lee R, Lee K-S, Ng L-C, Lam-Phua S-G. 2014. DNA barcoding: complementing morphological identification of mosquito species in Singapore. *Parasites & Vectors* **7(1)**:569 DOI 10.1186/s13071-014-0569-4.

Chao A, Ma KH, Hsieh TC, Chiu C-H. 2016. SpadeR: species-richness prediction and diversity estimation with R. *Available at https://cran.r-project.org/web/packages/SpadeR/SpadeR.pdf*.

Chimeno C, Hausmann A, Schmidt S, Raupach MJ, Doczkal D, Baranov V, Hübner J, Höcherl A, Albrecht R, Jaschhof M, Haszprunar G, Hebert PDN. 2022. Peering into the darkness: DNA barcoding reveals surprisingly high diversity of unknown species of *Diptera* (*Insecta*) in Germany. *Insects* **13(1)**:82 DOI 10.3390/insects13010082.

Chimeno C, Hübner J, Seifert L, Morinière J, Bozicevic V, Hausmann A, Schmidt S, Müller J. 2023. Depicting environmental gradients from Malaise trap samples: is ethanol-based DNA metabarcoding enough? *Insect Conservation and Diversity* **16(1)**:47–64 DOI 10.1111/icad.12609.

Chimeno C, Morinière J, Podhorna J, Hardulak L, Hausmann A, Reckel F, Grunwald JE, Penning R, Haszprunar G. 2019. DNA barcoding in forensic entomology—establishing a DNA

reference library of potentially forensic relevant Arthropod Species. *Journal of Forensic Sciences* **64(2)**:593–601 DOI 10.1111/1556-4029.13869.

Cranston P. 2008. Keys to the adult male *Chironomidae* of Britain and Ireland—edited by P.H. Langton and L.C.V. Pinder and a guide to the identification of genera of Chironomid Pupal Exuviae occurring in Britain and Ireland (including common genera from Northern Europe) and their use in monitoring Lotic and Lentic fresh waters—edited by R.S. Wilson and L.P. Ruse. *Systematic Entomology* **33**:215–217 DOI 10.1111/j.1365-3113.2007.00412_5.x.

Cranston PS, Ang YC, Heyzer A, Lim RBH, Wong WH, Woodford JM, Meier RUDOLF. 2013. The nuisance midges (*Diptera*: *Chironomidae*) of Singapore's Pandan and Bedok reservoirs. *The Raffles Bulletin of Zoology* **61(2)**:779–793.

Cranston PS, Hardy NB, Morse GE. 2012. A dated molecular phylogeny for the *Chironomidae* (*Diptera*). *Systematic Entomology* **37(1)**:172–188 DOI 10.1111/j.1365-3113.2011.00603.x.

Dayrat B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* **85(3)**:407–415 DOI 10.1111/j.1095-8312.2005.00503.x.

Dobson SL. 2004. Evolution of Wolbachia cytoplasmic incompatibility. *Evolution* **58**:2156–2166 DOI 10.1111/j.0014-3820.2004.tb01594.x.

Dorić V, Pozojević I, Vučković N, Ivković M, Mihaljević Z. 2021. Lentic chironomid performance in species-based bioassessment proving: high-level taxonomy is not a dead end in monitoring. *Ecological Indicators* **121(4)**:107041 DOI 10.1016/j.ecolind.2020.107041.

Duron O, Bouchon D, Boutin S, Bellamy L, Zhou L, Engelstädter J, Hurst GD. 2008. The diversity of reproductive parasites among arthropods: Wolbachiado not walk alone. *BMC Biology* **6(1)**:27 DOI 10.1186/1741-7007-6-27.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32(5)**:1792–1797 DOI 10.1093/nar/gkh340.

Egger J. 1863. Dipterologische Beiträge. *Verhandlungen der Zoologisch-Botanischen Gesellschaft in Wien* **13**:1101–1110.

Ekrem T. 2002a. A revision of the Tanytarsus Eminulus, Gregarius, Lugens and Mendax species groups (*Insecta*: *Diptera*: *Chironomidae*). PhD Thesis Thesis. University of Bergen, Museum of Zoology.

Ekrem T. 2002b. A review of selected South- and East Asian Tanytarsus v.d. Wulp (*Diptera*: *Chironomidae*). *Hydrobiologia* **474(1/3)**:1–39 DOI 10.1023/A:1016527603086.

Ekrem T. 2007. A taxonomic revision of the genus *Stempellinella* (*Diptera*: *Chironomidae*). *Journal of Natural History* **41(21–24)**:1367–1465 DOI 10.1080/00222930701437360.

Ekrem T, Stur E. 2009. A review of the genus Zavrelia (*Diptera*: *Chironomidae*). *European Journal of Entomology* **106(1)**:119–144 DOI 10.14411/eje.2009.016.

Ekrem T, Stur E, Hebert PDN. 2010. Females do count: documenting *Chironomidae* (*Diptera*) species diversity using DNA barcoding. *Organisms Diversity & Evolution* **10(5)**:397–408 DOI 10.1007/s13127-010-0034-y.

Ekrem T, Stur E, Orton MG, Adamowicz SJ. 2019. DNA barcode data reveal biogeographic trends in Arctic non-biting midges. *Trends in DNA Barcoding and Metabarcoding* **1**:787–796 DOI 10.1139/gen-2018-0100@gen-dna.issue01.

Ekrem T, Willassen E, Stur E. 2007. A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution* **43(2)**:530–542 DOI 10.1016/j.ympev.2006.11.021.

Engel MS, Ceríaco LMP, Daniel GM, Dellapé PM, Löbl I, Marinov M, Reis RE, Young MT, Dubois A, Agarwal I, Lehmann AP, Alvarado M, Alvarez N, Andreone F, Araujo-Vieira K,

Ascher JS, Baêta D, Baldo D, Bandeira SA, Barden P, Barrasso DA, Bendifallah L, Bockmann FA, Böhme W, Borkent A, Brandão CRF, Busack SD, Bybee SM, Channing A, Chatzimanolis S, Christenhusz MJM, Crisci JV, D'elía G, Da Costa LM, Davis SR, De Lucena CAS, Deuve T, Fernandes Elizalde S, Faivovich J, Farooq H, Ferguson AW, Gippoliti S, Gonçalves FMP, Gonzalez VH, Greenbaum E, Hinojosa-Díaz IA, Ineich I, Jiang J, Kahono S, Kury AB, Lucinda PHF, Lynch JD, Malécot V, Marques MP, Marris JWM, Mckellar RC, Mendes LF, Nihei SS, Nishikawa K, Ohler A, Orrico VGD, Ota H, Paiva J, Parrinha D, Pauwels OSG, Pereyra MO, Pestana LB, Pinheiro PDP, Prendini L, Prokop J, Rasmussen C, Rödel M-O, Rodrigues MT, Rodríguez SM, Salatnaya H, Sampaio Í, Sánchez-García A, Shebl MA, Santos BS, Solórzano-Kraemer MM, Sousa ACA, Stoev P, Teta P, Trape J-F, Dos Santos CV-D, Vasudevan K, Vink CJ, Vogel G, Wagner P, Wappler T, Ware JL, Wedmann S, Zacharie CK. 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society* **193(2)**:381–387 DOI 10.1093/zoolinnean/zlab072.

Failla AJ, Vasquez AA, Hudson P, Fujimoto M, Ram JL. 2016. Morphological identification and COI barcodes of adult flies help determine species identities of chironomid larvae (*Diptera, Chironomidae*). *Bulletin of Entomological Research* **106(1)**:34–46 DOI 10.1017/S0007485315000486.

Gadawski P, Montagna M, Rossaro B, Giłka W, Pešić V, Grabowski M, Magoga G. 2022. DNA barcoding of *Chironomidae* from the Lake Skadar region: reference library and a comparative analysis of the European fauna. *Diversity and Distributions* **28(12)**:1–20 DOI 10.1111/ddi.13504.

Geiger MF, Moriniere J, Hausmann A, Haszprunar G, Wägele W, Hebert PDN, Rulik B. 2016. Testing the Global Malaise Trap Program—how well does the current barcode reference library identify flying insects in Germany? *Biodiversity Data Journal* **4(1)**:e10671 DOI 10.3897/BDJ.4.e10671.

Giłka W. 2011. Ochotkowate—*Chironomidae*: plemię Tanytarsini, postaci dorosłe, samce/oprac. Biologica Silesiae. *Available at https://repozytorium.bg.ug.edu.pl/info/article/UOGc32f388322db4c10a6d29d2f6fd919c5/*.

Giłka W, Gadawski P. 2022. Tanytarsini (*Diptera*: *Chironomidae*) collected during Polish expeditions to North Korea. *Zootaxa* **5214(151)**:175 DOI 10.11646/zootaxa.5214.2.1.

Gratton C, Zanden MJV. 2009. Flux of aquatic insect productivity to land: comparison of lentic and lotic ecosystems. *Ecology* **90(10)**:2689–2699 DOI 10.1890/08-1546.1.

Hartop E, Srivathsan A, Ronquist F, Meier R. 2022. Towards large-scale integrative taxonomy (LIT): resolving the data conundrum for dark taxa. *Systematic Biology* **71(6)**:1404–1422 DOI 10.1093/sysbio/syac033.

Hausmann A, Godfray HCJ, Huemer P, Mutanen M, Rougerie R, van Nieukerken EJ, Ratnasingham S, Hebert PDN. 2013. Genetic patterns in european geometrid moths revealed by the barcode index number (BIN) system. *PLOS ONE* **8(12)**:e84518 DOI 10.1371/journal.pone.0084518.

Hausmann A, Krogmann L, Peters R, Rduch V, Schmidt S. 2020. GBOL III: dark taxa. *iBOL Barcode Bulletin* **10(1)**:6242 DOI 10.21083/ibol.v10i1.6242.

Hawlitschek O, Morinière J, Lehmann GUC, Lehmann AW, Kropf M, Dunz A, Glaw F, Detcharoen M, Schmidt S, Hausmann A, Szucsich NU, Caetano-Wyler SA, Haszprunar G. 2017. DNA barcoding of crickets, katydids and grasshoppers (*Orthoptera*) from Central Europe with focus on Austria, Germany and Switzerland. *Molecular Ecology Resources* **17(5)**:1037–1053 DOI 10.1111/1755-0998.12638.

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular Poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLOS Genetics* **6(2)**:e1000834 DOI 10.1371/journal.pgen.1000834.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270(1512)**:313–321 DOI 10.1098/rspb.2002.2218.

Hebert PDN, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, Pedersen S, Jannetta P, deWaard JR. 2016. Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371(1702)**:20150333 DOI 10.1098/rstb.2015.0333.

Herren CM, Webert KC, Drake MD, Jake Vander Zanden M, Einarsson Á, Ives AR, Gratton C. 2017. Positive feedback between chironomids and algae creates net mutualism between benthic primary consumers and producers. *Ecology* **98(2)**:447–455 DOI 10.1002/ecy.1654.

Hilsenhoff WL, Thorp JH, Covich AP. 2001. Ecology and classification of North American freshwater invertebrates. *Diversity and Classification of Insects and Collembola.* Cambridge: Academic Press, 661–721.

Hirvenoja M. 1973. Revision der Gattung Cricotopus van der Wulp und ihrer Verdandten. *Annales Zoologici Fennici* **10**:1–363.

Holzhauer S, Franke S, Kyba C, Manfrin A, Klenke R, Voigt C, Lewanzik D, Oehlert M, Monaghan M, Schneider S, Heller S, Kuechly H, Brüning A, Honnen A-C, Hölker F. 2015. Out of the dark: establishing a large-scale field experiment to assess the effects of artificial light at night on species and food webs. *Sustainability* **7(11)**:15593–15616 DOI 10.3390/su71115593.

Hsieh TC, Ma KH, Chao A. 2020. iNEXT: interpolation and extrapolation for species diversity. *Available at https://cran.r-project.org/web/packages/iNEXT/iNEXT.pdf.*

Hsieh TC, Ma KH, Chao A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* **7(12)**:1451–1456 DOI 10.1111/2041-210X.12613.

Hölker F, Vanni MJ, Kuiper JJ, Meile C, Grossart H-P, Stief P, Adrian R, Lorke A, Dellwig O, Brand A, Hupfer M, Mooij WM, Nützmann G, Lewandowski J. 2015. Tube-dwelling invertebrates: tiny ecosystem engineers have large effects in lake ecosystems. *Ecological Monographs* **85(3)**:333–351 DOI 10.1890/14-1160.1.

Karima Z. 2021. *Chironomidae*: biology, ecology and systematics. In: *The Wonders of Diptera.* London, United Kingdom: IntechOpen.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28(12)**:1647–1649 DOI 10.1093/bioinformatics/bts199.

Keck F, Vasselon V, Tapolczai K, Rimet F, Bouchez A. 2017. Freshwater biomonitoring in the information age. *Frontiers in Ecology and the Environment* **15(5)**:266–274 DOI 10.1002/fee.1490.

Kirk-Spriggs AH, Sinclair BJ. 2017. *Manual of afrotropical diptera volume 1.* Pretoria: South African National Biodiversity Institute.

Kmiec B, Woloszynska M, Janska H. 2006. Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. *Current Genetics* **50(3)**:149–159 DOI 10.1007/s00294-006-0082-1.

Krell F-T. 2004. Parataxonomy *vs* taxonomy in biodiversity studies—Pitfalls and applicability of morphospecies sorting. *Biodiversity and Conservation* **13(4)**:795–812 DOI 10.1023/B:BIOC.0000011727.53780.63.

Langton PH, Pinder LCV. 2007. *Keys to the adult male Chironomidae of Britain and Ireland, Vol 1 and 2.* Ambleside: The Freshwater Biological Association.

Lehmann J. 1970. Revision der Europäischen Arten (Imagines ♂♂) der Gattung Parachironomus Lenz (*Diptera, Chironomidae*). *Hydrobiologia* **36(1)**:129–158 DOI 10.1007/BF00751287.

Lencioni V, Marziali L, Rossaro B. 2012. Chironomids as bioindicators of environmental quality in mountain springs. *Freshwater Science* **31(2)**:525–541 DOI 10.1899/11-038.1.

Lin X, Stur E, Ekrem T. 2015. Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. *PLOS ONE* **10(9)**:e0138993 DOI 10.1371/journal.pone.0138993.

Lundström JO, Brodin Y, Schäfer ML, Vinnersten TZP, Östman Ö. 2010. High species richness of *Chironomidae* (*Diptera*) in temporary flooded wetlands associated with high species turn-over rates. *Bulletin of Entomological Research* **100(4)**:433–444 DOI 10.1017/S0007485309990472.

Mandelik Y, Roll U, Fleischer A. 2010. Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *Journal of Applied Ecology* **47(6)**:1179–1188 DOI 10.1111/j.1365-2664.2010.01864.x.

Manfrin A, Singer G, Larsen S, Weiß N, van Grunsven RHA, Weiß N-S, Wohlfahrt S, Monaghan MT, Hölker F. 2017. Artificial light at night affects organism flux across ecosystem boundaries and drives community structure in the recipient ecosystem. *Frontiers in Environmental Science* **5**:61 DOI 10.3389/fenvs.2017.00061.

Marziali L, Armanini DG, Cazzola M, Erba S, Toppi E, Buffagni A, Rossaro B. 2010. Responses of Chironomid larvae (*Insecta, Diptera*) to ecological quality in Mediterranean river mesohabitats (South Italy). *River Research and Applications* **26(8)**:1036–1051 DOI 10.1002/rra.1303.

Meier R, Blaimer BB, Buenaventura E, Hartop E, von Rintelen T, Srivathsan A, Yeo D. 2022. A re-analysis of the data in Sharkey et al.'s, (2021) minimalist revision reveals that BINs do not deserve names, but BOLD Systems needs a stronger commitment to open science. *Cladistics* **38(2)**:264–275.

Meier R, Shiyang K, Vaidya G, Ng PK. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55(5)**:715–728.

Meigen J. 1818. Nouvelle classification des mouches a deux ailes (*Diptera* L.). In: Fuchs JJ, ed. Paris: D'apres un plan tout nouveau.

Milošević D, Milosavljević A, Predić B, Medeiros AS, Savić-Zdravković D, Stojković Piperac M, Kostić T, Spasić F, Leese F. 2020. Application of deep learning in aquatic bioassessment: towards automated identification of non-biting midges. *Science of the Total Environment* **711(1)**:135160 DOI 10.1016/j.scitotenv.2019.135160.

Montagna M, Mereghetti V, Lencioni V, Rossaro B. 2016. Integrated taxonomy and DNA barcoding of Alpine Midges (*Diptera*: *Chironomidae*). *PLOS ONE* **11(3)**:e0149673 DOI 10.1371/journal.pone.0149673.

Morinière J, Balke M, Doczkal D, Geiger MF, Hardulak LA, Haszprunar G, Hausmann A, Hendrich L, Regalado L, Rulik B, Schmidt S, Wägele J-W, Hebert PDN. 2019. A DNA barcode library for 5,200 German flies and midges (*Insecta*: *Diptera*) and its implications for metabarcoding-based biomonitoring. *Molecular Ecology Resources* **19(4)**:900–928 DOI 10.1111/1755-0998.13022.

Morinière J, de Araujo BC, Lam AW, Hausmann A, Balke M, Schmidt S, Hendrich L, Doczkal D, Fartmann B, Arvidsson S, Haszprunar G. 2016. Species identification in Malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLOS ONE* **11(5)**:e0155497 DOI 10.1371/journal.pone.0155497.

Pamilo P, Viljakainen L, Vihavainen A. 2007. Exceptionally high density of NUMTs in the honeybee genome. *Molecular Biology and Evolution* **24(6)**:1340–1346 DOI 10.1093/molbev/msm055.

Pfenninger M, Nowak C, Kley C, Steinke D, Streit B. 2007. Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (*Diptera*) species. *Molecular Ecology* **16(9)**:1957–1968 DOI 10.1111/j.1365-294X.2006.03136.x.

Pillot HKM. 2008. Identification and ecology of the genus Smittia Holmgren in the Netherlands (*Diptera*: *Chironomidae*). *Tijdschrift voor Entomologie* **151**:245–270.

Pires A, Marinoni L. 2010. DNA barcoding and traditional taxonomy unified through Integrative Taxonomy: a view that challenges the debate questioning both methodologies. *Biota Neotropica* **10(2)**:339–346 DOI 10.1590/S1676-06032010000200035.

Porter TM, Gibson JF, Shokralla S, Baird DJ, Golding GB, Hajibabaei M. 2014. Rapid and accurate taxonomic classification of insect (class *Insecta*) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Molecular Ecology Resources* **14(5)**:929–942 DOI 10.1111/1755-0998.12240.

Proulx I, Martin J, Carew M, Hare L. 2013. Using various lines of evidence to identify Chironomus species (*Diptera*: *Chironomidae*) in eastern Canadian lakes. *Zootaxa* **3741(4)**:401–458 DOI 10.11646/zootaxa.3741.4.1.

Puillandre N, Brouillet S, Achaz G. 2021. ASAP: assemble species by automatic partitioning. *Molecular Ecology Resources* **21(2)**:609–620.

Ratnasingham S, Hebert PDN. 2007. BOLD: the barcode of life data system. *Molecular Ecology Notes* **7(3)**:355–364 DOI 10.1111/j.1471-8286.2007.01678.x.

Raunio J, Heino J, Paasivirta L. 2011. Non-biting midges in biodiversity conservation and environmental assessment: findings from boreal freshwater ecosystems. *Ecological Indicators* **11(5)**:1057–1064 DOI 10.1016/j.ecolind.2010.12.002.

Raupach MJ, Astrin JJ, Hannig K, Peters MK, Stoeckle MY, Wägele J-W. 2010. Molecular species identification of Central European ground beetles (*Coleoptera*: *Carabidae*) using nuclear rDNA expansion segments and DNA barcodes. *Frontiers in Zoology* **7(1)**:26 DOI 10.1186/1742-9994-7-26.

Rosenberg DM. 1992. Freshwater biomonitoring and *Chironomidae*. *Netherland Journal of Aquatic Ecology* **26(2–4)**:101–122 DOI 10.1007/BF02255231.

Saether OA. 1971. Nomenclature and phylogeny of the genus Harnischia (*Diptera*: *Chironomidae*). *The Canadian Entomologist* **103(3)**:347–362 DOI 10.4039/Ent103347-3.

Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH. 2010. Integrative taxonomy: a multisource approach to exploring biodiversity. *Annual Review of Entomology* **55(1)**:421–438 DOI 10.1146/annurev-ento-112408-085432.

Sheth BP, Thaker VS. 2017. DNA barcoding and traditional taxonomy: an integrated approach for biodiversity conservation. *Genome* **60(7)**:618–628 DOI 10.1139/gen-2015-0167.

Silva F, Wiedenbrug S. 2015. Amazonimyia gigantea gen. n., sp. N., a new *Tanypodinae* (*Diptera*: *Chironomidae*) from the Neotropical Region. *Zootaxa* **3947**:275–281 DOI 10.11646/zootaxa.3947.2.9.

Sinclair CS, Gresens SE. 2008. Discrimination of Cricotopus species (*Diptera*: *Chironomidae*) by DNA barcoding. *Bulletin of Entomological Research* **98(6)**:555–563 DOI 10.1017/S0007485308005865.

Siri A, Brodin Y. 2014. Cladistic analysis of Rheochlus and related genera, with description of a new species (*Diptera*: *Chironomidae*: *Podonominae*). *Acta Entomologica Musei Nationalis Pragae* **54**:361–375.

Song C, Lin X-L, Wang Q, Wang X-H. 2018. DNA barcodes successfully delimit morphospecies in a superdiverse insect genus. *Zoologica Scripta* **47(3)**:311–324 DOI 10.1111/zsc.12284.

Stoeckle M. 2003. Taxonomy, DNA, and the bar code of life. *BioScience* **53(9)**:796–797 DOI 10.1641/0006-3568(2003)053[0796:TDATBC]2.0.CO;2.

Stur E, Ekrem T. 2000. Tanytarsus usambarae, spec. nov. from West Usambara Mts., Tanzania, East Africa (*Insecta*, *Diptera*, *Chironomidae*). *Spixiana* **23**:219–223.

Stur E, Ekrem T. 2006. A revision of West Palaearctic species of the Micropsectra atrofasciata species group (*Diptera*: *Chironomidae*). *Zoological Journal of the Linnean Society* **146(2)**:165–225 DOI 10.1111/j.1096-3642.2006.00198.x.

Stur E, Ekrem T. 2011. Exploring unknown life stages of Arctic Tanytarsini (*Diptera*: *Chironomidae*) with DNA barcoding. *Zootaxa* **2743(1)**:27–39 DOI 10.11646/zootaxa.2743.1.2.

Stur E, Ekrem T. 2015. A review of Norwegian Gymnometriocnemus (*Diptera*, *Chironomidae*) including the description of two new species and a new name for Gymnometriocnemus volitans (Goetghebuer) sensu Brundin. *ZooKeys* **508**:127–142 DOI 10.3897/zookeys.508.9874.

Stur E, Spies M. 2011. Description of Chaetocladius longivirgatus sp. n., with a review of C. suecicus (Kieffer) (*Diptera*: *Chironomidae*). *Zootaxa* **2762(1)**:37–48 DOI 10.11646/zootaxa.2762.1.3.

Stur E, Wiedenbrug S. 2005. Two new Orthoclad species (*Diptera*: *Chironomidae*) from cold water springs of the Nationalpark Berchtesgaden. *Germany Aquatic Insects* **27(2)**:125–131 DOI 10.1080/01650420400019817.

Sæther OA. 1977. Female genitalia in *Chironomidae* and other *Nematocera*: morphology, phylogenies, keys. Department of Fisheries and the Environment, Fisheries and Marine Service.

Tang H, Cheng Q, Han W, Cranston PS. 2022. Integrative taxonomy: molecular phylogenetics of Polypedilum (Cerobregma) and revisited morphology of Yaethauma and Collartomyia (*Diptera*: *Chironomidae*) reveals synonymy and supports new classification. *Zoological Journal of the Linnean Society* **194**:102–119 DOI 10.1093/zoolinnean/zlaa187.

Wiedenbrug S, Lamas C, Trivinho-Strixino S. 2012. A review of the genus Corynoneura Winnertz (*Diptera*: *Chironomidae*) from the Neotropical region. *Zootaxa* **3574(1)**:1–61 DOI 10.11646/zootaxa.3574.1.1.

Wiedenbrug S, Lamas C, Trivinho-Strixino S. 2013. A review of Neotropical species in Thienemanniella Kieffer (*Diptera*, *Chironomidae*). *Zootaxa* **3670(2)**:215–237 DOI 10.11646/zootaxa.3670.2.7.

Wiedenbrug S, Silva F. 2016. Diplosmittia caribensis, a new *Orthocladiinae* (*Diptera*: *Chironomidae*) from the Dominican Republic. *Zootaxa* **4103(1)**:71 DOI 10.11646/zootaxa.4103.1.8.

Wiederholm T. 1989. *Chironomidae* of the Holarctic region. Keys and diagnoses. Part 3-Adult males. Vol. 3. Lund: Entomological Society of Lund, 1–532.

Will KW, Mishler BD, Wheeler QD. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology* **54(5)**:844–851 DOI 10.1080/10635150500354878.

Wirta HK, Vesterinen EJ, Hambäck PA, Weingartner E, Rasmussen C, Reneerkens J, Schmidt NM, Gilg O, Roslin T. 2015. Exposing the structure of an Arctic food web. *Ecology and Evolution* **5(17)**:3842–3856 DOI 10.1002/ece3.1647.

Zahiri R, Lafontaine JD, Schmidt BC, deWaard JR, Zakharov EV, Hebert PDN. 2014. A transcontinental challenge — a test of DNA barcode performance for 1,541 species of Canadian *Noctuoidea* (*Lepidoptera*). *PLOS ONE* **9(3)**:e92797 DOI 10.1371/journal.pone.0092797.

Publication V:

Depicting environmental gradients from Malaise trap samples: Is ethanol based DNA metabarcoding enough?

Published in Insect Conservation and Diversity

Chimeno, C., Hübner, J., Seifert, L., Morinière, J., Bozicevic, V., Hausmann, A., Schmidt, S. & Müller, J. (2023). Depicting environmental gradients from Malaise trap samples: Is ethanol based DNA metabarcoding enough?. *Insect Conservation and Diversity*, *16*(1), 47-64.
https://doi.org/10.1111/icad.12609

Received: 18 October 2021 | Accepted: 21 September 2022
DOI: 10.1111/icad.12609

ORIGINAL ARTICLE

Depicting environmental gradients from Malaise trap samples: Is ethanol-based DNA metabarcoding enough?

Caroline Chimeno[1] | Jeremy Hübner[1] | Linda Seifert[2] | Jérôme Morinière[3] | Vedran Bozicevic[3] | Axel Hausmann[1] | Stefan Schmidt[1] | Jörg Müller[2,4]

[1]SNSB-Zoologische Staatssammlung München, Munich, Germany
[2]Bavarian Forest National Park, Grafenau, Germany
[3]Advanced Identification Methods GmbH (AIM), Leipzig, Germany
[4]Field Station Fabrikschleichach, Department of Animal Ecology and Tropical Biology, Biocenter, University of Würzburg, Rauhenebrach, Germany

Correspondence
Caroline Chimeno, SNSB-Zoologische Staatssammlung München, Münchhausenstr. 21, 81247 Munich, Germany.
Email: chimeno@snsb.de

Funding information
Bavarian Forest National Park Authority; Federal Ministry of Education (BMBF), Grant/Award Number: 6LI1901B

Editor: Yves Basset and Associate Editor: Nusha Keyghobadi

Abstract

1. DNA metabarcoding is revolutionising biodiversity research, as it offers researchers a holistic taxonomic approach across lineages. Many studies are dedicated to testing its application and optimising workflows. One topic of discussion is the nature of samples used for sequencing and comparing taxonomic results.
2. However, in ecological and environmental studies, where scientists always work with subsets of species, it may be less important whether different methods provide different subsets but more important if ecological and environmental information is conserved equally.
3. Numerous studies have successfully applied destructive and non-destructive metabarcoding approaches to evaluate patterns in biodiversity and in this respect, we aim to determine for the very first time whether environmental information is also conserved in the preservative ethanol of terrestrial arthropod bulk samples.
4. To test this, we applied DNA metabarcoding on tissue DNA and on ethanol-based DNA of the same Malaise trap samples. The arthropod material was collected with eight traps located in three different habitats: forest, meadow, and riparian.
5. We identified more than 3000 operational taxonomic units and demonstrate that ethanol-based DNA sequencing did not provide information on ecological gradients, except for the case of seasonal patterns, which was well conserved for some taxa.
6. The conserved seasonality is an interesting starting point for further investigations. Until future research has provided more successful results, we recommend researchers dealing with terrestrial ecosystems to be careful when using ethanol DNA.

KEYWORDS
arthropod communities, arthropod tissue, DNA barcoding, ecological gradients, ethanol-based DNA, Malaise traps, metabarcoding, preservative ethanol, tissue-based DNA

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Insect Conservation and Diversity* published by John Wiley & Sons Ltd on behalf of Royal Entomological Society.

INTRODUCTION

Evaluating the state of an ecosystem requires adequate monitoring of biodiversity (Liu et al., 2021). This includes having knowledge on the inhabiting communities at one or more ecological levels and assessing changes over time and space (Coissac et al., 2012; Niemelä, 2000). Arthropods are especially suitable as ecological indicators, because they are abundant, species rich, and sensitive to slight environmental changes due to their functionality in an ecosystem (Medhi et al., 2021; Schowalter, 2017). However, identifying arthropod species using conventional morphological approaches is challenging, often dependent on specialised taxonomists (the availability of which is in decline), and time-consuming (Chimeno et al., 2022; Ji et al., 2013; Morinière et al., 2016; Yu et al., 2012).

Following the advent of DNA barcoding (Hebert et al., 2003), molecular approaches have become more frequent in biomonitoring surveys (Cristescu, 2014; Hardulak et al., 2020; Shokralla et al., 2012). One approach that is expediting biodiversity monitoring is DNA metabarcoding (Liu et al., 2020; Makiola et al., 2020). This method extends single species delimitation to the identification of entire communities holistically by extracting genetic material from entire bulk samples and sequencing a standard DNA marker via high-throughput sequencing (HTS) (Aylagas et al., 2018; Cristescu, 2014; Hardulak et al., 2020; Ji et al., 2013; Keck et al., 2017; Meusnier et al., 2008; Taberlet et al., 2012; Yu et al., 2012). Not only does DNA metabarcoding enable highly standardised, reliable and cost-efficient community analysis, it also enables biodiversity assessments of larger community subsets across a broad range of ecosystems (Liu et al., 2021; Morinière et al., 2016). Analysis of biodiversity patterns driven by ecological gradients is therefore much more comprehensive than in conventional biomonitoring where scientists are often limited to the evaluation of few key taxa (Bohan et al., 2017; Keck et al., 2017; Mandelik et al., 2010; Porter et al., 2014).

Although DNA metabarcoding has become a well-established method (Shum & Palumbi, 2021), a consensus workflow is still lacking in some fields of research (e.g. studies on terrestrial arthropods; see Elbrecht & Leese, 2015). Numerous studies are therefore dedicated to testing its robustness across protocols (Deagle et al., 2014; Hardulak et al., 2020; Ji et al., 2013; Marquina et al., 2019). One subject of debate, for example, is the nature of samples used for sequencing. Homogenisation of arthropod tissue has quickly become a favoured approach, because most DNA is released upon tissue destruction. More DNA, however, comes at a cost of losing the specimen's structural integrity, which erases any possibility for subsequent morphological analysis (Aylagas et al., 2016, 2018).

Due to its non-destructive nature and easy application, the interest for ethanol-based DNA sequencing has greatly increased in recent years. Instead of regarding ethanol as a mere preservative that is discarded upon specimen analysis, it could be poured out, filtered and its contents subjected to molecular analysis. Thus, ethanol-based DNA metabarcoding can provide an extensive community analysis all while keeping the specimens intact (Erdozain et al., 2019; Marquina et al., 2019). Studies testing the consistency of taxonomic results between the use of specimen tissue and preservative ethanol are still sparse, and those that have provide divergent results. Studies conducted on freshwater benthic macroinvertebrates were overall more successful (Hajibabaei et al., 2012; Zizka et al., 2019) than those conducted on terrestrial arthropods (see Kirse et al., 2022; Linard et al., 2016; Marquina et al., 2019), and when examining real-life Malaise trap samples of terrestrial arthropods, Marquina et al. (2019) recovered significantly different arthropod communities with each approach, displaying little to almost no overlap between OTUs of the same samples. The authors therefore concluded that when dealing with Malaise trap samples, the ethanol-based DNA should not be used a sole substitute to tissue DNA, but at most be regarded as a complementary source of information (Marquina et al., 2019).

In this study, we also aim at comparing detected arthropod communities across methods but in a different context. In ecology, where researchers always work with subsets of communities, identical taxonomic recovery may not always be as crucial as the conservation of ecological and environmental information. Ji et al. (2013) were the first to examine the reliability of metabarcoding for depicting ecological trends among the homogenised tissue of arthropod communities. Since then, numerous studies have successfully applied destructive metabarcoding approaches to evaluate patterns in biodiversity (see Barsoum et al., 2019; Liu et al., 2021; Watts et al., 2019). In this respect, we aim to determine for the very first time whether environmental information is also conserved in the preservative ethanol of terrestrial arthropod bulk samples. We compare results of tissue homogenate metabarcoding with that of the preservative ethanol of the same samples to see whether we obtain similar ecological patterns among our communities. If this were the case, the preservative ethanol can in fact be regarded as a valuable non-destructive source of DNA for metabarcoding applications in environmental research. To answer our question, we set up Malaise traps to capture arthropod communities from different localities and habitats. For direct comparison we performed, for each bulk sample, metabarcoding on (1) the homogenised arthropod tissue and (2) the ethanol-based DNA.

MATERIALS AND METHODS

Arthropod sampling

In 2019, we installed eight Malaise traps in the Bavarian Forest National Park, which is located in southeast Germany along the border with the Czech Republic (Figure 1). Six traps, ranging from 650 to 800 m.a.s.l., were set in the catchment areas of the streams Kolbersbach, Grosse Ohe, and Kleine Ohe: one was installed directly above each stream using wooden beams, and one in the surrounding forest. Two further traps were installed in open meadows located in Kolbersbach and Bergerau. All traps were in operation from the end of April to September. The collection bottles were replaced every 2 weeks with new ones that were distinctive to the specific trap. All collection bottles had been bleached prior to the start of the experiment, and between collection events, the bottles were cleaned with distilled

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

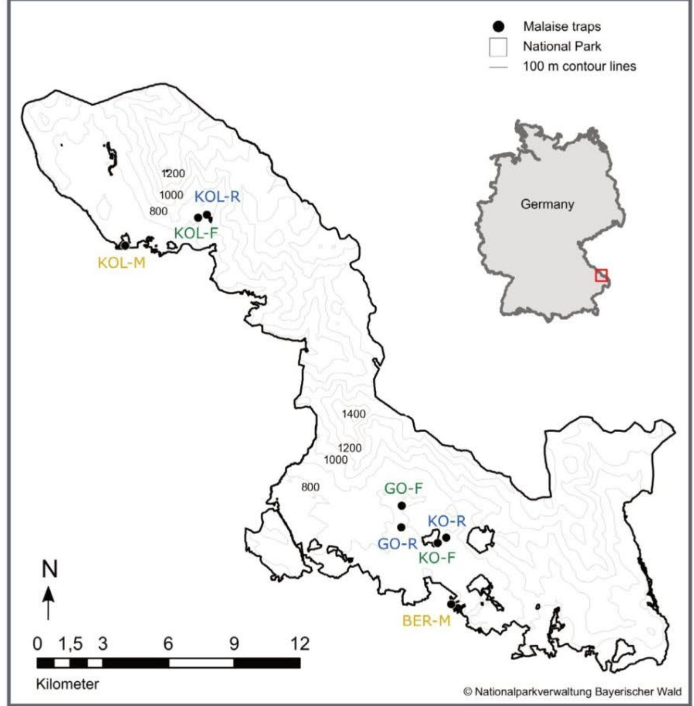


FIGURE 1 Location of the eight Malaise traps that were set up in the Bavarian Forest National Park and in operation from April to September 2019. KOL (Kolbersbach); GO (Große Ohe); KO (Kleine Ohe); BER (Bergerau); R (riparian); F (forest); M (meadow)

water and ethanol. The 80% ethanol (1 vol% MEK) was used for arthropod sampling.

Laboratory procedures

In the laboratory, we processed each sample individually to avoid cross-contaminations. We used cellulose tea bags to separate the arthropod tissue from its preservative ethanol (first phase ethanol used for sampling). A fresh bag was used for each sample. We weighed the tissue and transferred it to fresh 96% ethanol. We subsampled 50 ml of the ethanol (after thorough mixing) which we filtered (using sterile cellulose nitrate filters for vacuum filtration, 0.45 μm) and stored individually in 96% ethanol at −30°C until analysis (Advanced Identification Methods GmbH, Leipzig, Germany). The arthropod tissue and the ethanol filters were dried separately overnight in an oven at 60–70°C to remove all residual ethanol. We homogenised the arthropod tissue and the ethanol filters separately with stainless steel beads in a FastPrep 96 (MP Biomedicals) and used a 90:10 solution of animal lysis buffer (buffer ATL, Qiagen DNeasy Tissue Kit, Qiagen, Hilden, Germany) and Proteinase K for lysis, which was performed overnight in a 56°C oven. All samples were cooled to room temperature for subsequent DNA extraction. We took 200 μl aliquots of each lysate from which DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. PCR was performed using 5 μl of the extracted genomic DNA, 12.5 μl Plant MyTAQ (Bioline, Luckenwalde, Germany), and 1 μl HTS adapted mini-barcode primers mlCOIntF 5′–GGW ACW GGW TGA ACW GTW TAY CCY CC–3′ and dgHCO 5′–TAA ACT TCA GGG TGA CCA AAR AAY CA–3′ (see Morinière et al., 2016, 2019). We used the following PCR profile of 95°C for 5 min; 3 cycles of 96°C for 15 s; 48°C for 30 s; 65°C for 90 s; then 30 cycles of 96°C for 15 s; 55°C for 30 s; 65°C for 90 s and a final extension of 76°C for 10 min.

We examined amplification success and fragment lengths via gel electrophoresis, cleaned up the amplified DNA using ExoSap (Thermo Fisher), and resuspended it in 50 μl molecular grade water for each sample. Illumina Nextera XT (Illumina Inc., San Diego, USA) indices were indexed to the samples using a second PCR reaction. We used standard Illumina i5/i7 indices. Here, the same annealing temperature (55°C) was used as in the first PCR reaction, but with fewer cycles (7). Ligation success was confirmed by gel electrophoresis and DNA

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

concentrations were measured using a Qubit fluorometer (Life Technologies, Carlsbad, USA), which resulted in ~52 ng/μl for the tissue samples and ~24 ng/μl for the ethanol samples. We measured the DNA concentrations for each tagged sample, then pooled samples together (taking each PCR product into account) in order to obtain 40 μl pools that comprised concentrations of 100 ng/μl DNA each. The pools were purified using MagSi-NGSprep Plus (Steinbrenner Laborsysteme GmbH) beads. A final elution volume of 20 μl was used for HTS, which was performed on an Illumina MiSeq using v3 chemistry (2 × 300 bp, 600 cycles, maximum of 25 million paired-end reads). We aimed at obtaining 250 k RAW reads (125 k paired-end after merging) per sample. Overall, we used six negative controls per 96-well plate: two negative controls of DNA extractions, two amplicon PCR negative controls, and two indexing PCR negative controls.

Bioinformatic analysis

Briefly, we merged the paired-end reads using USEARCH v11.0.667_i86linux32 (Edgar, 2010). We trimmed adapters using CUTADAPT (Martin, 2011) and all reads that did not contain them were filtered out. Quality filtering, de-replication, chimera filtering, and clustering were carried out using the VSEARCH suite v2.9.1 (Rognes et al., 2016). We quality-filtered all reads containing more than one expected error per read, and then de-replicated them, first at the sample level, and then again at the combined dataset level after concatenating all sample files into one large FASTA file. This file was also filtered for singletons (reads that only occur once in the entire dataset). To save processing power, we pre-clustered the reads at 98% identity before chimera filtering using the VSEARCH centroids algorithm. As recommended by Rognes et al. (2016), we then carried out de novo chimera filtering, followed by the final round of clustering into OTUs at 97% identity.

In order to create the OTU table, the reads had to be mapped back to the created OTUs. To do this, we used a Perl script obtained from Rognes et al. (2016) to recover all quality- and chimera-filtered reads from the individual samples, including singletons, as well as reads that were previously removed by the two rounds of de-replication (https://github.com/torognes/vsearch/wiki/VSEARCH-pipeline). To reduce likely false positives, we excluded read counts in the OTU table that constituted less than 0.01% of the total number of reads in the sample. We then blasted the OTUs in Geneious (v.10.2.5; Biomatters, Auckland, New Zealand) and following methods described in the study by Morinière et al. (2016). We first blasted against a local copy of the NCBI nucleotide database (downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/) and then also against a custom database built from data downloaded from BOLD (www.boldsystems.org; Ratnasingham & Hebert, 2007, 2013), including taxonomy and BIN information. We exported the resulting CSV files from Geneious, including the OTU ID and NCBI/BOLD annotations for each detected OTU, and then combined them with the OTU table generated by the bioinformatic pre-processing pipeline. To provide another measure of control other than BLAST, we then classified OTUs into taxa using the Ribosomal Database Project (RDP) naïve Bayesian classifier (Wang et al., 2007) trained on a cleaned COI dataset of Arthropods and Chordates (plus outgroups; Porter & Hajibabaei, 2018). We filtered out all OTUs where the combined number of reads in the negative control samples constituted more than 20% of the total number of reads. Finally, we annotated the OTUs using NCBI taxonomic information (downloaded from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/).

Statistical analysis

All analyses were performed using R version 3.6.3 (R Core Team, 2012), and the packages *vegan* version 2.5-7 (Oksanen et al., 2020), *iNEXT* version 2.0.20 (Hsieh et al., 2020), *rtk* version 0.2.6.1 (Saary et al., 2017), *stats* version 3.4.3 (included in the standard R). An example R script and input data sets are deposited on Figshare (doi: https://doi.org/10.6084/m9.figshare.c.5666860.v2). We evaluated metabarcoding results of tissue- and ethanol-based DNA for all arthropods, then individually for each of the top five biodiverse arthropod orders in our dataset. For all arthropods and each individual order, we created an OTU × sample table with associated environmental variables (sites, habitats, seasonality) and sample type (tissue and ethanol). All reads were converted to presence/absence (Yu et al., 2012).

For statistical testing, the OTU dataset was rarefied to the lowest number of reads to equalise the sampling effort (via *rtk*; *rtk* package). To test whether community compositions differ based on associated environmental variables, we performed permutation multivariate analysis of variance (PERMANOVA) (via *adonis2*; *vegan* package; Jaccard dissimilarity method; 999 permutations). This method is best for testing compositional differences among multiple factors (Anderson, 2017). To differentiate between location and dispersion effects, we applied a beta dispersion test analogous to Levene's test (via *betadisper*; *vegan* package) and an *F*-test (via *permutest*; *vegan* package). In cases of unequal dispersion, we used a Tukey test (via *TukeyHSD*; *stats* package) to locate the variables responsible for inner group variation.

To visualise and compare environmental trends between the tissue- and the ethanol-based DNA communities, we used non-metric dimensional scaling ordinations (NMDS; via *metaMDS*; *vegan* package) or multidimensional scaling (PCoA; via *cmdscale*; *stats* package) of Jaccard dissimilarity matrices. We used the functions *vegdist* (to calculate a dissimilarity matrix), *ordiplot* (plotting function), *ordiellipse* (to add ellipses to ordination plot), and *ordispider* (to add spider graphs to the plot) from the *vegan* package. We created an ordination of each sample type (tissue and ethanol) for all arthropods, then for each of the top five most abundant arthropod orders.

We performed an alpha-diversity analysis (via *iNEXT*) of tissue- and ethanol-based DNA for the entire arthropod dataset. *iNEXT* uses observed sample incidence data (presence–absence data) to compute diversity estimates for sample-size and coverage-based rarefaction and extrapolation (R/E) curves using Hill numbers (Chao & Chiu, 2016): Indices such as the Shannon index and Simpson diversity have always been used by biologists to portray biological diversity in a

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Malaise trap sample information

Location	Habitat	Malaise traps	Malaise trap samples
Kleine Ohe	Forest	1	11
	Riparian	1	11
Große Ohe	Forest	1	11
	Riparian	1	11
Kolbersbach	Forest	1	11
	Riparian	1	11
	Meadow	1	10
Bergerau	Meadow	1	11

Note: Number of bulk samples obtained for each location and habitat type. Abbreviation: PERMANOVA, permutation multivariate analysis of variance.

given system; however, researchers have demonstrated that the nonlinearity of these metrics can mislead researchers when evaluating their results. Thus, diversity values were converted into equivalents, also known as Hill numbers, to overcome these shortcomings (Chao & Chiu, 2016; Cox et al., 2017; Jost, 2006). Hill numbers differ among themselves only by an exponent q, providing results for species richness ($q = 0$), Shannon diversity ($q = 1$) and Simpson diversity ($q = 2$). Chao and Jost (2012) established coverage-based R/E methods, which standardise samples by completeness rather than by size in sample-based approaches (see Colwell et al., 2012), which is highly dependent on the sampling effort. Integrating both approaches offers the best of both worlds: a consolidated framework for (1) estimating species richness and (2) statistical conclusions. For each sample type (tissue and ethanol), we constructed a list of presence–absence data for each habitat (samples × OTUs) to obtain the correct input format. All three measures of Hill numbers (q) were used in our analysis, but we only look at the species richness ($q = 0$) in this study. We created three plots for each sample type: a sample-size-based R/E curve plot, a sample completeness curve plot, and a coverage-based R/E plot.

RESULTS

Overall, we collected 87 Malaise trap samples throughout the season (Table 1). In total, 174 samples were sequenced: 87 tissue samples and their corresponding ethanol. From all samples, we detected 3636 OTUs belonging to six phyla, namely Arthropoda (3620 OTUs), Annelida (5 OTUs), Chordata (3 OTUs), Platyhelminthes (4 OTUs), Mollusca (3 OTUs), and Tardigrada (1 OTU). Limiting our analyses to arthropods, we recovered 2725 OTUs from tissue-based DNA, 1823 OTUs from ethanol-based DNA, and 934 (25.8%) from both (Figure 2a). These belong to 31 orders, of which the top five most diverse are (from most to least diverse): Diptera (1554 OTUs), Lepidoptera (610 OTUs), Hymenoptera (555 OTUs), Coleoptera (392 OTUs) and Hemiptera (132 OTUs) (Figure 2b). Together, these orders represent 89.6% of all arthropod OTUs. In total, 49.5% more arthropod OTUs were recovered from analysis of tissue-based DNA than ethanol-based DNA.

Tissue-based DNA sequencing results of all arthropods

PERMANOVA analysis found a significant difference in community compositions based on trap site, habitat type, and seasonality (all adonis2 $p = 0.001$) (Table 2). The measured significance among sites, however, also includes dispersion effects that are caused by uneven sample distribution among the trap site Große Ohe. Interaction effects were significant between habitats and sites (adonis2 $p = 0.002$), habitats and seasonality (adonis2 $p = 0.001$), and sites and seasonality (adonis2 $p = 0.032$) but not for all three together. Consistent with the statistical results, the NMDS plot (Figure 3a; Figure S1a) reveals clear distinctions in communities based on habitat type and along a chronological seasonal gradient. Differences based on trap sites are not as prominent.

Sample-size-based rarefaction curves show that the forests (1743 OTUs) are the richest habitats that we sampled, followed by riparian (1413 OTUs) and lastly the meadow habitats (1401 OTUs) (Figure 4a; left). Extrapolation to double the sampling units reveals that both terrestrial habitats (forest, meadow) display a higher species richness than the riparian habitats and that at least 25% more OTUs could have been obtained for each habitat type (forest +25.9% OTUs; riparian +26.9%; meadow +27.9%). Sample coverage was highest for samples collected in the forest (90.6%) and riparian (90.5%) habitats (Figure 4a; right). Doubling the sampling effort would not have provided a much higher coverage for these habitats. Sample coverage was lowest for the meadow landscapes (85.8%), which is due to the lower number of sampling units for this habitat type; our extrapolation curve shows that a very similar coverage would have been obtained with more sampling effort. Coverage-based rarefaction and extrapolation curves show similar results when comparing to sample-size-based R/E: the highest species richness was found among the forest habitats. Furthermore, species diversity within the terrestrial habitats was higher than that of the riparian habitats at equivalent coverage levels.

Ethanol-based DNA sequencing results of all arthropods

Hypothesis testing of the ethanol-based DNA results found no significant difference in community compositions based on trap site (adonis2 $p = 0.463$) nor habitat type (adonis2 $p = 0.073$; with dispersion effects) (Table 1). Testing for seasonality revealed a significant difference in community composition (adonis2 $p = 0.001$) with the inclusion of dispersion effects. Tukey testing revealed that samples collected from Week 18 to 24 are dispersed highly differently than those from Week 26 to 38. Interaction effects were significant between trap sites and seasonality (adonis2 $p = 0.005$). In the NMDS plot, samples are plotted into two distinct groups, with those on the left side being more dispersed than those on the right (Figure 3b; Figure S1b). There is no clear distinction between the different habitats nor between trap sites, and samples are not plotted along a chronological seasonal gradient. Furthermore, gradient lines for collection Weeks 32–38 are missing.

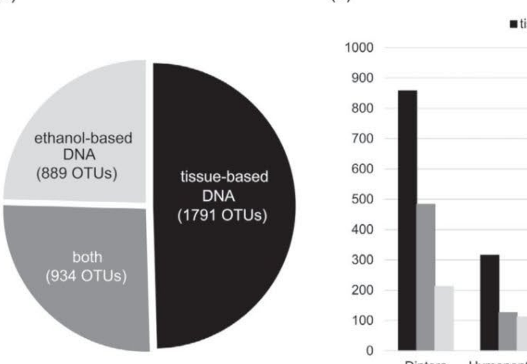


FIGURE 2 (a) Source of detected OTUs. Pie chart displaying the number of OTUs found with each metabarcoding approach. (b) OTU abundances per order. The bar chart shows results for the top five most abundant orders; together, these represent 89.6% of all arthropod OTUs detected throughout all samples.

TABLE 2 Statistical analysis of all arthropod OTUs (rarefied)

		PERMANOVA					Permutest
DNA source	Variables	Df	SS	R^2	*F*	Pr (>*F*)	*P*
Tissue-based DNA	Site	3	24.271	0.07801	3.0618	0.001***	0.001***
	Habitat	2	36.995	0.11890	7.0004	0.001***	0.891
	Week	1	28.128	0.09040	10.6448	0.001***	0.996
	Site:Habitat	2	0.9315	0.02994	1.7626	0.002**	
	Site:Week	3	10.294	0.03309	1.2986	0.039*	
	Habitat:Week	2	10.358	0.03329	1.9600	0.001***	
	Site:Habitat:Week	2	0.4163	0.01338	0.7877	0.913	
	Residuals	71	187.608	0.60299			
	Total	86	311.133	100.000			
Ethanol-based DNA	Site	3	0.6735	0.02768	0.9504	0.483	0.394
	Habitat	2	0.7430	0.03054	1.5726	0.054	0.023*
	Week	1	32.637	0.13413	13.8157	0.001***	0.002**
	Site:Habitat	2	0.3951	0.01624	0.8362	0.588	
	Site:Week	3	14.996	0.06163	2.1160	0.006**	
	Habitat:Week	2	0.5635	0.02316	1.1927	0.216	
	Site:Habitat:Week	2	0.4215	0.01732	0.8921	0.521	
	Residuals	71	167.725	0.68931			
	Total	86	243.325	100.000			

Note: Results of PERMANOVA (testing for differences in OTU community compositions) and permutation tests (P. test) via permutest (checking for homogeneity of multivariate dispersion) based on 999 permutations. Significance codes: 0 "***"; 0.001 "**"; 0.01 "*"; 0.05 "."; 1 " ".
Abbreviation: PERMANOVA, permutation multivariate analysis of variance.

Sample-size-based rarefaction curves show that the forests are the richest habitats that we sampled (1349 OTUs), followed by the riparian (1207 OTUs) and the meadow habitats (747 OTUs) (Figure 4b; left). Extrapolation to double the sampling units shows similar curves for the forest and riparian habitat, displaying a much higher species richness than the meadow habitats. At least 29% more species could have been obtained for each habitat (forest +30.3% OTUs; riparian + 29.8% OTUs; meadow +>36.8% OTUs) when doubling the sampling effort. Sample coverage was highest among the riparian habitats (91.7%), followed by the forest (91.6%) and lastly the

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

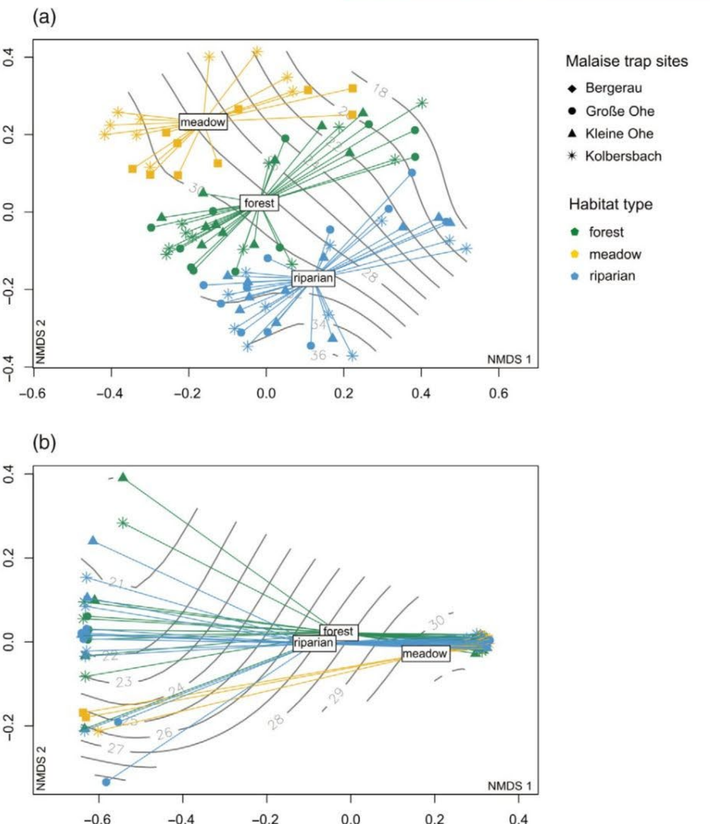


FIGURE 3 Non-metric dimensional scaling (NMDS) plots of arthropod community compositions of samples collected from four sites (Bergerau, Große Ohe, Kleine Ohe, Kolbersbach) covering three habitat types (riparian, forest and meadow). Sites are different symbols and habitats are different colours. Points nearest in plot space have similar species assemblages. In the NMDS plots, seasonality is displayed with ordisurf and ranges from calendar Week 18 to 39. (a) Arthopod communities of tissue-based DNA; NMDS of tissue-based DNA sequencing (3D analysis; stress = 0.1492). (b) Arthopod communities of ethanol-based DNA; NMDS of ethanol-based DNA sequencing (3D analysis; stress = 0.039). Ellipses are 95% CI of centroids for each sample type.

meadow (88.9%) habitats (Figure 4b; right). Doubling the sampling effort would not have provided a much higher coverage for these habitats.

Analysis of the most abundant orders

We performed individual statistical analyses for each of the top five most abundant arthropod orders in our dataset. For each of the five orders, analysis of the tissue-based OTUs depicted highly significant differences in community compositions based on each of the three environmental variables (Table 3). The majority of the significant results are driven by location effects only: For almost all orders, sample dispersion was homogenous among habitat types (exception: Hymenoptera) and among collection events. Sample dispersion, was not homogenous among site types for Diptera, Hymenoptera, and Coleoptera. Consistent with these statistical results, the NMDS/PCoA plots reveal clear distinctions in communities based on habitat type and along a chronological seasonal gradient (Figures 5a,c,e and 6c; Figures S2a,c,e and S3a,c). Sample clustering based on trap site is not clearly visible.

For every order, analysis of the ethanol-based OTUs displayed no significant differences in community compositions based on sites nor based on habitats (Table S1). Accordingly, in the NMDS/PCoA plots, there is no clear clustering as samples originating from different habitats and sites overlap one another (Figures 5b,d,f and 6b,d; Figures S2b,d,f and S3b,d). Testing for community differences based

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

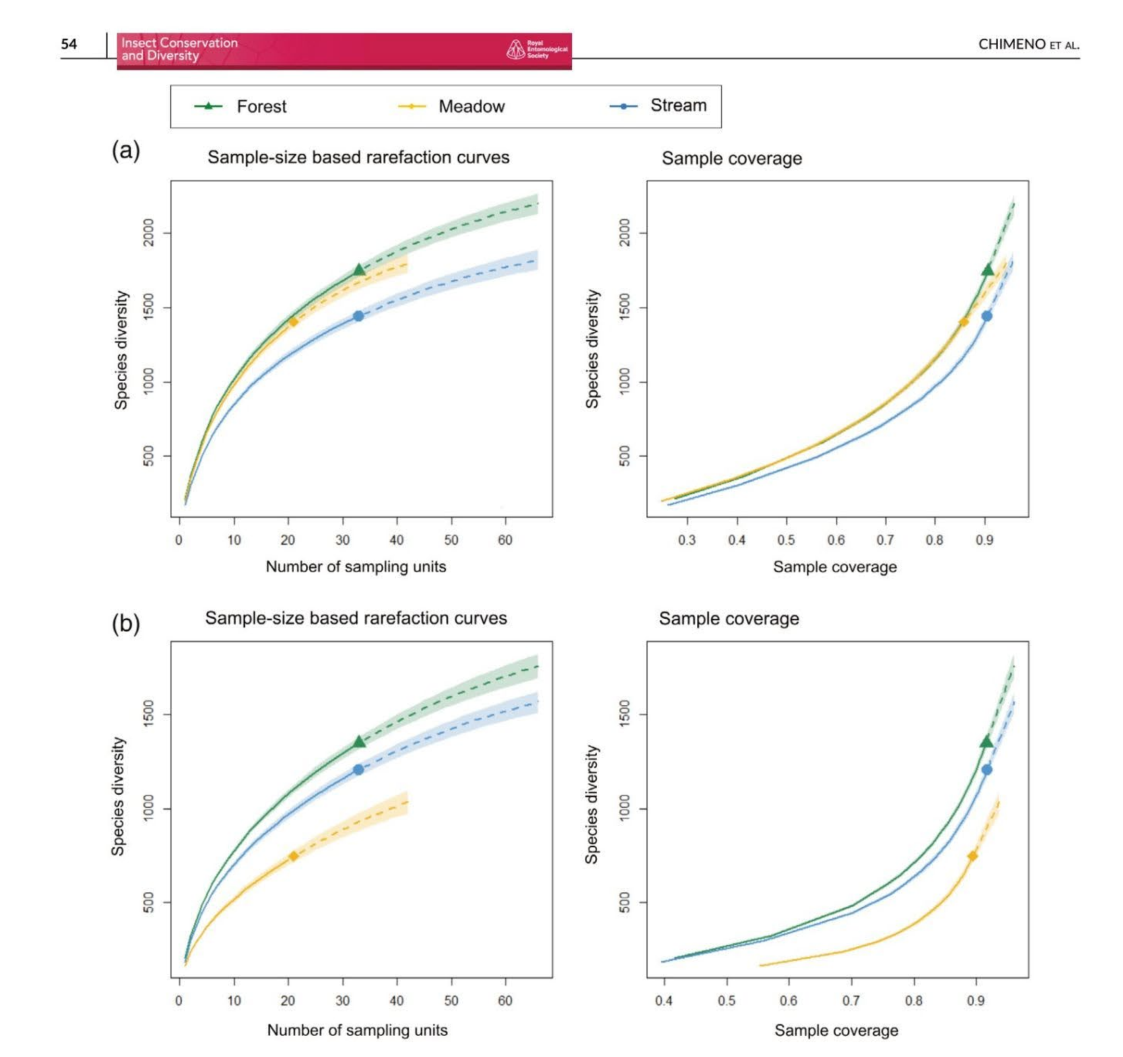


FIGURE 4 Rarefaction and extrapolation curves for $q = 0$ (species richness): (a) Arthopod communities of tissue-based DNA; sample-size-based rarefaction and extrapolation; (b) Arthopod communities of ethanol-based DNA; coverage levels for each habitat (meadow, forest, riparian). Solid lines represent rarefaction, while dashed lines represent extrapolation up to the double of sampling units. Shaded areas represent the 95% confidence interval using the bootstrap method on the basis of 100 repetitions.

on seasonality revealed highly significant results (all adonis2 $p = 0.001$) for all orders except for Hemiptera (adonis2 $p = 0.102$). Accordingly, samples of Hymenoptera and Coleoptera are plotted along a clear chronological seasonal gradient, however, less so for Diptera, and not at all for Lepidoptera. Samples of Lepidoptera were not homogenously dispersed throughout collection events. Although statistical analysis depicted no significant difference in community compositions of Hemiptera based on seasonality, samples are plotted along a chronological gradient in the Principal Coordinate Analysis (PCoA) ordination (Figure 6d; Figure S3d).

DISCUSSION

Discrepant arthropod communities

We detected completely different arthropod communities based on the DNA source used for sequencing (Figure 3c), which is congruent with findings of previous studies (see Elbrecht et al., 2017; Kirse et al., 2022; Marquina et al., 2019). For example, the preservative ethanol of an arthropod sample is more likely to contain the DNA of soft-bodied individuals because they release their DNA more freely into

TABLE 3 Statistical analysis of individual arthropod orders (rarefied)

Diptera	Variables	PERMANOVA Df	SS	R^2	F	Pr (>F)	Permutest P
Tissue-based DNA	Site	3	22.200	0.07504	2.8960	0.001***	0.001***
	Habitat	2	34.308	0.11597	6.7132	0.001***	0.709
	Week	1	26.279	0.08883	10.2844	0.001***	0.911
	Site:Habitat	2	0.9900	0.03347	1.9373	0.001***	
	Site:Week	3	0.9553	0.03229	1.2461	0.064	
	Habitat:Week	2	0.8364	0.02827	1.6365	0.006**	
	Site:Habitat:Week	2	0.3811	0.01288	0.7457	0.935	
	Residuals	71	181.423	0.61325			
	Total	86	295.838	100.000			
Ethanol-based DNA	Site	3	0.9106	0.02918	0.8816	0.771	0.353
	Habitat	2	0.7696	0.02466	1.1176	0.222	0.531
	Week	1	17.614	0.05643	5.1159	0.001***	0.905
	Site:Habitat	2	0.5214	0.01671	0.7572	0.937	
	Site:Week	3	13.953	0.04470	1.3509	0.029*	
	Habitat:Week	2	0.6922	0.02218	1.0052	0.435	
	Site:Habitat:Week	2	0.7158	0.02293	1.0395	0.346	
	Residuals	71	244.450	0.78321			
	Total	86	312.113	100.000			
Hymenoptera	**Variables**	**Df**	**SS**	R^2	F	**Pr (>F)**	P
Tissue-based DNA	Site	3	1.979	0.05536	1.7995	0.001***	0.005***
	Habitat	2	2.262	0.06327	3.0852	0.001***	0.01**
	Week	1	1.918	0.05366	5.2331	0.001***	0.84
	Site:Habitat	2	0.985	0.02755	1.3435	0.021*	
	Site:Week	3	1.084	0.03032	0.9857	0.511	
	Habitat:Week	2	0.705	0.01972	0.9618	0.567	
	Site:Habitat:Week	2	0.790	0.02210	1.0775	0.274	
	Residuals	71	26.029	0.72802			
	Total	86	35.753	100.000			
Ethanol-based DNA	Site	3	0.9530	0.03167	0.9032	0.632	0.534
	Habitat	2	0.7099	0.02359	1.0093	0.416	0.488
	Week	1	20.410	0.06783	5.8030	0.001***	0.649
	Site:Habitat	2	10.128	0.03366	1.4398	0.066	
	Site:Week	3	12.143	0.04036	1.1508	0.218	
	Habitat:Week	2	0.8514	0.02830	1.2104	0.176	
	Site:Habitat:Week	2	0.7971	0.02649	1.1331	0.251	
	Residuals	64	225.099	0.74810			
	Total	79	300.895	100.000			
Coleoptera	**Variables**	**Df**	**SS**	R^2	F	**Pr (>F)**	P
Tissue-based DNA	Site	3	1.918	0.05018	1.5507	0.002***	0.002***
	Habitat	2	1.938	0.05073	2.3513	0.001***	0.056*
	Week	1	2.107	0.05513	5.1109	0.001***	0.383
	Site:Habitat	2	0.867	0.02269	1.0517	0.321	
	Site:Week	3	1.320	0.03455	1.0677	0.227	

(Continues)

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 (Continued)

Coleoptera	Variables	Df	SS	R^2	F	Pr (>F)	P
	Habitat:Week	2	1.212	0.03171	1.4699	0.004***	
	Site:Habitat:Week	2	0.822	0.02151	0.9970	0.442	
	Residuals	68	28.031	0.73350			
	Total	83	38.215	100.000			
Ethanol-based DNA	Site	3	10.180	0.03768	1.0945	0.292	0.061
	Habitat	2	0.4528	0.01676	0.7303	0.870	0.219
	Week	1	15.747	0.05828	5.0793	0.001***	0.559
	Site:Habitat	2	0.4536	0.01679	0.7316	0.865	
	Site:Week	3	11.681	0.04323	1.2559	0.136	
	Habitat:Week	2	0.9149	0.03386	1.4754	0.065	
	Site:Habitat:Week	2	0.6636	0.02456	1.0702	0.357	
	Residuals	67	207.721	0.76883			
	Total	82	270.179	100.000			
Lepidoptera	**Variables**	**Df**	**SS**	**R^2**	**F**	**Pr (>F)**	**P**
Tissue-based DNA	Site	3	1.879	0.05602	1.4913	0.001***	0.108
	Habitat	2	1.734	0.05169	2.0642	0.001***	0.147
	Week	1	1.280	0.03817	3.0485	0.001***	0.884
	Site:Habitat	2	0.804	0.02398	0.9577	0.597	
	Site:Week	3	1.563	0.04660	1.2406	0.010**	
	Habitat:Week	2	1.362	0.04062	1.6222	0.001***	
	Site:Habitat:Week	2	0.981	0.02926	1.1684	0.093	
	Residuals	57	23.936	0.71367			
	Total	72	33.540	100.000			
Ethanol-based DNA	Site	3	0.5955	0.02709	0.9858	0.430	0.452
	Habitat	2	0.6804	0.03095	1.6893	0.075	0.028*
	Week	1	38.188	0.17369	18.9634	0.001***	0.013*
	Site:Habitat	2	0.3038	0.01382	0.7542	0.634	
	Site:Week	3	15.042	0.06841	2.4898	0.006**	
	Habitat:Week	2	0.4878	0.02219	1.2112	0.226	
	Site:Habitat:Week	2	0.2980	0.01355	0.7399	0.666	
	Residuals	71	142.978	0.65031			
	Total	86	219.863	100.000			
Hemiptera	**Variables**	**Df**	**SS**	**R^2**	**F**	**Pr (>F)**	**P**
Tissue-based DNA	Site	3	27.244	0.10038	2.5727	0.001***	0.313
	Habitat	2	20.499	0.07553	2.9036	0.001***	0.19
	Week	1	0.7811	0.02878	2.2128	0.009**	0.959
	Site:Habitat	2	0.8555	0.03152	1.2118	0.190	
	Site:Week	3	13.171	0.04853	1.2437	0.121	
	Habitat:Week	2	0.8404	0.03096	1.1904	0.207	
	Site:Habitat:Week	2	0.9230	0.03401	1.3074	0.113	
	Residuals	50	176.493	0.65029			
	Total	65	271.405	100.000			
Ethanol-based DNA	Site	3	11.624	0.06741	0.8683	0.784	0.712
	Habitat	2	0.8268	0.04794	0.9264	0.623	0.889

(Continues)

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 (Continued)

Hemiptera	Variables	Df	SS	R^2	F	Pr (>F)	P
	Week	1	0.5979	0.03467	1.3398	0.118	0.004**
	Site:Habitat	2	0.7798	0.04522	0.8737	0.721	
	Site:Week	3	16.973	0.09843	1.2679	0.054	
	Habitat:Week	2	0.7129	0.04134	0.7988	0.882	
	Site:Habitat:Week	2	0.7580	0.04395	0.8493	0.774	
	Residuals	24	107.098	0.62104			
	Total	39	172.449	100.000			

Note: Results of PERMANOVA (testing for differences in OTU community compositions) and permutation tests (P. test) via permutest (checking for homogeneity of multivariate dispersion) based on 999 permutations. Significance codes: 0 "***"; 0.001 "**"; 0.01 "*"; 0.05 "."; 1 " ".
Abbreviation: PERMANOVA, permutation multivariate analysis of variance.

the preservative fluid than higher sclerotised individuals (Elbrecht et al., 2017; Morinière et al., 2016). In general, small-bodied or fragile individuals are also more apt to be detected in the preservative fluid because their bodies (or detached parts) may pass through the mesh of the filter during sample processing (Marquina et al., 2019). Comparing each community, we find that metabarcoding results of tissue DNA resulted in sample compositions that are coherent with typical catchings of Malaise traps: predominantly Diptera and Hymenoptera, followed by other orders in much lower abundances (Geiger et al., 2016; Gressitt & Gressitt, 1962; Karlsson et al., 2020; Matthews & Matthews, 2017; Moeed & Meads, 1987; Schmidt et al., 2019; Skvarla, 2015). In contrast, we recovered a strikingly high proportion of Lepidoptera when metabarcoding the ethanol-based DNA. Of all lepidopteran OTUs that we recovered in total, more than half of these were detected exclusively in the ethanol-based DNA, making Lepidoptera the most abundant order after Diptera. Interestingly, Lepidoptera is also the only (abundant) order for which we recovered more OTUs from the ethanol-based DNA than from the tissue DNA. We believe that this may be explained by several interacting factors: First, Lepidoptera possess soft-bodied abdomens, meaning that the DNA of these individuals is easily released into the preservative ethanol (Elbrecht et al., 2017; Morinière et al., 2016). Second, a large proportion of Germany's lepidopteran fauna are small-bodied microlepidoptera meaning that the DNA of these individuals is more likely to be concealed by that of larger ones in the tissue (Herrich-Schäffer & Hübner, 1843; Marquina et al., 2019). Thus, these individuals are likely underrepresented in the tissue and overrepresented in the ethanol. Third, both macro- and microlepidoptera serve as important food sources for other arthropods (Strazanac & Butler, 2005). Because some species are known to regurgitate their stomach contents when coming in contact with ethanol (Marquina et al., 2019), we believe that a substantial proportion of lepidopteran OTUs recovered in the ethanol may in fact be gut-based DNA.

Depicting ecological gradients

As expected, we found that communities recovered from the tissue DNA depicted clear biodiversity patterns based on environmental factors (see Barsoum et al., 2019; Liu et al., 2021; Watts et al., 2019). All statistical tests that we performed on tissue DNA revealed highly significant differences in communities for all three variables (sites, habitats, seasonality) individually, but also as a result of interaction effects. We created ordinations to obtain visual overviews of the sample data and in all cases, the environmental trends depicted in the plots were coherent with the statistical results. Environmental trends were strongest for Diptera, which was expected because Malaise traps are very efficient at catching flies; hence, sample size and sample representativeness are much higher for this order than for others (e.g. Coleoptera, Lepidoptera) (Matthews & Matthews, 2017).

Metabarcoding the ethanol-based DNA of the same Malaise trap samples demonstrated that ecological trends were only partly conserved in the preservative fluid. Habitats and sites had no effect on community compositions, but seasonality did. For all orders (except Hemiptera), statistical analysis depicted highly significant differences in communities driven by seasonality (adonis2 $p = 0.001$). Seasonal gradients were strongest among Hymenoptera and Coleoptera, and permutation testing validated that these differences were only driven by location effects. Gradients were not as prominent for Diptera and not at all visible for Lepidoptera. Although statistical testing found that seasonality had a significant effect on lepidopteran communities, we believe that this result is strongly driven by dispersion effects and that we may be dealing with a type II statistical error. Permutation testing revealed that communities collected in the first half of the season were more dispersed than those collected in the second half (permu $p = 0.009$), and the box plot of Tukey's results displayed absolutely no overlap between these groups. Interestingly, although we measured no significant difference in hemipteran community compositions based on seasonality, samples are clearly plotted along a chronological gradient in the ordination. In this case, we suspect that we may be possibly dealing with a type I statistical error, but further analyses are needed.

We are not certain as to why seasonal trends in the ethanol are better conserved among some groups and lesser so among others. However, we speculate that a group's trophic level may have a meaningful impact, as arthropod specimens that fall prey to other arthropods

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

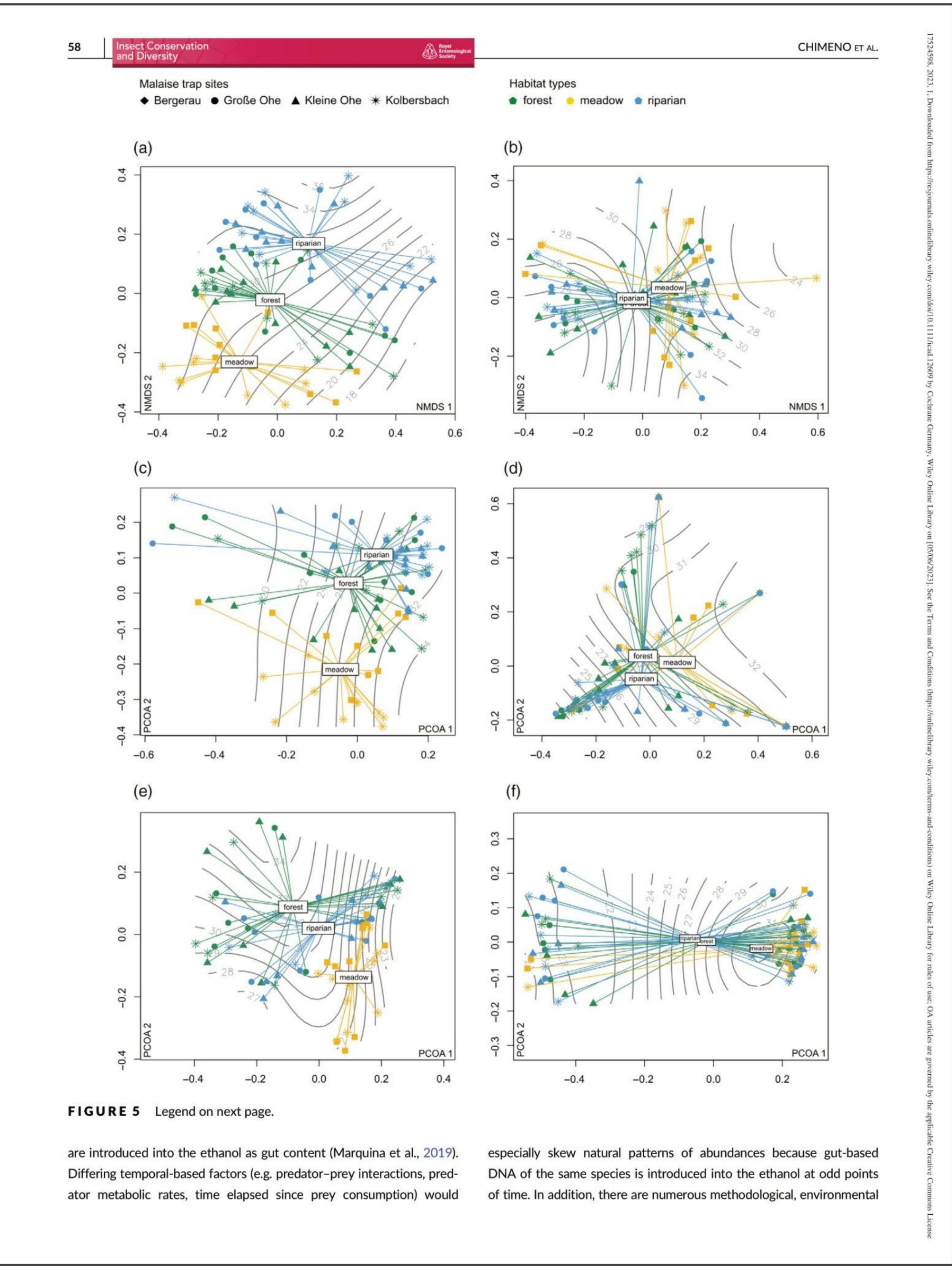


FIGURE 5 Legend on next page.

are introduced into the ethanol as gut content (Marquina et al., 2019). Differing temporal-based factors (e.g. predator–prey interactions, predator metabolic rates, time elapsed since prey consumption) would especially skew natural patterns of abundances because gut-based DNA of the same species is introduced into the ethanol at odd points of time. In addition, there are numerous methodological, environmental

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

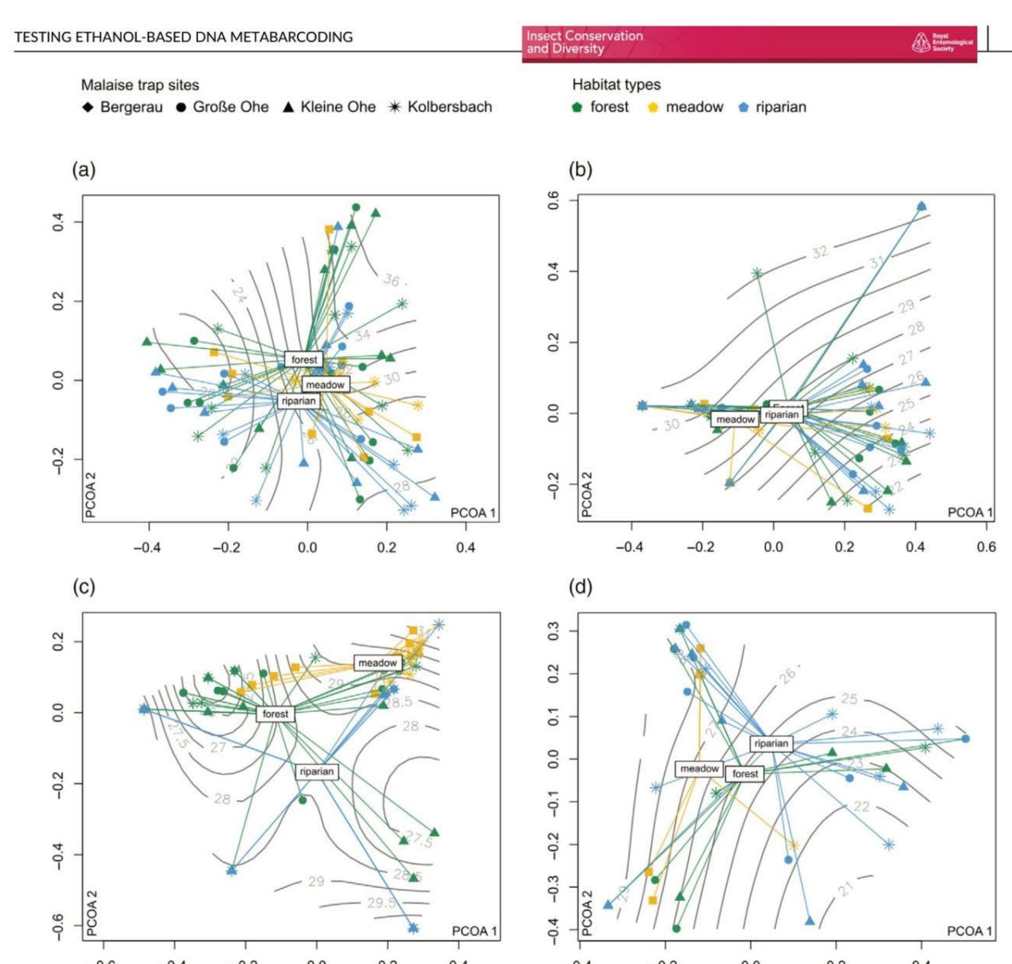


FIGURE 6 Non-metric dimensional scaling (NMDS)/PCoA plots of individual orders. Coleoptera: (a) tissue-based DNA sequencing (PCoA); (b) ethanol-based DNA sequencing (PCoA). Hemiptera: (c) tissue-based DNA sequencing PCoA; (d) ethanol-based DNA sequencing PCoA. Samples collected from four sites (Bergerau, Große Ohe, Kleine Ohe, Kolbersbach) covering three habitat types (riparian, forest and meadow). Sites are different symbols and habitats are different colours. Points nearest in plot space have similar species assemblages. In the NMDS plots, seasonality is displayed with ordisurf and ranges from calendar Week 18 to 39.

and biological/physiological factors that have a direct influence on success rates of gut content sequencing (Eitzinger et al., 2013; Greenstone et al., 2010; von Berg et al., 2008). With too many sources of bias that are introduced into the analysis of ethanol-based DNA, and no possibility of discriminating between ingested and captured arthropods, seasonal patterns are especially prone to distortion among groups that include many prey species. In our study, seasonal gradients were best depicted among Hymenoptera, Coleoptera, and Hemiptera, but lesser

FIGURE 5 NMD/PCoA plots of individual orders. Diptera: (a) tissue-based DNA sequencing (3D analysis; stress = 0.1649); (b) ethanol-based DNA sequencing (2D analysis; stress = 0.1533). Hymenoptera: (c) tissue-based DNA sequencing (PCoA); (d) ethanol-based DNA sequencing (PCoA) and Lepidoptera: (e) tissue-based DNA sequencing (PCoA); (f) ethanol-based DNA sequencing (PCoA). Samples collected from four sites (Bergerau, Große Ohe, Kleine Ohe, Kolbersbach) covering three habitat types (riparian, forest and meadow). Sites are different symbols and habitats are different colours. Points nearest in plot space have similar species assemblages. In the non-metric dimensional scaling (NMDS) plots, seasonality is displayed with ordisurf and ranges from calendar Week 18 to 39.

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

so (or not at all) for Diptera and Lepidoptera. We believe that because the former orders encompass species that are less susceptible to falling prey to other arthropods, they are also less likely to be introduced into the ethanol of our samples as gut content. Typical predators of Coleoptera, Hymenoptera, and Hemiptera are, for example, birds, bats, and frogs (Britannica, 2022). Other arthropods that predate on these taxa include Odonata and Araneae, both of which are lesser represented in our dataset. In contrast, predators of Diptera and Lepidoptera are very well represented in our Malaise trap samples, as these include many taxa of Hymenoptera, Coleoptera, Diptera, and Araneae (Flint & Dreistadt, 1998).

Sequencing ethanol-based DNA failed at depicting spacial patterns. We detected no significant differences among trap sites nor among habitats for all orders. Consistent with previous findings, alpha-diversity assessment demonstrated that the ethanol-based DNA (1) failed at discriminating between the terrestrial and riparian habitats and (2) underrepresented the magnitude of arthropod diversity within every single habitat (see Erdozain et al., 2019; Linard et al., 2016). Recently, Zenker et al. (2020) conducted DNA metabarcoding exclusively on the preservative ethanol of automatic light trap samples to compare the alpha and beta diversity of arthropod communities in Brazil. Unfortunately, they did not examine or process the tissue of these samples at all, so no reference was available as a guideline to their interpretations. Observing our alpha-diversity curves, we strongly believe that the sole use of preservative ethanol can clearly lead to false conclusions, and we therefore discourage its sole use until further research has been conducted.

Overall, we find that ethanol-based DNA sequencing did not provide information on ecological gradients, except for the case of seasonal patterns. The conserved seasonality among some taxa is an interesting starting point for further investigations but until more research has provided more successful results, we recommend researchers dealing with terrestrial ecosystems to be careful when using ethanol-based DNA. It is important to note that in this study, we used 80% ethanol (1 vol% MEK) for arthropod sampling. We conducted DNA extractions in spring 2020 following the collection season (April–October 2019). According to Marquina et al. (2021), this concentration of ethanol is too low for ideal DNA preservation over time. We therefore highly encourage others to use 95% ethanol for sampling to guarantee optimal DNA preservation.

Non-destructive DNA extractions as a promising alternative

A striking subject of today's (and the future's) research concerns the advancing methodology of non-destructive DNA extractions. Numerous studies dedicated to the development of non-destructive methodologies for sequencing are emerging, showing that it is possible to extract DNA (although in smaller quantities) from specimens while keeping their structural integrity intact (Batovska et al., 2021; Carew et al., 2018; Kirse et al., 2022; Marquina et al., 2022; Martins et al., 2019; Martoni et al., 2022; Nielsen et al., 2019). Such protocols roughly consist of leaching DNA from whole individuals by temporarily submerging them in a digestive buffer (Castalanelli et al., 2010; Krosch & Cranston, 2012; Nielsen et al., 2019; Porco et al., 2010; Wong et al., 2014). While various studies have tested non-destructive DNA extractions on single arthropod specimens or samples of mock communities (see Castalanelli et al., 2010; Marquina et al., 2022; Nielsen et al., 2019), we only found one study that did so on real-life bulk samples of terrestrial arthropods from Malaise traps (see Kirse et al., 2022). Malaise traps are especially challenging to process as they can contain hundreds to thousands of individuals (Geiger et al., 2016), each displaying various degrees of sclerotisation, which require different incubation times for adequate non-destructive DNA extraction (Elbrecht et al., 2017). Moreover, there are many options in which non-destructive DNA extractions can be performed, ranging from an optional step of sample sorting, to the choice of lysis buffer, to incubation times of specimen in the fluid, to the protocol used for extraction (Kirse et al., 2022; Marquina et al., 2022; Martoni et al., 2022). With so many factors, numerous researchers are in the process of testing these different options in determining which combination is most effective. One very recent study is especially interesting as the authors conducted comparative analysis on real-life (however sorted) Malaise trap samples (see Kirse et al., 2022). The authors were able to demonstrate that when choosing the right protocol, non-destructive analysis can provide comparable results in terms of species richness and community composition.

On the basis of these results, we believe that in time, non-destructive DNA extractions will become the preferred technique for obtaining DNA from terrestrial arthropod bulk samples. Not only is the sample integrity conserved for further studies, this technique is also quick and provides a lower contamination risk in comparison to traditional tissue-based approaches (Kirse et al., 2022). On this note, we highly encourage future work to test whether ecological trends are also conserved in the OTUs recovered from such analyses. We strongly believe that this is the case as Kirse and authors have shown that they recovered comparable OTU communities in their study using both methods.

CONCLUSION

Returning to the topic of ethanol-based DNA, we recommend researchers dealing with terrestrial ecosystems to be careful when using this approach. These results are not comparable to those obtained using the traditional destructive approaches. However, we do invite researchers in the field of aquatic ecology to look into our research question. Overall, preservative ethanol sequencing on aquatic macroinvertebrates has provided better results as these communities are dominated by soft-bodied specimens—thus, it would be expected that environmental trends are better conserved in the ethanol of such samples.

AUTHOR CONTRIBUTIONS

Caroline Chimeno: Formal analysis (lead); software (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead).

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Jeremy Hübner: Formal analysis (equal); methodology (supporting); writing – original draft (equal); writing – review and editing (supporting). **Linda Seifert:** Investigation (lead); methodology (lead); writing – review and editing (supporting). **Jérôme Morinière:** Methodology (lead); software (equal); supervision (lead); writing – original draft (supporting); writing – review and editing (supporting). **Vedran Bozicevic:** Data curation (equal); methodology (equal); software (equal); writing – original draft (supporting); writing – review and editing (supporting). **Axel Hausmann:** Supervision (equal); writing – original draft (supporting); writing – review and editing (supporting). **Stefan Schmidt:** Supervision (equal); writing – original draft (supporting); writing – review and editing (supporting). **Jörg Müller:** Conceptualization (equal); funding acquisition (equal); project administration (equal); supervision (equal); writing – review and editing (equal).

ACKNOWLEDGEMENTS

The authors would like to thank the National Park Authority for funding the metabarcoding of our Malaise trap samples and Josef Nußhardt for running the traps. The authors would also like to thank the Federal Ministry of Education and Research (BMBF) for funding the German Barcode of Life Project: Dark Taxa, in which framework this study has been evaluated, written, and submitted. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available on Figshare under the following dois:Chimeno et al. (2022): R Script accompanying Manuscript. figshare. Software (https://doi.org/10.6084/m9.figshare.20222322.v2); Chimeno et al. (2022): Entire Arthropod Dataset (Input data for R script).xlsx. figshare. Dataset (https://doi.org/10.6084/m9.figshare.19397132.v1); Chimeno et al. (2022): Sample Metadata. figshare. Dataset (https://doi.org/10.6084/m9.figshare.19377122.v1); Chimeno et al. (2022): OTU Table. figshare. Dataset (https://doi.org/10.6084/m9.figshare.15029157.v2); Chimeno et al. (2022): Fastq Files. figshare. Dataset (https://doi.org/10.6084/m9.figshare.19376666.v1).

ORCID

Caroline Chimeno https://orcid.org/0000-0001-9933-6492

REFERENCES

Anderson, M.J. (2017) Permutational multivariate analysis of variance (PERMANOVA). In: *Wiley StatsRef: statistics reference online*, pp. 1–15. https://doi.org/10.1002/9781118445112.stat07841

Aylagas, E., Borja, Á., Irigoien, X. & Rodríguez-Ezpeleta, N. (2016) Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Frontiers in Marine Science*, 3, 96. https://doi.org/10.3389/fmars.2016.00096

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

Barsoum, N., Bruce, C., Forster, J., Ji, Y.-Q. & Yu, D.W. (2019) 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. *Ecological Indicators*, 101, 313–323. https://doi.org/10.1016/j.ecolind.2019.01.023

Batovska, J., Piper, A.M., Valenzuela, I., Cunningham, J.P. & Blacket, M.J. (2021) Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches. *Scientific Reports*, 11, 7946. https://doi.org/10.1038/s41598-021-85855-6

Bohan, D.A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A.J. & Woodward, G. (2017) Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in Ecology & Evolution*, 32, 477–487. https://doi.org/10.1016/j.tree.2017.03.001

Britannica. (2022) *coleopteran*. Available at: https://www.britannica.com/animal/beetle [Accessed 17th March 2022].

Carew, M.E., Coleman, R.A. & Hoffmann, A.A. (2018) Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? *PeerJ*, 6, e4980. https://doi.org/10.7717/peerj.4980

Castalanelli, M.A., Severtson, D.L., Brumley, C.J., Szito, A., Foottit, R.G., Grimm, M. et al. (2010) A rapid non-destructive DNA extraction method for insects and other arthropods. *Journal of Asia-Pacific Entomology*, 13, 243–248. https://doi.org/10.1016/j.aspen.2010.04.003

Chao, A. & Chiu, C.-H. (2016) Nonparametric estimation and comparison of species richness. *Els*, 1–11. https://doi.org/10.1002/9780470015902.a0026329

Chao, A. & Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93, 2533–2547. https://doi.org/10.1890/11-1952.1

Chimeno, C., Hausmann, A., Schmidt, S., Raupach, M.J., Doczkal, D., Baranov, V. et al. (2022) Peering into the darkness: DNA barcoding reveals surprisingly high diversity of unknown species of Diptera (Insecta) in Germany. *Insects*, 13, 82. https://doi.org/10.3390/insects13010082

Coissac, E., Riaz, T. & Puillandre, N. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, 21, 1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. et al. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5, 3–21. https://doi.org/10.1093/jpe/rtr044

Cox, K.D., Black, M.J., Filip, N., Miller, M.R., Mohns, K., Mortimor, J. et al. (2017) Community assessment techniques and the implications for rarefaction and extrapolation with hill numbers. *Ecology and Evolution*, 7, 11213–11226. https://doi.org/10.1002/ece3.3580

Cristescu, M.E. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29, 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, 10, 20140562. https://doi.org/10.1098/rsbl.2014.0562

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

Eitzinger, B., Micic, A., Körner, M., Traugott, M. & Scheu, S. (2013) Unveiling soil food web links: new PCR assays for detection of prey DNA in the gut of soil arthropod predators. *Soil Biology and Biochemistry*, 57, 943–945. https://doi.org/10.1016/j.soilbio.2012.09.001

Elbrecht, V. & Leese, F. (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

17524598, 2023, 1, Downloaded from https://resjournals.onlinelibrary.wiley.com/doi/10.1111/icad.12609 by Cochrane Germany, Wiley Online Library on [05/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sequence relationships with an innovative metabarcoding protocol. *PLoS One*, 10, e0130324. https://doi.org/10.1371/journal.pone.0130324

Elbrecht, V., Peinert, B. & Leese, F. (2017) Sorting things out - assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, 7, 6918–6925. https://doi.org/10.7287/peerj.preprints.2561

Erdozain, M., Thompson, D.G., Porter, T.M., Kidd, K.A., Kreutzweiser, D.P., Sibley, P.K. et al. (2019) Metabarcoding of storage ethanol vs. conventional morphometric identification in relation to the use of stream macroinvertebrates as ecological indicators in forest management. *Ecological Indicators*, 101, 173–184. https://doi.org/10.1016/j.ecolind.2019.01.014

Flint, M.L. & Dreistadt, S.H. (1998) *Natural enemies handbook: the illustrated guide to biological Pest control.* Oakland: University of California Press.

Geiger, M.F., Moriniere, J., Hausmann, A., Haszprunar, G., Wägele, W., Hebert, P.D.N. et al. (2016) Testing the global malaise trap program – how well does the current barcode reference library identify flying insects in Germany? *Biodivers Data Journal*, 4, e10671. https://doi.org/10.3897/BDJ.4.e10671

Greenstone, M.H., Szendrei, Z., Payton, M.E., Rowley, D.L., Coudron, T.C. & Weber, D.C. (2010) Choosing natural enemies for conservation biological control: use of the prey detectability half-life to rank key predators of Colorado potato beetle: ranking predators for biocontrol. *Entomologia Experimentalis et Applicata*, 136, 97–107. https://doi.org/10.1111/j.1570-7458.2010.01006.x

Gressitt, J.L. & Gressitt, M.K. (1962) An improved Malaise trap. *Pacific Insects*, 4, 87–90.

Hajibabaei, M., Spall, J.L., Shokralla, S. & van Konynenburg, S. (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, 12, 1–10. https://doi.org/10.1186/1472-6785-12-28

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

Hebert, P.D.N., Cywinska, A., Ball, S.L. & de Waard, J.R. (2003) Biological identifications through DNA barcodes. *Proceeding of the Royal Society of London. Series B: Biological Sciences*, 270, 313–321. https://doi.org/10.1098/rspb.2002.2218

Herrich-Schäffer, G.A.W. & Hübner, J. (1843) *Systematische Bearbeitung der Schmetterlinge von Europa, zugleich als Text, Revision und Supplement zu Jakob Hübner's Sammlung europäischer Schmetterlinge*. Regensburg: In Commission bei G. J. Manz.

Hsieh, T.C., Chao, K.H.M. & Anne, C. (2020) iNEXT: interpolation and extrapolation for species diversity. Available at: http://chao.stat.nthu.edu.tw/wordpress/software_download/

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A. et al. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16, 1245–1257. https://doi.org/10.1111/ele.12162

Jost, L. (2006) Entropy and diversity. *Oikos*, 113, 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Karlsson, D., Forshage, M., Holston, K. & Ronquist, F. (2020) The data of the Swedish malaise trap project, a countrywide inventory of Sweden's insect fauna. *Biodiversity Data Journal*, 8, e56286. https://doi.org/10.3897/BDJ.8.e56286

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F. & Bouchez, A. (2017) Freshwater biomonitoring in the information age. *Frontiers in Ecology and the Environment*, 15, 266–274. https://doi.org/10.1002/fee.1490

Kirse, A., Bourlat, S., Langen, K., Zapke, B. & Zizka, V. (2022) Comparison of destructive and non-destructive DNA extraction methods for the metabarcoding of arthropod bulk samples. *Molecular Ecology Resources*, 05. https://doi.org/10.1111/1755-0998.13694

Krosch, M.N. & Cranston, P.S. (2012) Non-destructive DNA extraction from Chironomidae, including of fragile pupal exuviae, extends analysable collections and enhances vouchering. *CHIRONOMUS Journal of Chironomidae Research*, 25, 22–27. https://doi.org/10.5324/cjcr.v0i25.1532

Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A. & Vogler, A.P. (2016) Lessons from genome skimming of arthropod-preserving ethanol. *Molecular Ecology Research*, 16, 1365–1377. https://doi.org/10.1111/1755-0998.12539

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

Liu, M., Jordan, G.J., Burridge, C.P., Clarke, L.J. & Baker, S.C. (2021) Metabarcoding reveals landscape drivers of beetle community composition approximately 50 years after timber harvesting. *Forest Ecology and Management*, 488, 119020. https://doi.org/10.1016/j.foreco.2021.119020

Makiola, A., Compson, Z.G., Baird, D.J., Barnes, M.A., Boerlijst, S.P., Bouchez, A. et al. (2020) Key questions for next-generation biomonitoring. *Frontiers in Environmental Science*, 7, 197. https://doi.org/10.3389/fenvs.2019.00197

Mandelik, Y., Roll, U. & Fleischer, A. (2010) Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *Journal of Applied Ecology*, 47, 1179–1188. https://doi.org/10.1111/j.1365-2664.2010.01864.x

Marquina, D., Buczek, M., Ronquist, F. & Łukasik, P. (2021) The effect of ethanol concentration on the morphological and molecular preservation of insects for biodiversity studies. *PeerJ*, 9, e10799. https://doi.org/10.1101/2020.03.05.978288

Marquina, D., Esparza-Salas, R., Roslin, T. & Ronquist, F. (2019) Establishing arthropod community composition using metabarcoding: surprising inconsistencies between soil samples and preservative ethanol and homogenate from malaise trap catches. *Molecular Ecology Resources*, 19, 1516–1530. https://doi.org/10.1111/1755-0998.13071

Marquina, D., Roslin, T., Łukasik, P. & Ronquist, F. (2022) Evaluation of non-destructive DNA extraction protocols for insect metabarcoding: gentler and shorter is better. *Metabarcoding & Metagenomics*, 6, 187–201. https://doi.org/10.3897/mbmg.6.78871

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17, 10–12.

Martins, F.M.S., Galhardo, M., Filipe, A.F., Teixeira, A., Pinheiro, P., Paupério, J. et al. (2019) Have the cake and eat it: optimizing nondestructive DNA metabarcoding of macroinvertebrate samples for freshwater biomonitoring. *Molecular Ecology Resources*, 19, 863–876. https://doi.org/10.1111/1755-0998.13012

Martoni, F., Piper, A.M., Rodoni, B.C. & Blacket, M.J. (2022) Disentangling bias for non-destructive insect metabarcoding. *PeerJ*, 10, e12981. https://doi.org/10.7717/peerj.12981

Matthews, R.W. & Matthews, J.R. (2017) The Malaise trap: its utility and potential for sampling insect populations. *The Great Lakes Entomologist*, 4, 7.

Medhi, J., Dutta, J. & Chandra Kalita, M. (2021) Biomonitoring ecosystem: modelling relationship with arthropods. In: Eduardo Rebolledo Ranz, R. (Ed.) *Arthropods - are they beneficial for mankind?* IntechOpen. https://doi.org/10.5772/intechopen.94313

Meusnier, I., Singer, G.A., Landry, J.-F., Hickey, D.A., Hebert, P.D. & Hajibabaei, M. (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, 9, 214. https://doi.org/10.1186/1471-2164-9-214

Moeed, A. & Meads, M.J. (1987) Seasonality of arthropods caught in a malaise trap in mixed lowland forest of the Orongorongo Valley,

After having read this thesis, I do have one question.

“Do you understand the words that are coming out of my mouth?”

– Detective Carter from Rush Hour