# Development and Improvement of Next Generation Sequencing Pipelines for Mixed and Bulk Samples of German Fauna

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Biologie der Ludwig-Maximilians-Universität München


staatliche
naturwissenschaftliche
sammlungen bayerns


Vorgelegt von Laura Anne Hardulak
München, 2020

Diese Dissertation wurde angefertigt
unter der Leitung von Prof. Dr. Gerhard Haszprunar
im Bereich von der Zoologische Staatssammlung München
an der Ludwig-Maximilians-Universität München

1. Gutachter: Prof. Dr. Gerhard Haszprunar
2. Gutachter: Prof. Dr. Roland Melzer

Tag der Abgabe: 23.07.2020

Tag der mündlichen Prüfung: 12.10.2020

**Eidesstattliche Erklärung**

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

München, den 16.6.2020

_____

Laura Anne Hardulak

**Erklärung**

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 16.6.2020

_____

Laura Anne Hardulak

Work presented in this dissertation was performed in the laboratory of Dr. Gerhard Haszprunar, Director of the Bavarian State Collection of Zoology, Munich, Germany.

Work was performed under the supervision of Prof. Dr. Gerhard Haszprunar, Director of Bavarian State Collection of Zoology, Munich, Germany and Dr. Axel Hausmann, Head of Department Entomology of the Bavarian State Collection of Zoology, Munich, Germany.

# Table of Contents

# Abbreviations

BIN……..……..……..…….Barcode Index Number
BOLD……..……..……..Barcode of Life Database
bp……..……..……...……...…….base pair
eDNA……..……..…….environmental Deoxyribonucleic Acid
GMTP……..……..……..…..Global Malaise Trap Program
HTS…….....High Throughput Sequencing
LGL……..……..……..…..Landesamt für Gesundheit und Lebensmittelsicherheit
MOTU ……...……..……..…… Molecular Operational Taxonomic Unit
NGS……..……..……..…….Next Generation Sequencing
NJ……...……..……..……..……...Neighbor Joining
NPBW ……..……..……..……. .Nationalpark Bayerischer Wald
OTU ……..……..……..……...Operational Taxonomic Unit
PMI ……..……..……..……..….Postmortem interval
rDNA……..……..……..ribosomal Deoxyribonucleic Acid
ZSM ……...…....……..……..…... Zoologische Staatssammlung München

# List of publications and Declaration of contribution as a co-author

**Publication I:** "DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species"

Laura Hardulak performed a majority of the laboratory work (sample preparation, DNA extraction, PCRs, and NGS library preparation) for samples collected in 2016. For samples collected in 2018, Laura assisted Jerome Morinière and Dr. Axel Hausmann in supervising the laboratory work, designed the extraction method comparison tests and performed a major portion of them in the laboratory. Laura was responsible for implementing a customized bioinformatic pipeline, performed all statistical tests, generated all plots, and wrote a majority of the manuscript.

**Publication II:** "A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring"

Laura Hardulak performed the bioinformatic analysis of sequence data for the metabarcoded samples, implementing a customized pipeline for obtaining OTUs from raw NGS data. Laura ran the Automated Barcode Gap Discovery analysis on the generated OTUs, created the presence-absence diagrams, contributed to writing the metabarcoding and data analysis section of the paper, and provided English language editing.

**Publication III**: "DNA Barcoding in Forensic Entomology – Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Species"

Laura Hardulak participated in support and supervision of the field setup and laboratory work (preparing bulk samples for NGS). Laura performed the bioinformatic analysis of sequence data for the metabarcoded samples, implementing a customized pipeline for processing raw NGS sequence data into OTUs with molecular taxonomic identifications. Laura oversaw the subsequent analysis of the data, wrote a section of the paper on data analysis, and contributed edits to the final manuscript.

**Publication IV:** "High Throughput Sequencing as a novel quality control method for industrial yeast starter cultures"

Laura Hardulak collaborated with TUM researchers in designing the experiment, and assisted in supervision of the laboratory work to prepare yeast samples for NGS. Laura performed the bioinformatic analysis of the NGS sequence data, wrote the bioinformatics secion of the paper, and provided English language editing.

.

# Summary

## Abstract (English)

As a relatively new technology, DNA metabarcoding has already shown potential for a wide variety of practical applications. Biodiversity monitoring is a discipline of particular importance currently, as hundreds or thousands of species become extinct each year, and most extant species remain undescribed. Metabarcoding can greatly assist in increasing the speed and decreasing the cost of large-scale biodiversity monitoring campaigns, but development and improvement of techniques involved in the steps of a metabarcoding pipeline, from DNA extraction through taxonomic identification of sequence data, are still needed. Projects presented in this thesis cover a range of applications of DNA metabarcoding, from biodiversity monitoring of terrestrial invertebrates, to forensic entomology, reverse taxonomy, and the quality control of food, beverage, and novel food products. A multi-year biomonitoring survey with a special focus on early detection of invasive and/or pest species was conducted in the largest national park in Europe. Results demonstrate the effectiveness of metabarcoding for characterizing biodiversity patterns and phenologies, with Principal Component Analyses and ANOSIM tests showing a significant difference in BIN compositions between groups of samples taken from inside of versus outside of the park, for each of the two study years (2016 r = 0.2, p = 2e-04; 2018 r = 0.239, p = 1e-04). Results of the same study also provide support for employing multiple methods of DNA extraction from bulk samples (i.e. homogenizing the specimens themselves, and utilization of the preservative ethanol as a source of genetic material), as well as combining multiple reference sequence databases, in order to improve the chances of detecting species of interest. An attempt was made to counter the issue of specimen size bias, by pre-sorting specimens according to size, but was not successful for the smallest specimens. The invasive pest *Lymantria dispar* (Linnaeus, 1758) was detected in an ethanol-extracted sample, representing the first detection of this species in the Bavarian Forest National Park. In another project, a DNA barcode library was created, with records for 2,453 named species and 5,200 total BINs, whereby metabarcoding sequence clusters were able to be assigned to “dark” taxa, or taxa which have not yet been described, but are known only by BIN or MOTU, in a reverse taxonomic approach. For families containing “dark taxa”, an inverse correlation was discovered between body size and percentage of unnamed taxa (r = -0.41, p = 4e-04). A pilot study in DNA barcoding for forensic entomology resulted in the contribution of 120 high quality COI barcode sequences to the ZSM reference library, with 46 newly added species belonging to 11 orders. Metabarcoding facilitated the characterization of insect material collected on decomposing porcine corpses, with 469 species identified molecularly from HTS data. Metabarcoding of food and brewing yeasts was also performed. It was demonstrated that metabarcoding can be successfully applied as

a non-targeted approach to detecting differing species in supposedly pure yeast starter cultures, using the 26S rDNA D1/D2 region of chromosome XII in *Saccharomyces* spp. All of the work herein contributes to the growing knowledge bases of describing the earth's biodiversity, as well as, from a practical standpoint, the refinement of methods involved in the process of DNA metabarcoding for molecular taxon identification.


## Abstract (German)

Als relativ neue Technologie hat das DNA-Metabarcoding bereits Potenzial für eine Vielzahl praktischer Anwendungen gezeigt. Die Überwachung der biologischen Vielfalt ist derzeit eine Disziplin von besonderer Bedeutung, da jedes Jahr Hunderte oder Tausende von Arten aussterben und die meisten vorhandenen Arten unbeschrieben bleiben. DNA-Metabarcoding kann erheblich dazu beitragen, die Geschwindigkeit zu erhöhen und die Kosten für groß angelegte Kampagnen zur Überwachung der biologischen Vielfalt zu senken. Die Entwicklung und Verbesserung von Techniken, die an den Schritten einer Metabarcodingpipeline von der DNA-Extraktion bis zur taxonomischen Identifizierung von Sequenzdaten beteiligt sind, sind jedoch weiterhin erforderlich. Die in dieser Arbeit vorgestellten Projekte decken eine Reihe von Anwendungen der DNA-Metabarcoding Technologie ab, von der Überwachung der biologischen Vielfalt terrestrischer Wirbelloser über forensische Entomologie, umgekehrter Taxonomie bis hin zur Qualitätskontrolle von Lebensmitteln, Getränken und neuartigen Lebensmitteln. Im größten Nationalpark Europas wurde ein mehrjähriges Biomonitoring-Projekt mit besonderem Schwerpunkt auf der Früherkennung invasiver und / oder Schädlingsarten durchgeführt. Die Ergebnisse zeigen die Wirksamkeit des DNA-Metabarcodings für die Charakterisierung von Biodiversitätsmustern und -phänologien, wobei Hauptkomponentenanalysen und ANOSIM-Tests einen signifikanten Unterschied in der BIN-Zusammensetzung zwischen Gruppen von Proben zeigen, die innerhalb und außerhalb des Parks für jedes der beiden Studienjahre (2016) entnommen wurden ($r = 0,2$, $p = 2e-04$; 2018 $r = 0,239$, $p = 1e-04$). Die Ergebnisse derselben Studie unterstützen auch die Anwendung mehrerer Methoden zur DNA-Extraktion aus Massenproben (beziehungsweise Homogenisierung der Proben selbst und Verwendung des Konservierungsmittels Ethanol als Quelle für genetisches Material) sowie die Kombination mehrerer Referenzsequenzdatenbanken, um die Chancen zu verbessern, alle Arten zu entdecken. Es wurde versucht, dem Problem der Abweichung der Probengröße durch Vorsortieren der Proben nach Größe entgegenzuwirken, was jedoch bei den kleinsten Proben nicht erfolgreich war. Der invasive Schädling *Lymantria dispar* (Linnaeus, 1758) wurde in einer mit Ethanol extrahierten Probe nachgewiesen, was den ersten Nachweis dieser Art im Nationalpark Bayerischer Wald darstellt. In einem anderen Projekt wurde eine DNA-Barcode-Bibliothek mit Aufzeichnungen für 2.453 benannte Arten und insgesamt 5.200 BINs erstellt, wobei Metabarcoding-Sequenzcluster „dunklen" Taxa oder Taxa zugeordnet werden konnten, die noch nicht beschrieben wurden, aber nur

bekannt sind von BIN oder MOTU in einem reverse-taxonomy Ansatz. Für Familien mit „dunklen Taxa“ wurde eine inverse Korrelation zwischen Körpergröße und Prozentsatz unbenannter Taxa entdeckt (r = -0,41, p = 4e-04). Eine Pilotstudie zum DNA-Barcodierung für die forensische Entomologie ergab den Beitrag von 120 hochwertigen COI-Barcode-Sequenzen zur ZSM-Referenzbibliothek, wobei 46 neu hinzugefügte Arten zu 11 Ordnungen gehörten. Das Metabarcoding erleichterte die Charakterisierung von Insektenmaterial, das bei der Zersetzung von Schweinen gesammelt wurde, wobei 469 Arten molekular aus HTS-Daten identifiziert wurden. Metabarcoding von Lebensmitteln und Brauhefen wurde ebenfalls durchgeführt. Es wurde gezeigt, dass das Metabarcoding erfolgreich als nicht zielgerichteter Ansatz zum Nachweis unterschiedlicher Arten in vermeintlich reinen Hefestarterkulturen unter Verwendung der 26S-rDNA-D1 / D2-Region von Chromosom XII in *Saccharomyces* spp. angewendet werden kann. Alle hierin enthaltenen Arbeiten tragen zu den wachsenden Wissensgrundlagen zur Beschreibung der biologischen Vielfalt der Erde sowie zur praktischen Verfeinerung der Methoden bei, die am Prozess des DNA-Metabarcodings  zur Identifizierung molekularer Taxa beteiligt sind.

# General Introduction

## Biodiversity Monitoring

### Why is biodiversity important?

Biodiversity is all of the life on Earth, and our ability to monitor and conserve it is critical to the continuation of human life on this planet. Every species in every ecological system, whether terrestrial, marine, or freshwater, plays a role in the overall functioning of the ecosystem and its ability to provide services upon which our lives depend. Declines in biodiversity have recently been accelerated by human activity. In fact, it has been estimated that the current extinction rate is 1000 times that of the natural background extinction rate (Coleman, 2015). In light of these figures, it is difficult to argue against the viewpoint that we must be responsible for minimizing the irreversible damage that we have been causing it.

Ecosystems provide basic services to the planet, such as nutrient cycling, soil formation, and the stabilization of water resources. Biodiversity provides us with food and medicinal resources, and raw materials we need to create clothing, shelter, fuel, and the products we use every day. Furthermore, biodiversity enriches our lives through the vital importance of many animals, plants, and natural landscapes to culture, recreation, and tourism. Replacing these resources once lost

would be either impossible or extremely difficult. Therefore, efforts in biodiversity monitoring and conservation must continue to expand.

The most serious threats to biodiversity include climate change, habitat destruction, pollution, natural resource exhaustion, and invasive species. Furthermore, it is difficult to predict the future impacts the interactions of these factors will have (Bellard et al., 2012); (Mantyka-Pringle, Martin, and Rhodes 2013; Segan, Murray, and Watson, 2016); however hundreds, if not thousands, of species are currently becoming extinct each year (Chivian and Bernstein, 2008). Monitoring changes in species' numbers and distributions can be an overwhelming task, especially in light of the fact that only a fraction of all species that currently exist have been described, and estimates of how many there are, are still widely variable. Currently, approximately 1.5 million species have been described, and estimates of the total number on Earth range anywhere from 2 million (Mora et al., 2011) to 2 billion (Larsen et al., 2017). In a recent study, Larsen et al. (2017) suggest that 78% of all species on Earth are bacteria, and that there are roughly 2 billion total species on Earth. This may come as a surprise, as less than 1% of all described species are bacteria, and insects comprise the largest taxonomic group of described species, with many previous estimates agreeing on a figure of approximately 6.8 million total insect species in existence. However, recent advances in molecular methods of species identification based on DNA sequences provide evidence for different species boundaries of insects than was previously thought. Based on this new information, the researchers estimated the total count of insect species at approximately 40 million. Additionally, they estimated that each insect species likely hosts a unique species of mite, as well as of nematode, microsporidian, and apicomplexan protist; but most importantly, that each insect species likely hosts at least 10 bacterial species that are not found anywhere else. Therefore, a new projection of the estimated total number of species on Earth could be placed at about 2 billion.

Another reason why biodiversity is critically important to human life, is its role in mitigating the spread of pandemics. The increasing frequency of infectious disease outbreaks over the past few decades has been linked to biodiversity loss and climate change (Banu et al., 2014; Karvonen et al., 2010; Ostfeld 2009; Zell 2004). Over this time period, deforestation (due to all causes) has been steadily increasing (World Bank Data, 2016). Deforestation drives animals out of their natural ecological habitats and into closer proximity with human settlements, creating ideal conditions for vectors to breed and spread infectious diseases (Gottwalt, 2014). Some well-known examples of zoonoses include the bubonic plague, the West Nile virus, swine and avian influenza viruses, Lyme disease, malaria, dengue fever, and SARS-CoV-1 and -2. Deforestation changes the ecology of disease vectors and other medically important insects, and it has been predicted that rapid changes in disease transmission patterns would pose problems for public health services (Burkett-Cadena and Vittor 2018; Walsh et al., 1993). Zoonoses would be less common if animals were not brought en masse into contact with humans and ecosystems were allowed to recover.

A fundamental understanding of biodiversity and ability to make informed decisions toward its management is provided by taxonomy. Being that we cannot conserve what we do not know, the above-mentioned findings underscore the importance of biodiversity monitoring and the enormity of the proportions of this undertaking on a global scale. Furthermore, the importance of insects is highlighted for the key role each species plays in supporting biodiversity; and also, the role of DNA sequences in species identification is of critical importance. Molecular methods of biodiversity monitoring and taxonomy, therefore, should continue to be developed and pursued in the furtherance of conservation ecological goals.

## The past and future of biodiversity monitoring

Traditionally, biodiversity monitoring has relied upon the morphological identification and visual quantification of the organisms present in a sample ecosystem. The limitations of this approach are underscored by the recent knowledge of how low a proportion of all species which are known is likely to be. Not even all species which are described can be morphologically identified to the species level. In any case, there are many sources of error when measuring biodiversity, as it is a concept and not a simple quantity that can be measured with complete objectivity, such as physical or chemical quantities like temperature, air pressure, or pH, for example. Estimating the biodiversity of a given environment requires adherence to sampling protocols, but a consensus on the standardization of these methods in the scientific community is lacking (Archaux, 2011).

Generally, a subset of species of interest, such as keystone species of a given ecosystem (if they are known) is used as a proxy for quantifying the total species abundance. This is, of course, not completely accurate, and is biased towards species which are already not only described but are known to exist in the area under study. The possibilities of taxa becoming threatened and going extinct without our knowledge are now more evident than ever before. When targeted species lists for monitoring are expanded, this increases the accuracy of the results, but also increases the amounts of time and funds necessary.

Two factors which introduce bias into conventional biodiversity monitoring efforts are the taxonomic bias in biodiversity data and societal preferences (Troudet et al., 2017). A relatively small number of species attract most of the public, governmental, and scientific attention, while a vast majority remain unstudied and unknown. Namely, plants and vertebrates, are overrepresented in biology. This can be harmful because small, rare, or unappealing organisms do play key roles in ecosystem functioning; and it has been shown that focusing on only a few species inhibits reaching a global consensus and the development of efficient plans (Troudet et al., 2017). For example, the global bias in interest in mammals and birds at risk of extinction undermines efforts to secure funding for thousands of other threatened species, which continue to go overlooked (Davies et al., 2018).

Furthermore, the use of indicator species to monitor environmental changes and assess the efficacy of management, has been criticized, mostly with regard to the methods of choosing which species to use as indicators, and studies have shown that indicator species selection and identification of the relationship between these indicators and their specific applications remains challenging (Siddig et al., 2016). In order to collect large volumes of data, citizen science is sometimes employed for this form of biodiversity monitoring. However, it is subject to spatial bias related to human infrastructure and population density (Geldmann et al., 2016).

### Major approaches to species identification

Another issue relevant to biodiversity monitoring is related to taxonomy and the ability to identify species. Classifying species based on their morphological characteristics has been practiced for over 250 years, having its roots in comparative anatomy. Species are defined based on their phenotypic traits which characterize them as distinct from one another. Morphological taxonomy forms the basis for all hypotheses of phylogenetic relationships forming the tree of life. As most species that have ever existed are now extinct, in most cases, extinct organisms can only be characterized by their fossil records, limiting the extent to which they can be characterized. Morphological characterization, furthermore, is most effective on well-preserved adult male specimens. Even then, there are cases in which two or more species are visually indistinguishable from each other, or sexual dimorphism or other visual variations within species lead to incorrect multiple identifications. Some of these limitations can be overcome with molecular methods, however. As the relatively new field of molecular methods of species identification has become increasingly sophisticated, there has been somewhat of a divide among taxonomists regarding whether the morphological or molecular should be the most universally accepted and highly regarded method of identifying species. An overall consensus in the global scientific community on a standard method of species identification would be adventitious, as it would aid and streamline research collaboration and information exchange and transmission.

Molecular methods lend a degree of objectivity to species identification, which, when based only on morphology, is subject to disagreement among experts on how to distinguish particular traits and species from one another and the terminology to use. Disparity exists not only with respect to anatomical features, but also with the taxonomic levels of assignment. The concept of what constitutes a species continues to evolve as new information is incorporated. Several definitions of the term “species” exist, corresponding to different biological disciplines. The most widely accepted of which is the biological species concept: *“Species are groups of actually or potentially interbreeding populations that are reproductively isolated from other groups”* (Mayr, 1942). Whether individual specimens are reproductively compatible with one another cannot always be determined through morphological analysis, but can be via molecular systematics (Duellman and Venegas,

2005). This is particularly true in the case of cryptic species, or multiple species which appear identical to the naked eye. This phenomenon has been observed in most insect orders. There are also populations within single species which have been observed not to interbreed, for which molecular verification would also be helpful. Other cases where the morphological approach may fail are when only immature life stages, such as larvae or eggs, only parts or remains of specimens are available. In these cases, molecular methods are clearly helpful, with DNA being the biological molecule with the greatest record of success for identifying species. Overall, both morphological and molecular methods have advantages and disadvantages, and the complementary use of both has proven successful (e.g. Best et al., 1986; Duellman and Venegas, 2005; Miyamoto 1981; Shaklee and Tamaru 1981). Being that it is important to ensure that biodiversity is as representatively sampled as possible, the need for complementary approaches to biodiversity monitoring is growing.

## Taxonomic Impediment

Another major issue impacting our ability to inventory--and thus conserve--all of the life on earth is that of taxonomic impediment. Encompassed by the term “taxonomic impediment” are several problems:  the lack of completion in our knowledge of the global biodiversity, the insufficient numbers of expert taxonomists and their uneven distribution throughout the tree of life around the world, and the deficiency in global taxonomic infrastructure. It is compounded by a negative outlook in light of the increases in extinction rates and in the estimated number of species which are still unknown or undescribed. The pressure to advance progress towards the ultimate goal of describing all species on the planet is increasing, as the numbers of taxonomists are decreasing. According to a poll of taxonomists by Mora et al. (2011, cited in Coleman, 2015), 79% of respondents believed the number of professional taxonomists in their respective fields to be decreasing.

There may be exceptions to this trend, however, in which taxonomists are growing in numbers, particularly in Latin America. While the numbers of doctorates awarded in botany and zoology have been decreasing in the United States of America (de Carvalho et al., 2005), for example, the emphasis of Brazil on training undergraduate students of biology in cladistics has contributed to its higher numbers of systematic ichthyologists, entomologists, and botanists than most countries. Some of the countries with greater interest among the younger generations in systematic biology, are also rich areas of biodiversity. Unfortunately, however, the lack of employment opportunities in developing countries causes some graduates to seek employment in developed countries, but the United States and other countries with great overall opportunities for employment have been de-emphasizing organismic biology. As financial interests shift away from systematic biology, collections suffer from losses of funding, which could also be used to hire new taxonomists.

The dearth of taxonomists and the lack of political initiative to prioritize taxonomic and conservation biology goals can therefore be seen as a vicious cycle. To combat this, the first Convention on Biological Diversity was opened for signature at the Earth Summit in Rio de Janeiro in 1992, emphasizing the issue of taxonomic impediment, as the large gaps in the cumulative taxonomic knowledge, deficiencies in taxonomic infrastructure, and declining numbers of taxonomists. Since then, researchers in systematic biology have emphasized the necessity of taxonomic research for the urgent conservation of biodiversity, as the latter has gained in public awareness lately, as biodiversity loss has accelerated even more.

For example, as Buckley (2015) articulates, the Earth supports over 7 billion people, and the numbers of some other species are up to 9 orders of magnitude less. Although humans can make individual choices, such as driving and flying less, consuming and wasting less, and even having fewer children, this goes against our evolved instinct to expand the species, as well as the financial interests of many businesses and corporations. Therefore, the only way for us to willingly bring about changes necessary to save biodiversity is through “social machines” using politics, government, finance, and communications. In order for governmental agencies to take action, they must be advised by scientific research communicated clearly. As Mace (2004) points out, species lists designed for conservation planning are often used to determine where to focus conservation actions, but a new collaboration between conservationists and taxonomists is needed in order to overcome the barriers to conservation caused by the shortage of taxonomic information and skills and the confusion over the delimitation of what a “species” is.

Taxonomical misinformation can be a serious problem when it underpins decisions made by governmental and intergovernmental organizations. In a controversial commentary, Garnett and Christidis (2017) argue that conservation is being hampered by too much freedom for taxonomists to define species at their own discretion. They mention the inconsistencies between how scientists studying different classes, e.g. mammals and birds, delineate one species from another, resulting in different approaches to the splitting and lumping of groups and hence inconsistencies in figures reported. Splitting species into smaller units, for example, results in more species being designated as threatened, potentially misinforming decisions for investment and land use. Misinforming the public can further harm biodiversity; according to Garnett and Christidis (2017), increased splitting of certain vertebrate taxa could encourage trophy hunters to target more animals in order to have a representative of every perceived species.

The opposing perspective of what should be done to control the spread of taxonomical misinformation is expressed by Thomson et al. (2018). They argue that an increase in legislative restrictions on the publication of research will be harmful to biodiversity because it impedes taxonomic research, and taxonomic research is needed in order to achieve conservation goals. They assert that the position taken by Garnett and Christidis (2017) is based on a fundamental

misunderstanding of the scientific basis of taxonomy, formalized nomenclature, and the relationship between them; and that their assertion that an "assumption that species are fixed entities underpins every international agreement on biodiversity conservation" demonstrates their failure to understand taxonomy and the ever-changing view of what constitutes a species. The overly narrow understanding of taxonomy is perceived by Thomson et al., (2018) to be a trend, which may be due in part to the decrease in emphasis on the teaching of taxonomy and nomenclature at universities, as research priorities shift away from these systematic fields.

For all these reasons, it is vital not only for taxonomic research to be continued without undue impediment, but also for scientific rigor to be applied, in order to facilitate the flow of accurate information between researchers as well as to policy makers. Though the best way to achieve it is debatable, it would be reasonable to conclude that a greater degree of objectivity would benefit taxonomy and conservation efforts. Molecular taxonomic methods, such as DNA barcoding, may be of use here.

## DNA Barcoding

Starting in its early development, DNA Barcoding was designed with the objective to accelerate the rate of species discovery, in order to combat the problem of taxonomic impediment in striving towards the ultimate goal of inventorying all of the life on Earth. Utilizing standardized regions of genes as species tags, DNA barcoding was introduced with the publication “Biological identifications through DNA barcodes” (Hebert et al., 2003). These researchers at the University of Guelph, Ontario, Canada proposed a segment of the mitochondrial gene cytochrome c oxidase I (COI), amplified by primers created by Folmer et al. (1994), to serve as the core of a global database of reference sequences belonging to correctly identified specimens, to which sequences from unknown specimens from around the world can be compared. These sequences thus serve as a “barcode” indication of the species, in much the same way as a Universal Product Code identifies product from among millions of others.

The authors promoted DNA barcoding for its ability to overcome many of the limitations of traditional taxonomy, such as cryptic diversity, phenotypic plasticity, whole specimen availability, and the fact that a team of many thousands of traditional taxonomists would be required in perpetuity to identify all species estimated to exist. Their methods to demonstrate the ability of DNA barcoding were as follows. First, 655 COI sequences were obtained (some from GenBank; others from their own laboratory extractions). Next, three COI profiles were created: one for the seven dominant animal phyla, one for eight of the most diverse insect orders, and another for 200 closely related species of lepidopterans. The profiles served to provide an overview of COI diversity within each taxonomic group, and then as the basis for identifying “unknown” sequences to the phylum, order, or

species level based on their congruences to species included in profiles. The insect COI profile was based on a single representative from each of 100 different families, and it was used to correctly assign 50 new samples to their correct orders. Then, profiles were tested for their ability to correctly place additional test ("unknown") sequences. The profile consisting of 200 closely allied species of lepidopterans was used to test species-level identifications. Neighbor-joining (NJ) analysis was used for both analysis of relationships of taxa within the profiles, as well as for assigning test taxa.

Results showed Hebert et al. (2003)'s COI phylum profile to have good resolution of the major taxonomic groups, recovering monophyletic assemblages for three phyla, and the chordate lineage forming a cohesive group. The order profile also showed a high rate of cohesion, with seven of the eight orders forming monophyletic assemblages. The profile consisting of 200 species of lepidopterans showed distinct clustering into each of the three superfamilies on a MDS plot, with further evidence of clustering of related species shown by NJ analysis. Regarding the "test" sequences, 53 of the 55 were assigned to the correct phylum on the phylum profile; the ordinal profile could correctly assign all 50 insect sequences to their correct orders; and at the species level, all 150 sequences were correctly assigned by the lepidopteran species profile.

### Why COI?

DNA encodes all of the heritable biological information of an organism, not only recognizable phenotypic traits, and it can also provide evidence of evolutionary relationships between taxa. Only very small amounts of biological material are required in order to extract DNA and subsequently sequence an organism's genome. While each organism's own genome is unique, examining differences in the nucleotide or amino acid sequences between specimens can provide clues to their degrees of evolutionary relatedness. DNA is a relatively stable molecule and is significantly easier to work with in the laboratory before it degrades, compared to RNA. In many cases, traces of DNA left in the environments with which organisms had come into contact can be utilized for sequencing, and also DNA can be sampled from dead organisms, especially if they have been preserved.

In order for a genetic marker to serve as a DNA barcode, it must be short enough to be sequenced in a single reaction, yet have sufficient between-species variation as well as a conserved region for the binding of universal primers (Ferri et al., 2009; Savolainen et al., 2005). Mitochondrial DNA is often preferred because the genome is haploid and small, it lacks introns, and has had limited exposure to genetic recombination. Mitochondrial DNA evolves rapidly, however, and has therefore been useful in phylogenetic population studies (Ferri et al., 2009; Hartl et al., 1997). Additionally, robust primers enable specific sections of the mitochondrial genome to be recovered (Folmer et al., 1994). Previous studies utilizing mitochondrial genes encoding ribosomal DNA (e.g. 12S, 16S subunits) have demonstrated that the presence of insertions and deletions in these genes complicates the sequence alignment process (because they would shift the reading frame), making

them of limited use for broad taxonomic analyses (Doyle and Gaut, 2000). Conversely, the lack of indels of the 13 protein-coding genes of the mitochondrial genome makes them better suited for barcoding.

As a mitochondrial gene, COI is thought to exist in every animal, in every cell. Hebert et al. (2003) explain that two important advantages make it stand out. The first is the existence of very robust universal primers for this gene, which allow recovery of its 5' end from most, if not all, animal phyla. The second advantage is its apparent possession of the greatest range of phylogenetic signal. This is attributed to its high rate of evolution, seen in the high rate of base substitutions of its third-position nucleotides--approximately three times greater than 12s or 16s rDNA (Knowlton and Weigt, 1998). A phylogenetic signal of such range allows distinguishing of closely related species, and in some cases, even geographical population groups within single species (Cox and Hebert, 2001; Trontelj, Machino, and Sket, 2005).

With DNA having 4 possible variants at each position (A, G, C, or T), the sequencing of a stretch of only 15 nucleotide positions makes possible 1 billion different sequences; but, as Hebert et al. (2003) explain, the biological reality is that functional constraints hold some positions constant across taxa, while other positions exhibit intraspecific--not just interspecific--variation. Utilizing a protein-coding gene reduces the impact of functional constraints on variation, as four-fold degeneracy at the third nucleotide position of the codons makes them only weakly constrained by selection. Fortunately, it is just as easy to obtain DNA fragments hundreds of bp long, as it is for 45 bp (the length needed to create 1 billion possible unique barcode labels with four-fold degeneracy at the third position of each codon). This is influenced by two other biological phenomena: A-T or G-C bias, and the fact that there is less variation at most nucleotide positions in closely related species than would be expected by random chance. At a modest rate of sequence evolution, 12 diagnostic nucleotide differences can be expected in a 600 bp stretch when comparing species which have been reproductively isolated from each other for a million years.

### BOLD, OTUs, and BINs

The largest global database of reference sequences for DNA barcoding is the Barcode of Life Data System (BOLD, Ratnasingham and Hebert, 2007). Launched in 2005, BOLD is publicly available over the internet (http://v4.boldsystems.org/), as a workbench and repository supporting a growing community of researchers. As DNA barcoding has expanded, the amount of sequence and specimen data on BOLD had increased vastly. As of February 2020, it contains over 4.6 million sequences representing over 500 000 species. Most of the sequences are COI-5P. Users may contribute their own data, as long as specific criteria are met.

New versions of BOLD have been developed over the years as the sequence data has grown. The large amounts of data present bioinformatic challenges, and BOLD version 4 features an

Advanced Programming Interface (API) which enables researchers to utilize this data to test hypotheses and construct models on a larger scale than was previously possible.

Operational Taxonomic Units (OTUs) act as proxies for species not yet described. As the vast majority of life on Earth is undescribed, OTUs serve a pragmatic interim purpose, facilitating information exchange between researchers all over the world. OTUs are computationally derived clusters of sequences which show similar patterns among themselves within a cluster, but dissimilar patterns of the sequences belonging to other OTU clusters (Bhat et al., 2017). OTU clustering arose as a means to manage the “big data” proportions of sequences generated by microbial metagenomics, and today it is also used in metabarcoding of eukaryotic organisms. Such clusters generated from molecular data from NGS are known as MOTUs (molecular Operational Taxonomic Units).

Where to set the divergence threshold when performing clustering, in order to yield MOTUs which most closely resemble the barcodes of actual species, depends on the taxa. In the animal kingdom, studies have shown that COI sequence divergences within named species rarely exceed 2%, and that over 95% of species tested possessed a diagnostic COI sequence array (Ratnasingham and Hebert, 2013). However, it is important to keep in mind that no gene is “the speciation gene,” and care must be taken with respect to the “barcoding gap” when attempting to molecularly define OTUs.

An advantage of experimentation in macrofauna is that a great deal of morphological data is accessible, so that individual MOTU species hypotheses may be tested against their morpho- or biological species. Blaxter et al. (2005), who performed extensive macrofaunal surveys utilizing multiple markers, showed MOTU and breeding-based biological species hypotheses to be congruent, while morphologically based analyses had internal disagreement. Microfauna (animals of body sizes smaller than ~ 1 mm), on the other hand, are not very visually informative, and therefore have similar limitations to microbes, often only having MOTUs and not any other species hypotheses against which to validate them. For these and other poorly studied fauna, additional steps must be taken to optimize MOTU concordance with species.

In order to provide a system of standardized protocols for the delineation of animal OTUs, and to develop a registration program which enables the comparison of results from all over the world, the Barcode Index Number (BIN) system was introduced (Ratnasingham and Hebert 2013). The BIN system is a dynamic system based on state-of-the-art algorithms. New sequences are assigned to OTUs or signified as founders. As more sequences are continually added to BOLD, the BIN clustering algorithm is re-run, resulting in BINs with increasing concordance with species.

One must bear in mind, however, that BINs are not species. But even so, there is pragmatic value in using taxa defined by their barcodes in the ongoing effort of cataloging life. This view was already expressed by Blaxter et al. (2005), who regarded observing specimens’ sequences

clustering closely together as a clue for closer examination, rather than something to be taken as a foregone conclusion. This idea agreed with that of the inventors of DNA barcoding . . . although not everyone in the taxonomic community was of a similar mind.

### Criticisms of DNA Barcoding

As is often the case with new technologies which promise to change established traditions, DNA barcoding was met with a degree of opposition and criticism. Some taxonomists became concerned that the new technology would endanger the 250-year-old field as the world had known it. Overall, the main cause of concern appears to have its foundations in the viewpoint that barcoding was intended to--or could perhaps have the unintended consequence of--supplanting traditional taxonomy. This belief grew in part from the presumption that barcoding would compete with taxonomy for funding, giving rise to accusations that it was “anti-taxonomy” and would ultimately impede our understanding of biodiversity, by accelerating the decline of the well-established and much-needed discipline, leaving humanity with an inadequate replacement in its stead. This does appear to be the foundation for the critical publications “DNA barcoding is no substitute for taxonomy” (Ebach and Holdrege, 2005) and others. These early objections were addressed by Hebert and Gregory (2005), who asserted that the concerns and criticisms stemmed from fundamental misconceptions about the aims of the DNA barcoding effort.

Some researchers took in-depth oppositional viewpoints of the science of DNA barcoding and how it cannot meet what they perceived to be its goals if it were a comprehensive taxonomic and phylogenetic system. Will and Rubinoff (2004) began by admitting that morphological taxonomy is plagued by just as many complications as molecular taxonomy, but asserted that morphological taxonomists have a whole suite of anatomical characteristics at their disposal in order to overcome these difficulties, whereas someone trained only in DNA barcoding has only part of a single gene. This was based on the (unfounded) presumption that DNA barcoding would be performed by people lacking in taxonomic expertise. They go on to enumerate points they see as illustrating that barcoding falls short of its goals. The first being that it fails to recover accurate phylogenetic species trees. DNA barcoding, they assert, has an inherent inability to accurately place unknown sequences which are not exact matches to the database in cladograms, and internal attachment points are ambiguous or incorrect. They illustrate their point by highlighting noninformative internal branches of the tree diagrams from Hebert et al. (2003). Even though recovering phylogenetic relationships has never been a goal of barcoding, Will & Rubinoff (2004) appear to argue from the standpoint that phylogeny is indeed the goal, and dismiss Hebert et al. (2003)’s use of the word “profile”, rather than “phylogeny”, as an attempt to dodge criticism. This controversy was addressed by Hebert et al.

(2005), reiterating that DNA barcodes seek instead to identify specimens to known taxa and aid in the discovery of new ones.

A success in this aspect was illustrated with a new neighbor-joining analysis of Kimura 2 parameter (K2P) distances for barcode sequences of two genera of moths. Even though it did not recover taxonomic information below the suprageneric level, Hebert et al. (2005) write, it highlights taxonomic assignments in need of scrutiny. Namely, the close placement of *Simyra henrici* (Grote, 1873) to certain species of *Acronicta* revealed a previously unknown close relationship, which could be further supported by larval morphology, ecological niche, and other characteristics. Thus, DNA barcoding was able to refine existing hypotheses on taxonomic relationships, as well as provide new insights into evolutionary patterns.

Will & Rubinoff (2004)'s objection to DNA barcoding, and Hebert et al (2005)'s response illustrate that, contrary to the fear that barcoding was attempting to take over or replace traditional taxonomy--and would be inadequate in doing so--its creators never intended that to be the goal. Instead, it has always been intended as another tool to add to the existing taxonomic toolbox, build alliances between morphological and molecular taxonomists, and make the traditional Linnaean system more accessible to anyone interested in exploring biodiversity, from professionals to schoolchildren. Currently, this is evidenced by the tremendous amounts of not only sequence, but detailed specimen data on BOLD.

Early fears and misunderstandings notwithstanding, DNA barcoding has continued to be widely applied for identifying unknown biological specimens and aiding species discovery through species delimitation. A decade after its premier, an opinion piece was published titled "The seven deadly sins of DNA barcoding" (Collins and Cruickshank, 2013). Rather than being an opposition to barcoding, however, the authors expressed a positive view of its broad benefits to research as well as regulatory science, and aimed to suggest potential solutions to the main scientific issues they had observed related to shortcomings in the experimental designs of studies utilizing it.

These issues stem in large part, Collins & Cruickshank (2013) believe, from confusion between specimen identification and species discovery and the effect this has had on the ways in which hypotheses have been formed and tested. They see both of these aims as uncontroversial, however, as long as they are properly defined. But failure to test clear hypotheses constitutes the first "sin" of barcoding and may result in some of the other "sins", such as inappropriate use of neighbor joining trees, of bootstrap resampling, and of fixed-distance thresholds. Although Hebert (2005) referred to DNA barcoding as being "discovery-driven" as opposed to "hypothesis-driven", it is possible to construct hypotheses such as the identification success of a reference library via barcoding, and then test it by simulating a quantified identification scenario.

Another barcoding "sin" which contributes to problems downstream is inadequate a priori identification of specimens. The quality of the reference library is, of course, of pivotal importance to

any barcoding activity, and voucher specimens must not have been incorrectly identified if their sequences are to be utilized for comparison with sequences from unknown samples. On a positive note, the first years and decades of DNA barcoding have demonstrated its utility, and the problems that have arisen throughout its practical application are--at least theoretically--possible to overcome.

## DNA Metabarcoding

The advent of High Throughput Sequencing (HTS) (formerly Next Generation Sequencing, also known as second-generation sequencing) technology made possible a new branch of DNA barcoding: DNA metabarcoding. Utilizing parallelization of the sequencing of genetic material, HTS instruments can produce millions of sequences in each run (Grada and Weinbrecht, 2013). The availability of HTS, beginning in 2006, decreased the cost per base sequenced by orders of magnitude, compared to standard dye terminator methods. This precipitated a rapid expansion in the areas of biology involving the analysis of genetic material. Genomics and the related "-omics" disciplines grew explosively, with the time and cost restrictions of sequencing even large genomes, largely alleviated. This has revolutionized biology, giving it a part in the emerging scientific fields dealing with the analysis of "big data", made possible by contemporary advances in the technology of computing. Advancements in various fields related to biology, such medicine, food quality control, international trade control, ecology, forensic biology, and more have come to fruition through applications of HTS (Grada and Weinbrecht, 2013).

Since its inception, HTS has been widely applied to the study of the genomes of bacteria as well as eukaryotes. In metagenomics, samples containing unknown mixtures of species are sequenced. Initially, most of the research focused on bacteria, for which morphological analysis is of limited use. For the first time, single species did not have to be cultured individually before being sequenced; a sample of "bacterial soup" taken from an environmental source--whether the environment was within a eukaryotic organism (e.g. its gut or skin microbiome), feces, a body of water, a facility with cleanliness standards such as a hospital, restaurant, or manufacturing facility for food or pharmaceutical products, or anything else--could be taken, the DNA extracted holistically, amplified, and sequenced. This provided vast new potential to study genetics in ways that had been previously intractable, such as environmental DNA (eDNA), for example. While genomics is more difficult and expensive to apply to eukaryotes due to their much bigger genomes, the advantages of HTS can still be readily applied to Eukaryota, thanks to DNA barcoding. Samples containing DNA from any number of unknown specimens can be holistically extracted, their barcode regions amplified, sequenced, and the output digitally analyzed for taxonomic composition, through the process of metabarcoding.

Metabarcoding utilizes samples of DNA from more than one organism, extracted either from environmental sources where organisms have left traces of their genetic material (eDNA *sensu stricto*), or directly, from bulk samples of captured invertebrate animals (also known as community DNA). In the latter case, the most established and generally reliable method for extracting DNA from mixed samples involves first grinding the specimens into a homogenous powder, as first demonstrated successfully by Hajibabaei et al. (2011). The disadvantage, however, is that the specimens are destroyed, preventing them from being available for subsequent morphological analyses. In the plant kingdom, on the other hand, metabarcoding can in many cases be applied to samples of pollen. A practical advantage of DNA barcodes is that they are short (at least when we limit the discussion to classical COI barcoding of Animalia). The short length is a fortunate advantage for metabarcoding, as the most widely employed HTS technologies currently produce many more but much shorter reads, compared to Sanger sequencing. This means that even though a sample may contain DNA from thousands of unidentified specimens, the produced sequences comprise strategic portions of the DNA barcode region, allowing their comparison to a reference library by algorithms such as BLAST. Matches to the database elucidate the likely taxonomic composition of the sample, with the highest-scoring matches providing the highest levels of taxonomic resolution.

## Applications of DNA barcoding and metabarcoding

### *Biodiversity monitoring*

The ability of DNA metabarcoding to vastly expand the scale of DNA barcoding is of paramount importance to biodiversity monitoring. In light of the urgency created by a rapidly and unpredictably changing planet, scientists' ability to rapidly and comprehensively assess ecosystem health by monitoring biodiversity changes and predicting their trajectories is critical. Fortunately, metabarcoding has also been remarkably successful here. In its first few years, metabarcoding was shown to recover significant portions of existing biodiversity (Aylagas, Borja, and Rodríguez-Ezpeleta, 2014; Yu et al., 2012) and to reveal unknown patterns of biodiversity (Leray and Knowlton, 2015), and it has been successfully applied to large-scale biodiversity assessments (e.g. Elbrecht et al., 2017; Epp et al., 2012; Hardulak et al., 2020 (see appendix); Hausmann et al., 2020; Ji et al., 2013; Morinière et al., 2016; Shokralla et al., 2012; Taberlet et al., 2012). Notably, Elbrecht et al. (2017) identified more than twice the number of taxa in freshwater invertebrate samples with metabarcoding, compared with morphology, and that at a higher taxonomic resolution.

Improving protocols for analyzing the biodiversity present in mixed samples is a current research objective, being that many undescribed small invertebrates are present in samples at low abundances and tend to be overlooked. Towards the aim of overcoming this problem, the "$1 DNA barcodes" HTS protocol was designed (Meier et al., 2016). These authors demonstrated a

generation of mini-barcode molecular markers for over 1000 species of midges from a mixed sample in which more than half of the putative species were rare, and some were represented by immature life stages. They were able to lower the chemical costs to less than 0.40 USD per specimen, through elimination of the DNA extraction step in a process known as “directPCR” (Wong et al., 2014), and being one of the earliest metabarcoding studies to use Illumina MiSeq technology for HTS, which is significantly less expensive than pyrosequencing.

Due to the increased urgency of the biodiversity crisis and taxonomic impediment, it has been suggested that the field of biomonitoring needs to evolve into “Biomonitoring 2.0” (Baird and Hajibabaei, 2012). In the opinion piece, “Biomonitoring 1.0” is the current state of the art, limited to a “binary outcome” of taxa being either impacted or not impacted when compared with control environments lacking environmental stressors, and it is not sufficient because it lacks diagnostic approaches which aim to clarify specific causal agents within complex stressor scenarios. These authors assert HTS-based techniques for taxonomic identification (such as metabarcoding) will play a key role in making this paradigm shift, on the basis that such a shift has thus far been precluded by the limits of how much taxonomic information is yielded by processing biological samples with traditional, morphologically-based methods. The limitedness of biodiversity characterizations of samples from study environments, such as river samples consisting of mostly insect larvae, impedes the identification and isolation of specific stressor responses against a background of multiple, interacting stressors. Continuing with the example of a river ecosystem, “Biodiversity 1.0” is unable to establish a causal relationship between a pesticide and changes in biodiversity in consideration of factors which co-vary with the intensity of agricultural land use, such as nutrient emissions, runoff, and increased temperature related to deforestation. In order to establish stressor-response relationships, Baird and Hajibabaei (2012) write, increased “resolvable information content” is necessary.

*Quality control in the food and brewing industries*

Early accomplishments of metabarcoding have been seen in its applications in diverse areas such as ecological monitoring, quality control of food products, international trade disputes, diet analysis through eDNA from feces, and plant-pollinator interactions. Several cases where DNA barcoding or metabarcoding discovered ingredients other than those stated by the manufacturers in consumable goods for human consumption, have gained widespread attention in recent years. Utilizing multiple marker genes of plants, teas suspected of ingredient substitution were intercepted at the border and were confirmed to contain, either completely or partially, plants other than barley (*Hordeum vulgare*) (Jian et al., 2014). Metabarcoding has likewise shown promise towards the advancement of standard procedures for the quality and purity testing of herbal supplements (Raclariu et al., 2018). Food fraud has also been a popular subject in recent years. In a highly

publicized scandal, horse (*Equus caballus*) meat was discovered in frozen lasagna, fueling consumer concern about the integrity of meat products, especially beef, throughout Europe (Iwobi et al., 2017). DNA barcoding has since facilitated the investigations of food fraud, notably seafood mislabeling (Willette et al., 2017). The power of DNA barcoding to determine which species are present in food products is taken to the next level by metabarcoding, in that samples of processed foods, beverages, or supplements can be processed holistically, and ideally, every species present can be detected, even in trace amounts (i.e., contamination).

Recently, the consumption of insects by humans as food, an ancient practice which has largely been abandoned in the Western world for centuries, has been gaining popularity. In light of ever-growing concerns about a changing planet with dwindling resources, insects have been hailed as the food of the future. They provide a comparable amount of protein to that of vertebrate livestock, but are much more sustainable to produce, with a much smaller ecological footprint (DeFoliart, 1992). It has even been claimed that entomophagy is more environmentally friendly, and even overall less harmful to sentient beings, than veganism, because the former requires less crop cultivation than eating plants directly (Fischer, 2016). Over the past decades, the UN Food and Agricultural Organization has brought together researchers, practitioners, and industrial representatives from around the world to international meetings to discuss the benefits of using insects as feed for livestock, as well as of humans consuming insects themselves, in order to combat existing world hunger and malnutrition in a growing population (Stamer, 2015). While approximately 70% of the world's population does consume insects as a regular part of the diet, many Westerners react disgustedly to the idea and do not feel willing to try it. Nevertheless, interest in the potential benefits of utilizing insects as a sustainable substitute for vertebrates for feeding animals and humans has been gaining traction in the west (Shockley and Dossey, 2014). Barcoding may soon play an important role in the quality control of such novel food products as well.

### *Forensic entomology*

Another novel application of DNA barcoding and metabarcoding is in the field of forensic entomology. A primary objective of this discipline is to estimate the time of death of deceased individuals by examining the insect colonizations of the corpses. The start of arthropod colonization is presumed to correspond approximately to the time of death (assuming that colonization is not impeded). This is useful because medical techniques in pathology which are used to establish the time of death are only applicable within approximately the first 72 hours postmortem, before subsequent decomposition leads to the disappearance of crucial cues needed for such analyses. Forensic entomology has traditionally relied on morphological analysis of the insect specimens. This tends to be time-consuming, however, being that a majority of the arthropod biomass discovered on decomposing corpses consists of larvae, eggs, and material not visually recognizable. In the case of

larvae, these must be collected from the crime scene and subsequently raised in the laboratory to the adult stage, from where they can be identified. Molecular methods have hence been increasingly applied to forensic investigations, since their debut (see Sperling, Anderson, and Hickey, 1994). Clearly, DNA barcoding can benefit these efforts, since all life stages as well as traces of organisms can be identified to high taxonomic levels. The main limitation, however, is the comprehensiveness of the reference library. A significant augmentation of the publicly available reference sequences for forensically relevant arthropod species was achieved recently by Chimeno et al. (2018; see appendix).

*Dark taxa*

It bears repeating that the taxonomic impediment is all the more urgent an issue to address in light of current declines in biodiversity. Not only do insects comprise at least 50% of all animal species, but many insect taxa are among the least well-studied groups of animals, due to them being less interesting to humans. In particular, taxa which are small in size and lacking in coloration attract the least attention of naturalists. Many such taxa are therefore “dark taxa”, as they were completely unknown before the advent of molecular identification methods such as DNA barcoding, and currently are known only by their BINs, or even only as OTU clusters of barcode sequences, awaiting BIN and species assignments.

Recent declines in insects--especially pollinator species--have made these invertebrates a subject of intense public concern (Hausmann et al., 2020). Studies employing mass collection methods have suggested a general decline in flying insect populations, with major losses having occurred over the last two decades (Habel et al., 2016; Hallmann et al., 2017; Sorg et al., 2013), or even within a few years (Lister and Garcia, 2018). Moreover, focused studies on economically important groups have linked declines in wild bees to pesticide contamination, habitat fragmentation and degradation (Potts et al., 2010; Vanbergen et al., 2013). More detailed knowledge is lacking, however, on the severity of the impact across all insect groups; and this failure to track the status of individual lineages reflects the fact that our knowledge of most insect species is still limited (Brix et al., 2015; Cruaud et al., 2017; Pante, Schoelinck, and Puillandre, 2015; Riedel et al., 2013; Wheeler, 2004).

The gap in taxonomic knowledge is particularly serious for the two hyperdiverse insect orders, Diptera and Hymenoptera (Geiger et al., 2016; Klausnitzer, 2006). These two orders are thought to comprise over half of insect alpha diversity (Völkl, 2004). It is likely that the true diversity of these two groups is greatly underestimated, and this belief is supported by the extraordinarily high numbers of DNA barcode clusters retrieved by metabarcoding insect collections at single monitoring sites in Germany and elsewhere. Only about 1000 of the roughly one million undescribed species of Diptera are described each year (Santos, Sampronha, and Santos, 2017), making Diptera one of the

“darkest” animal orders. For these reasons, a DNA barcoding campaign on German Diptera was carried out by ZSM researchers and other collaborators (Morinière et al., 2019, see appendix), aiming to address these gaps.

*Invasive Species*

One important area of focus involved in large-scale biodiversity monitoring on a changing planet is that of pest and invasive species monitoring. Invasive species pose a major threat to the conservation of biodiversity and to human economic activity and general livelihood. While introductions of neozoan flora and fauna have been facilitated by human activities for centuries, they have intensified with the globalization of trade and passenger travel. An estimated 1% of all species introduced to a new geographical area become invasive with serious economic impacts (Meyerson and Reaser, 2002; Williamson, 1996). Such invasions are now recognized as one of the major causes of biodiversity loss (Keller et al., 2011; Sala et al., 2000). Early detection of species known to be pests and/or invasive, therefore, is crucial in order for measures to be taken to control their spread before major damage is done.

Some taxa which are innocuous or only minor pests in their native regions have unforeseen consequences after arriving in new areas which are lacking competition or predators. For example, of the six most serious forestry pests introduced in North America, only the European gypsy moth had pest status in its indigenous range (Cock, 2003; cited in Armstrong and Ball, 2005). In New Zealand, the introduced painted apple moth, *Teia anartoides*, from Australia was predicted to cause 33-205 million Euros in damage if it was not eradicated (Armstrong and Ball, 2005). Additionally, some members of particular species complexes may be far more harmful than others. An example is the silverleaf whitefly, *Bemisia tabaci* (Gennadius, 1889). This species complex is thought to have originated in India, and in 1897 it was discovered in the United States, attacking sweet potatoes; it subsequently colonized Australia, Africa, and Europe. It damages diverse crops by feeding on them as well as by transmitting viral diseases. Within five years of discovery in the United States, the newer strain of this hemipteran species had caused over $100 million in damage to agricultural crops (Diaz-Soltero, 1995; Gould et al., 2008). These ecological properties underscore the critical importance of accurate, rapid, large-scale methods of biomonitoring in areas where bio-invasions may be suspected. Metabarcoding can be very adventitious for this purpose.

International standards for field collection, specimen preservation procedures, laboratory techniques, as well as data analysis of HTS data, are still needed in order to verify the reliability of results. It is difficult, however, to standardize procedures across variable ecosystems and research facilities all over the world. Nevertheless, the numerous studies successfully carried out in the less-than-two-decade-long history of DNA barcoding are a strong indication that this research should be

continued. Explanations of specific challenges in the bioinformatic (data generation and analysis) and in the laboratory aspects of metabarcoding are discussed in the next two sections.

## Bioinformatic challenges in the implementation of DNA metabarcoding

For metabarcoding, the Illumina MiSeq is one of the most commonly employed sequencers. Illumina, who produced the first HTS machine, continues to dominate the market today, providing sequencing technology that delivers high accuracy and low cost (Reuter, Spacek, and Snyder, 2015). The read length, however, remains short in comparison to those of other HTS sequencing technologies, such as the now defunct Roche 454 (Rothberg and Leamon, 2008), and PacBio (Rhoads and Au, 2015). The Illumina MiSeq produces reads of up to 300 bp in length, with run times between 6 hours and 3 days, making it ideal for small genome sequencing. Although Illumina sequencers were not designed for PCR amplicon sequencing, such as DNA barcodes, low sequence quality can be partly alleviated by employing paired-end sequencing (Unno, 2015). Paired-end sequencing is frequently used in metabarcoding, and it enables longer amplicons to be assembled, based on the partial overlaps in the 5' and 3' ends. This precludes sequencing of the entire 650-bp COI barcode region; but due to the nature of metabarcoding, shorter fragment lengths may actually be desirable. One reason is that DNA may be degraded, as is often the case with environmental DNA, or for mixed samples more generally, and another is that primers are needed which amplify target sequences of a broad range of species. The first such "universal" primer set was designed by Meusnier et al. (2008). All available COI reference sequences were bioinformatically analyzed to calculate the probabilities of having species-specific barcode regions of smaller sizes. A 130-bp long universal mini barcode was designed which successfully amplified a comprehensive set of taxa from all major eukaryotic groups. Many other mini-barcodes for large taxonomic groups (e.g. insects) have since been designed.

Due to the nature of the state of the art of HTS technology, a series of informatic steps, commonly referred to as a bioinformatic pipeline, must be applied to the sequences generated by HTS instruments. This is very different from the relatively simple manual editing and processing of each individual, full-length barcode sequences produced by Sanger sequencing. Bioinformatic pipelines for HTS generate MOTUs, either de novo, or based on a set of reference sequences of species expected to be in the sample, if one is available. Like any OTUs, these MOTUs are helpful when samples of unknown composition could contain species either without representation in reference databases or are undiscovered (although they cannot be uploaded directly to BOLD).

For processing of the sequence data all the way from raw HTS reads to molecular identification of OTUs, various DNA metabarcoding pipelines have been developed by bioinformaticians in recent years. Many such programs were developed for the analysis of environmental samples of microbes, but they can often also be utilized for COI metabarcoding, or

can be adapted to do so. Examples include Qiime (Caporaso et al., 2010), CD-HIT (Fu et al., 2012), USEARCH (Edgar, 2010), VSEARCH (Rognes et al., 2016), OBITools (Boyer et al., 2016), and JAMP (available online, www.github.com/VascoElbrecht/JAMP). Each of these software suites enable researchers to perform all of the necessary steps for the attainment of MOTUs from the raw sequence data produced by the HTS instrument. Most of them run on UNIX-like operating systems or are platform-independent and are available as free and open-source versions, requiring only a basic familiarity with the command line environment in order to use them. Alternatively, platforms such as Galaxy (Goecks et al., 2010) have been created so that researchers without any command-line knowledge may conduct computational biological analyses, including genomics and metabarcoding, entirely on a graphical user interface through a Web browser. Commercial software suites also exist, such as the QIAGEN® CLC Genomics Workbench, which offer a graphical user experience and run on either Microsoft Windows, Mac OS X, or Linux, although licenses are very expensive. This option is often sensible for private companies offering bioinformatic analysis of HTS data to their customers.

Whichever software tools are utilized, a prominent challenge in metabarcoding laboratories is to determine the optimal sequence of commands, and settings and parameters to use for any particular dataset. The main steps to be performed on metabarcoding data consist of demultiplexing, paired-end merging, removal of primers and adapter sequences, base-call quality filtering, chimera detection and removal, and clustering into MOTUs. Different algorithms are utilized by different software packages to achieve these objectives. Which algorithms are most optimal is a subject of ongoing research in bioinformatics and computational biology.

MOTU clustering, a step necessary in metabarcoding due to the hundreds of thousands of sequence reads typically produced in a HTS run, has been particularly well-researched through comparisons of different computational methods for creating clusters of sequences with the highest concordance to species (see Brannock and Halanych, 2015; Chen et al., 2013; Kopylova et al., 2016). Assuming a reference set of OTU barcodes of all species thought to potentially be in a sample is not available, scientists may opt for open-reference clustering if references for most of the species are available; but more commonly, clustering is performed de novo. De novo clustering is based on sequence similarities among all sequences, and is performed by one of three types of algorithms: (1) hierarchical clustering, used by software such as Mothur (Schloss et al., 2009); (2) heuristic clustering, used by software such as such as CD-HIT (Fu et al., 2012) and Uclust (Edgar, 2010); and (3) model-based clustering methods. To use hierarchical and heuristic methods, the user must set a minimum percent similarity threshold, between 95 and 99%, depending on the barcoding gap of their particular taxa. Some such programs, such as VSEARCH, also allow switching between distance-based and abundance-based clustering. To avoid having to set a hard cutoff in sequence similarity, whose optimal level may actually vary with the taxa within a sample, model-based

clustering algorithms were invented. The first model-based method, CROP, was proposed by Hao, Jiang, and Chen (2011). CROP uses an unsupervised probabilistic Bayesian clustering algorithm with a soft threshold for defining OTUs (Chen et al., 2013).

Frequently, some of the OTUs generated by metabarcoding pipelines do not have high percentage matches in reference databases. However, in these cases, a reverse taxonomical approach can still be taken. Recent studies, such as Morinière et al., 2019 (see appendix) demonstrate the ability of metabarcoding to register unknown as well as taxonomically challenging species, or "dark taxa". This is a reverse taxonomic approach in which sequenced taxa may be subsequently assigned a bionomen by taxonomic specialists. In this way, specimens belonging to unnamed molecular character-based units (MOTUs or BINs) can be assigned to known species, or to species new to science (see Page, 2016 and Geiger et al., 2016). In this case, it is especially important to utilize tools such as ABGD (Puillandre et al., 2012), which predict the likely barcode gaps in unidentified taxa. If these challenges are met, significant progress towards overcoming the taxonomic impediment can be made.

### Laboratory challenges in the implementation of DNA metabarcoding

Because there are multiple species--and often diversity at much higher taxonomic levels--present in each sample, amplifying as much of the DNA in them as possible is a challenge specific to metabarcoding, and it is a subject of ongoing research. Taxa which are small, rare, and underrepresented in environmental samples tend to be overlooked. Attempts to overcome this challenge by using DNA barcoding (with Sanger sequencing) are still very expensive. NGS can provide some solutions to this biodiversity challenge, but cost-reducing measures must also be taken in the laboratory procedures prior to sequencing (Elbrecht et al., 2017(a)). One of the main issues limiting the ability to detect the total biodiversity present in communities via metabarcoding is that of PCR primer bias (Elbrecht et al., 2017(b); Piñol et al., 2015). In order to decrease bias and amplify the range of taxa likely to be present in a sample, metabarcoding primers should be designed with optimal levels of degeneracy to maximize the species whose target sequences adhere to the primer sequence and can subsequently be amplified, while minimizing slippage, chimeric reads, and other instances of areas of DNA other than the intended ones adhering to the primers. However, as the COI barcode region exhibits high levels of degeneracy throughout its sequence, creation of “universal” primers is difficult (Deagle et al., 2014; Sharma and Kobayashi, 2014). Therefore, careful evaluation of primer sets has been advised (Elbrecht and Leese, 2017; Tedersoo et al., 2015). Multiple primer sets may also be utilized, in order to further increase the range of taxa amplified (Morinière et al., 2016).

An alternative way to eliminate PCR bias is to avoid locus-specific amplification entirely, and sequence genomic DNA directly. Such “PCR-free” methods have been tested for metabarcoding (e.g. Crampton-Platt et al., 2016; Krehenwinkel et al., 2017; Liu et al., 2016; Shokralla et al., 2016). Indeed, this approach does circumvent amplification bias and has been shown to recover more species from diverse samples, making it ideal for exhaustive community analyses. However, PCR-free methods are sensitive to copy number variations of the target genes, not to mention the much higher costs associated with sequencing entire genomes to sufficient depths, as well as more complicated laboratory and bioinformatic workflows. Because all of the DNA--genomic and mitochondrial--is amplified and sequenced, time-consuming computational work is required to separate out the desired barcode sequences, and algorithms may produce false positives in the cases of nuMTs. Thus, PCR-based approaches remain standard practice for large-scale community analyses (Krehenwinkel et al., 2017).

Along with primer bias, the other main issue impacting the ability of metabarcoding to recover representative sequences from 100% of the biodiversity present in a mixed sample is that of unequal specimen size (Elbrecht and Leese 2015; Elbrecht et al., 2017(b); Leray and Knowlton, 2015). This is due to the simple fact that larger specimens have more biomass and thus more DNA. Therefore, when DNA is extracted holistically, the largest individual specimens will contribute the most DNA, and the smallest individuals the least. Even when the goal of a metabarcoding survey is simply to determine which taxa are present or absent in a sample (without attempting to quantify their proportions), unequal specimen size can contribute to systematic failure to detect smaller taxa. Increasing the sequencing depth alleviates this problem to some extent (see Elbrecht et al., 2017(b)), but this of course increases the cost of sequencing, which may be a problem, especially for research facilities with limited budgets, such as those in developing countries.

The effects of unequal body size, or biomass, of taxa in typical bulk samples of invertebrates on the ability to maximize detection with metabarcoding had not been quantified until recently. While it is straightforward that a larger specimen contains more DNA than a smaller one, and hence contributes more genetic material to the holistic extraction from a mixed sample, in practicality it is time consuming to quantify the effects of specimen size bias on HTS yield. In a typical sample of invertebrates from the field, there are almost as many different sizes in a vial of collected insects as there are species. In many types of metabarcoding samples, even of macroscopic organisms, it can be difficult or impossible to visually identify or even quantify the specimens present. In some such cases, some of the DNA present exists as (exogenous) eDNA, parts of organisms, and/or immature life stages. In other cases, such as field samples from malaise traps, specimen numbers in the tens of thousands, especially with the presence of microhymenopterans and microdipterans, inhibit counting them.

Field samples collected by malaise trapping are often separated by e.g., order, and morphologically studied prior to being recombined and metabarcoded. With regard to this common practice, Morinière et al. (2016) performed an experiment with different primer sets for metabarcoding a sample of invertebrates from a malaise trap. Specimens were sorted by order, DNA was extracted from four major insect orders, and each order was sequenced separately. Morinière et al. observed each primer set's ability to detect BINs to vary by order, and accordingly supported the use of multiple primers in overcoming this amplification bias. However, they were not able to reach a robust conclusion regarding whether the pre-sorting of specimens by order itself contributed to increased taxon detections. Although they also saved aliquots of DNA extracted from each of the four orders and combined them into a pool which was additionally sequenced, it was sequenced at the same sequencing depth as each of the four orders individually. Whereas, in order to serve as a valid control, it would have to be sequenced at 4x the sequencing depth; or alternatively, the reads of the individual orders sequenced could be subsampled *in silico* to simulate ¼ sequencing depth, in a process known as rarefaction.

The first study to rigorously assess the effects of specimen size bias on OTU recovery was that of Elbrecht et al. (2017(b)). Working with freshwater bulk samples, they divided specimens into three classes of roughly equal body sizes, in order to approximately overcome this size bias. After separating the size classes, specimens were counted, and the total dry weights of each size class were recorded, to estimate biomass. Each size class was homogenized and lysed, and then from the lysed tissue, pools were created, from which DNA would be extracted. The "unsorted" pools contained lysate from each size category in proportion to its dry weight, to serve as a control. The "sorted" pools, on the other hand, were corrected for size bias, by pooling lysates in proportion to the total *numbers* of specimens in each size, rather than by total weights. Indeed, they observed that the sorted samples yielded 30% higher taxa detection than the unsorted. This method of correcting for size bias therefore enables sequencing depths to be reduced without losing in the detection ability of metabarcoding, an especial advantage when cost is an issue.

While presorting specimens by size and creating proportionate pools significantly increases recovery of OTUs, it also has the disadvantages of being much more time-consuming and labor-intensive, and all the more so with increased specimen numbers and decreased specimen sizes. When done correctly, the protocol is effective towards the goal of having every taxon in a sample represented by its barcode sequence; in practice, however, discretion should be used when deciding whether to use it. Obviously, it cannot be used in cases where destroying the specimens is either undesirable or against regulations. Furthermore, it may be contraindicated for samples containing anything other than whole specimens which can be reasonably counted. This often presents a problem for malaise trap samples or samples where eDNA is present; and in actuality, these are common scenarios. Ultimately, the most appropriate countermeasure to the bias caused by unequal

specimen size, can be considered to depend upon the individual circumstances of the particular metabarcoding effort.

Facing practical, financial, scientific, and potential regulatory concerns, nondestructive methods of DNA extraction for mixed and bulk samples have been increasing in popularity in recent years. These methods involve utilizing eDNA from the ethanol in which the bulk sample had been preserved, instead of extracting DNA from the specimens directly, or alternatively, immersing the sample into lysis buffer for only a short time ("semilysis"). The most optimal nondestructive methodology is a subject of ongoing research, with studies showing varied results. Hajibabaei et al. (2012) found similar species recovery rates when DNA was extracted following evaporation of preservative ethanol, compared with homogenization of benthic invertebrate communities. Their methods included taking 10 subsamples of the original preservative ethanol from each sample, evaporating them, and then re-dissolving the residues into molecular water, and using this as the source of template DNA for PCR. Zizka et al. (2019) tested a specialized filtration DNA extraction method combined with three protocols designed to enhance the release of DNA into the fixative prior to extraction. They observed lower success in recovery of Coleoptera, Gastropoda, and Trichoptera sequences from ethanol-derived than from homogenized bulk aquatic samples. One possible reason given is that the tough chitinous exoskeletons or shells inhibit release of DNA into the fixative. Small taxa were also underrepresented in the ethanol. Carew et al. (2018) also observed lower detection rates for heavily sclerotized taxa, such as Coleoptera, in samples extracted with a nondestructive semilysis method, compared to homogenized samples. Conversely, some taxa may be overrepresented in ethanol. It may be due to soft bodies, such as in Clitellata and Diptera, or the regurgitated stomach content of predator insects being released into the ethanol, enabling their prey species to be detected (Zizka et al., 2019). Overall, biases inherent in ethanol metabarcoding may be further overcome in the future, as protocols continue to be developed.

## Summary of Results

*Summary of Publication I: DNA Metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species*

We have undertaken a wide-range, multi-year survey with the goal of providing early warning of pest and invasive invertebrates in the Bavarian Forest National Park. Malaise traps were set up in 2016 and 2018 at the same sites: three outside and six inside the park, and were emptied ten times per year throughout the growing seasons. Each of the 180 samples was metabarcoded and sequence data analyzed. Metabarcoding is useful for early detection of potentially invasive pests in that DNA can be obtained from either the captured specimens themselves, or from preservative ethanol. The latter increases the chances of detecting invasions or pests at low levels, possibly

before they could be observed visually. To this end, a list of species of interest was compiled from various literature sources to include as many terrestrial invertebrates designated as either pests or actual or potential invasive species to Germany, as possible. Species of interest whose barcode sequences were available on BOLD were compiled into publicly available datasets with DOIs. Species of interest without sequences on BOLD were downloaded from NCBI Genbank if available there. All reference sequences from both sources were then combined to create a custom database for pest and invasive invertebrates. Sample OTU sequences were compared to this, as well as to a downloaded database of all arthropod sequences on BOLD, for standard biodiversity analytics.

Biodiversity data based on read abundances and 7-level taxonomic assignments was analyzed by Principal Component Analysis, revealing clustering in concordance with whether collection sites were located inside or outside of the National Park. Next, Jaccard distance matrices of the presences of BINs at the nine collection sites were constructed for each of the two survey years and were shown by a Mantel test to be significantly correlated with each other. This provides evidence in support of the repeatability of metabarcoding for terrestrial biodiversity analysis. Additionally, we compared the data obtained in this study with data obtained from a voucher-based German malaise trap survey under the Global Malaise Trap Program (GMTP), which was conducted in the NPBW during 2012 (see Geiger et al., 2016). Overall similar patterns in the presence of total arthropod BINs, as well as BINs belonging to four major arthropod orders across the study area, were observed in both survey years, and were also comparable with the GMTP 2012 data (downloaded from BOLD).

Of the BOLD BIN-based database records to which OTUs matched at >= 97%, roughly half had species-level taxonomic classifications in BOLD. Based on presence and absence of BINs, a Mantel test revealed a significant correlation between matrices of the mean Jaccard distances by trap sites in 2016 with those of 2018 ($r = 0.4995$, $p = 0.005$). Based on read abundances, Principal Component Analyses of biodiversities in each trap showed that traps located outside of the National Park clustered farther from the sites inside the park. ANOSIM tests also showed significant differences between BIN compositions in traps inside vs outside of the park for both years (2106 $r = 0.2$, $p = 2e\text{-}04$; 2018 $r = 0.239$, $p = 1e\text{-}04$).

Of the 402 species on the pest and invasive reference database, two were detected at >= 97% similarity, with one likely false positive due to BIN sharing, leaving one likely accurate detection. This conclusion was based on construction of neighbor-joining gene trees through the bold website (www.boldsystems.org). *Dendrolimus superans* (Butler, 1877) and *Ips duplicatus* (Sahlberg, 1836) are both listed on the Warning List of the German National Institute for Nature Conservation (Rabitsch et al., 2013). *D. superans* (BOLD:AAB6845), a species which has never been observed in Germany, matched at 99.55% identity in malaise trap sample T1-52 (inside the National Park), collection September I, 2016. However, it shares a BIN with *Dendrolimus pini (Linnaeus, 1758)*,

which is known as an occasional pest throughout most of Europe, including Germany, and the two can be observed clustering closely together (Appendix, Fig. 8). *D. pini* was also detected in the same sample at 100% identity. Therefore, it is highly unlikely that *D. superans* was actually in the sample.

Conversely, *I. duplicatus* matched at 98.64% identity to the database in malaise trap T3-50, collection July II, 2018, filtered ethanol sample. The genus *Ips* is commonly known as bark beetles, with *I. duplicatus* being endemic to northern Europe. It is a pest of pine trees (*Pinus* spp.) and is unknown if it poses a threat to biodiversity. It is was unknown in Germany at the time of publication of the warning list, but has recently been spreading southward, through central, eastern, and southern Europe (Fiala & Holuša, 2019). Although another congeneric species, *Ips typographus* (Linnaeus, 1758), a keystone species in the Bavarian Forest National Park (Müller et al., 2008), was also detected in the same trap at 100% identity, these two species' barcode sequences cluster less closely together, and they do not share a BIN. Therefore, it is likely to be a case of correct molecular identification of *I. duplicatus*, and to represent the first detection of this invasive saproxylic beetle in the Bavarian Forest National Park.

*Summary of Publication II: A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring*

This study provided a summary of the results of a DNA Barcoding campaign of German Diptera with a focus on dark taxa. The three main goals were to (1) provide a DNA barcode library for 5,200 species (BINs) of Diptera; (2) demonstrate by the example of bulk extractions from a malaise trap experiment that DNA barcode clusters, labelled with globally unique identifiers (such as OTUs and/or BINs), provide a pragmatic, accurate solution to the 'taxonomic impediment'; and (3) demonstrate that interim names based on BINs and OTUs obtained through metabarcoding is an effective method for studies on species-rich groups that are usually neglected in biodiversity research projects because of their unresolved taxonomy. A reference sequence library was created through the barcoding of over 45,000 individual specimens. This library consists of full-length barcode sequences of approximately 5,100 species (2,453 named species that were assigned to 2,500 BINs, and another 2,700 unnamed BINs--so-called "dark taxa"). These 5,200 BINs included representatives of 88 of 117 (75%) of the dipteran families known from Germany. More than a third (1,829) of the BINs were new to BOLD.

Previously, most dipteran families have been taxonomically inaccessible due to the lack of specialists and resultant taxonomic impediment. But with DNA barcoding and metabarcoding, we have accomplished the creation of an interim taxonomic system for half of all German Diptera. The metabarcoding portion of this study entailed extracting DNA from malaise trap samples, running an appropriate pipeline, and creating MOTUs de novo. Although these do not meet the requirements for BIN assignment, MOTUs can be assigned as interim taxonomic nomenclature and serve as

references for putative species. For dark taxa, the ABGD tool was used to validate MOTU clustering. For individually sequenced specimens, neighbor-joining trees and the TaxCl-approach for detecting taxonomic incongruences (Rulik et al., 2017), revealed high levels of congruence with morphology-based identifications. Among families, an inverse relationship between the number of dark taxa and average body size was also observed.

*Summary of Publication III: DNA Barcoding in Forensic Entomology – Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Species*

This study, which was conducted at the ZSM in collaboration with forensic biologists of the Bavarian State Police, has contributed significantly towards the aim of establishing more comprehensive reference databases for insects relevant to forensic research. Through a field experiment involving two decomposing porcine corpses, bulk samples of arthropods were collected using pit falls, net swings, and selective sampling. Samples were collected and data recorded at specific time points throughout each stage of decomposition (fresh, bloated stage, active decay stage, and advanced decay stage). It resulted in the contribution of 120 high quality sequences, with 46 newly added species belonging to 11 distinct orders.

The DNA barcode reference library at the ZSM was extended by 54.5% in terms of species-count through DNA extraction, PCR and Sanger Sequencing. Metabarcoding facilitated the volume of insect material sampled throughout the 9-month-long experiment; with HTS, a total of 469 species were identified within a time frame corresponding to three to four weeks of laboratory procedures. Based on resultant OTU data, detailed presence-absence diagrams were constructed for a subset of 137 species chosen based on their forensic relevance. Diagrams were constructed for these species throughout the stages of decomposition (fresh, bloated stage, active decay stage, and advanced decay stage). It was observed that, while some species were present throughout the entire experiment, others seemed to appear only on distinct days or during distinct periods. These observations enable more detailed pictures to emerge of the taxonomic characterizations of colonization and how they change over time.

*Summary of Publication IV: High Throughput Sequencing as a novel quality control method for industrial yeast starter cultures*

Today's fermentation products market owes much of its success to the use of pure starter cultures. In this study, we have made a contribution to the state-of-the-art technology of the quality control of industrial yeast starter cultures for the brewing of wine and special beer fermentations. Currently, methods consist of the use of either selective media or targeted approaches via Real-Time PCR. With these methods, however, researchers can only test for specific strains of potentially interfering spoilage yeasts, and if there is a spoilage yeast strain present which was not specifically

tested for, it would be overlooked. Here, we have demonstrated that metabarcoding of the 26 S rDNA D1/D2 regions of yeast chromosome XII can be effective at identifying which species are present in a given starter culture, without prior suspicion of their identities. In total, eight of 14 samples of supposedly pure starter cultures (7 samples of *Saccharomyces cerevisiae* and one sample of *Torulaspora delbrueckii*) were confirmed to be pure by 26 S rDNA metabarcoding, while six showed indications of contaminating spoilage yeast strains. The results show that it is possible to detect differing species in supposedly pure yeast cultures by application of the new method. Some strains showed potential traits of intraspecific hybridization, horizontal gene transfer or syntrophic cultures, which interfered with the results. The 26S rDNA D1/D2 region showed to be discriminative for only some species, indicating a need to additionally utilize more discriminative regions, such as ITS1. Moreover, a more comprehensive database needs to be built up in order to improve molecular identifications.

*Summary of Additional Project: Food Security*

We collaborated with researchers at the Bavarian Ministry for Health and Food Security (LGL, Landesamt für Gesundheit und Lebensmittelsicherheit) in a project to test the authenticity of the ingredients of several exotic meats sold over the internet. Consumers who want to purchase meats not typically available in supermarkets or other domestic sources have been buying them online for over a decade. However, despite often being higher-priced, these consumable goods are not subject to the same quality control as food products manufactured in Germany. In this study, we supplemented the LGL's standard methods of DNA-Chip and Sanger sequencing with DNA metabarcoding. The main advantage of DNA metabarcoding in this application is that it is a non-targeted approach. A DNA Chip, for example, allows testing for the presence of 24 species on each of eight samples. While testing is generally performed for most species suspected of being in the samples, there is always a possibility that species not tested for are present in the samples. Such cases are highly likely to be discovered by metabarcoding, with its ability to test mixed and bulk samples for even trace amounts of any taxa with representation in the reference libraries.

Results showed that the emu steak, python steak, and reindeer topside were all consistent with stated ingredients. The camel steak, however, which was stated as Bactrian camel, was instead found to contain dromedary camel meat. Four other samples in this study constituted clear cases of fraudulent labeling. These were dried sausages, produced by two different manufacturers, each sold as a different species of mammal from southern Africa: kudu, springbuck, oryx, and ibex. All four, however, were found to consist of red deer (*Cervus elaphus*) meat instead. Results of this project were published as the article *„Wilde“ Zustände beim Online-Handel?* in the German trade magazine Deutsche Lebensmittel-Rundschau: Zeitschrift für Lebensmittelkunde und Lebensmittelrecht 115(March):98–102.

# General Discussion

DNA barcoding and metabarcoding have greatly augmented the knowledge base of the molecular taxonomic characterizations of individuals and communities of all kinds. Metabarcoding has presented scientists with novel opportunities to scale up environmental biodiversity studies to paradigm-shifting levels, which is likely to be particularly important during a biodiversity crisis. It also has challenges in its implementation. How to optimize metabarcoding pipelines and methods is a subject of ongoing research, to which some of the results of the projects presented in this thesis have contributed. Through performing a variety of application studies employing DNA barcoding and metabarcoding, we have demonstrated what is possible, as well as experienced some of the potential pitfalls and practical limitations to implement these relatively new technologies in the fields of biodiversity monitoring, forensic entomology, and quality control of consumable products in the food and brewing industries. Moreover, we have performed preliminary testing of different methods for their ability to enhance HTS results by increasing DNA yield of bulk samples of invertebrates, as well as the necessity of destructive methods of DNA extraction. Explanations of difficulties encountered in each application study, and how achievements of the studies' objectives have enhanced the current state of knowledge, follow.

*Biodiversity survey in the Bavarian Forest National Park*

We were able to assess invertebrate biodiversity in the largest national park in Europe, via DNA metabarcoding. Using both presence-absence and read count-based analyses, we observed trends in frequencies of observations of taxa throughout across time, utilizing bulk samples from malaise traps at sites inside and outside of the park, de novo OTU generation, and existing reference libraries. The results of the NPBW survey are completely based on molecular taxonomic identification methods. In DNA barcoding, BLAST hits of >= 97% or 98% identity are commonly used to indicate species-level matches. Of course, this is only a general rule of thumb, and the barcode gap varies between taxa. Likewise, gene trees were valuable in assessing whether either of the two species of interest detected at >= 97% were true positives. A potential drawback, though, is that the specimens cannot be morphologically examined, due to the destructive nature of the DNA extraction methods utilized. Although this particular case does not leave much room for reasonable doubt, there could be other cases in which morphological examination is more critical in order to reach conclusions regarding the presence of taxa in bulk samples. For this reason, the exclusive use of nondestructive extraction methods may be desirable in some cases. Caution should be taken, however, when utilizing preservative ethanol as a source of DNA. It has been tested less extensively

than homogenization, and while some studies do show promising results, in practicality the quality of DNA obtained from preservative ethanol varies widely, being affected by preservation methods and storage conditions prior to extraction, as well as by the particular nondestructive extraction method used.

The main issues encountered in the pest and invasive species detection portion of this study included the limited availability of reference sequences for species of interest, and uncertainty regarding high but partial (above 97% but less than 100%) matches to the database, particularly when destructive methods of DNA extraction prevented specimens from further morphological analysis. Such are well-known limitations encountered in metabarcoding. Scientists are engaged in research for the development of improved techniques, so that metabarcoding pipelines can be applied on large scales with high reliability. This is particularly important with regard to research performed for governmental agencies, because protocols must be followed which have been proven to be sufficiently reliable in order that legislative action should be taken on their basis. A major question in the metabarcoding community is that of how to optimize the methods in ways applicable across different situations, i.e. ecological environments and taxonomic classifications of study organisms, as well as region-specific concerns such as compliance with regional or national laws regarding specimen collection, preservation, and destruction, and research facility-related concerns, such availability of laboratory and computing technologies. Overcoming these barriers will enable a higher level of reproducibility and reliability of DNA metabarcoding results.

*Comparison of nondestructive DNA extraction methods*

A comparison of three nondestructive DNA extraction methods was also attempted while the 2018 samples from the NPBW were in the laboratory. The nondestructive methods chosen were: 1. evaporation of subsamples of their original preservative ethanol and subsequent DNA extraction (evaporated ethanol), 2. filtration of their original preservative ethanol and extraction of DNA from the filters (filtered ethanol), and 3. submersion of the samples in lysis buffer solution for a short time and extracting DNA from the liquid (semilysis). Unfortunately, this experiment was not originally planned or incorporated into the budget, but was conceived of later on. Applying the semilysis method to all samples, for example, would have doubled the amounts of proteinase and insect lysis buffer required. Therefore, since we did not have adequate amounts of reagents, we only tested small subsets of the 90 samples from 2018. Also, having failed to design and agree upon a sound experimental methodology before beginning the laboratory work, we did not apply all extraction methods in question to the same samples. We tested only five samples with semilysis, five with ethanol filtration, and 45 with ethanol evaporation. As all samples were eventually homogenized for the biodiversity survey, the most extraction methods performed on a single sample was three: on sample T3-50B, evaporated ethanol, filtered ethanol, and tissue homogenization. On one other

sample, semilysis, evaporated ethanol, and homogenization were applied. While one sample is not enough to draw any conclusions worthy of inclusion in a methods paper, these comparisons provided some preliminary data. Additionally, ethanol evaporation being performed on 45 samples enabled a more adequate comparison with homogenization. The following is a summary of the unpublished results of the DNA extraction methods comparisons.

Evaporated ethanol yielded by far the fewest reads and OTUs of all methods. Taxonomic composition also varied the most in the evaporated ethanol samples, compared with the tissue. In these samples, we observed that greater numbers of OTUs and BINs were detected in most families of Diptera, Coleoptera, Lepidoptera, and Hymenoptera, with a few exceptions, including small flies such as Cecidomyiidae, Chironomidae, and Psychodidae, as well as highly sclerotized beetles such as Carabidae, Scirtidae, and Staphylinidae. In sample T3-50B, July II, for which ethanol evaporation, filtration, and tissue homogenization were performed, ethanol filtration yielded an order of magnitude more reads than evaporation, and the order composition of OTUs more closely resembled that of the homogenized tissue than did that of the evaporation method. In the sample for which semilysis as well as evaporated ethanol methods were performed in addition to homogenization, the largest overlap in BINs was observed between the semilysis and tissue homogenization methods. Although more of the BINs recovered by homogenization were also recovered with semilysis, the ethanol methods revealed more taxa which were not found by homogenizing the specimens.

As suggested by previous studies (e.g. Linard et al., 2016; Zizka et al., 2019), taxa detected by ethanol-based methods but not in semilysis or tissue homogenization methods may represent traces of genetic material such as regurgitated gut content of predatory arthropods, or other environmental DNA. Some taxa are underrepresented in ethanol and others overrepresented in ethanol, but this is also an issue with tissue homogenization due to primer bias and unequal specimen sizes. In our results, many taxa were missed entirely in either the tissue samples or the ethanol method(s) but detected in the other(s). Moreover, the invasive species, *I. duplicatus*, would have been missed in our study had we not utilized ethanol-based methods. Taking all factors into consideration, it may be advisable in many cases to utilize a combination of extraction methods for DNA metabarcoding. Semilysis and ethanol filtration may be an especially effective combination when complete specimen destruction is to be avoided. The exclusive use of ethanol may be necessary in some cases if the specimens must not be destroyed. In these cases, it is advisable to adhere to state-of-the art recommendations for maximizing DNA yield, as eDNA is currently a subject of a host of research. DNA yield has been shown to be heavily dependent on a combination of DNA capture, preservation, and extraction methods (Hinlo et al., 2017). It has also been observed that freezing bulk samples in ethanol prior to extraction increased DNA yield, but shaking or sonicating them did not (Zizka et al., 2019). Zizka et al. (2019), who achieved good results, changed the ethanol once prior to extraction; and notably, they utilized the filtration method to extract DNA

from the ethanol. Some studies have shown up to 100% recovery rates of sequences of species in bulk samples from the evaporative ethanol method (e.g. Hajibabaei 2012), but multiple subsamples of ethanol were used from each bulk sample. High numbers of subsamples may not be feasible for large scale biodiversity monitoring efforts. Filtration of entire samples, on the other hand, concentrates genetic material onto the filter paper. Filters can then be easily transported to the laboratory for extraction. Our own preliminary results do show much higher yields of reads and OTUs from filtered compared to evaporated ethanol; and previous experiences in our laboratory at the ZSM (J. Morinière, unpublished work) have shown severely diminished DNA yields from evaporated ethanol methods. However, these samples had been transported from the field as well as stored at room temperature during the summer due to practical constraints, and it is better for the preservation of DNA to store samples at lower temperatures, such as refrigerating or freezing them (see Hinlo et al., 2017).

Overall, the results of metabarcoding the malaise trap samples from the NPBW provide support for the recommendation that, whenever possible, a combination of extraction methods should be applied. This way, molecular identifications can be maximized, without resorting to methods which incur additional time and/or costs. Overall, metabarcoding of ethanol for large-scale biodiversity monitoring shows great promise, and--as long as appropriate care is taken to follow optimal protocols--its continued research, development, and real-world applications should be recommended.

*Should bulk environmental samples be pre-sorted by size?*

As specimen size bias is one of the major issues hampering complete recovery of barcode sequences from the biodiversity of bulk samples, we attempted to replicated Elbrecht et al. (2017)'s size sorting experiment on malaise trap samples as well. The goal was to test the effects of presorting specimens by size on metabarcoding efficiency. As would be expected, it was hypothesized that size sorting would increase OTU detection similarly as was shown by Elbrecht et al. (2017); however, applying the laboratory procedures of counting and weighing to malaise trap contents proved prohibitively difficult. Firstly, the trap contained an order of magnitude more specimens than Elbrecht et al. (2017(b))'s, and secondly, many specimens were barely visible to the naked eye, such as microhymenopterans and microdipterans. These factors contributed to unforeseen complications in the laboratory work.

Initially, specimens had been stored in ethanol, and it was first attempted to perform size sorting (using a series of sieves), dry the specimens, and then count the individuals in, and measure the dry weights of, each size class. Static electricity interfered with the process, however, preventing transference from PVC storage tubes. This also exacerbated specimen breakage (of legs, antennae, etc.), a known issue with size sorting. Subsequently, we attempted to count specimens in ethanol

before drying them. This was successful for all but the smallest size class (< 1 mm body size). Although a stereomicroscope made specimens visible, a technique was not able to be perfected to accurately count out thousands of tiny individuals floating in ethanol under the stereomicroscope, especially when broken pieces were not able to be distinguished from whole insects. This resulted in inaccurate counts. Weighing, in any case, must be performed after evaporating off the ethanol, as dry weights are desired in order to closely approximate biomasses. The smallest size class (XS) again proved problematic to implement. When dry, these were again affected by static electricity and subtle air currents, contributing to specimen loss and impairing transfer to and from a laboratory balance. Then, we attempted to estimate the weight of the XS contingent by a sampling-with-replacement method. The problem there, though, was that many specimens were of such low mass that they did not register on even the most sensitive balance in the laboratory. We were, therefore, unable to calculate an accurate weight or estimate with margin of error for this size class (although such an estimation is mathematically possible and was performed on the other size classes).

Furthermore, difficulties were encountered in applying lysis buffer. When volumes of lysis buffer solution added to the size classes were not proportionate to their respective weights, subsequent pooling of lysed specimen solutions was not proportionate. Rather, some lysates were more dilute than others, thereby invalidating the crux of the experiment. Although DNA concentrations are adjusted after extraction, it didn't alleviate the problem, possibly because the lysates had to be handled gently, rotating them only enough to distribute the buffer solution and specimens roughly evenly, and not turned upside down or shaken, in order to avoid losing or damaging the DNA. However, without thorough mixing, the DNA may have either sunk to the bottom, or otherwise been unevenly distributed throughout the lysis tubes, resulting in aliquots taken for pooling to vastly differ in the amounts of DNA contained, thereby losing representation of some taxa nonuniformly.

To avoid the problem of uneven distribution of DNA throughout lysis buffer solutions, proportionate pooling may be conducted in either of two other ways: pooling the ground tissue (after homogenization and before lysis), or pooling the already extracted DNA solutions. The latter was not attempted, due to another bias introduced by application of this method to mixed samples; namely, different taxa have different concentrations of mitochondria. So even when specimen size is accounted for, some taxa, particularly flying insects, are richer in mitochondrial DNA, therefore contributing greater amounts of COI and other mitochondrial genes to the extraction than other taxa (Elbrecht et al., 2017(b)).

Pooling ground tissue was attempted as a second method. Difficulties arose, however, upon realizing that not enough ground tissue from the XS size class remained after having created the original set of pools, due to the fact that a much higher proportion of it had to be taken for the Sorted simulation pool than Unsorted, in correcting for the high specimen number but small specimen size

of the XS size class. Therefore, the tissue pools were scaled down. As creation of tissue pools is indeed difficult (see Elbrecht et al., 2017(b)), scaling them down proved even harder to ensure the aliquots taken from each size class were representative, due to dealing with tiny amounts of ground insects. Although a high quality homogenizer had been used, inevitably some taxa cannot be safely homogenized completely, and highly sclerotized pieces of insects remained which were clearly visible to the naked eye. The smallest aliquot by far was taken from the XL fraction, as it had the highest total biomass but lowest specimen count; and it was apparently too small to be fully representative. Analysis of the HTS results revealed that most of the species and BINs known to have body sizes greater than 7.15 mm (mesh size used to separate the XL fraction of the samples), which were present in the lysis pools and Unsorted tissue pool, to be entirely missing from the Sorted tissue pool results.

Overall, our attempt to replicate Elbrecht et al. (2017(b))'s size sorting experiment on samples of terrestrial invertebrates from malaise traps proved to be infeasible from a practical standpoint. Pre-sorting by size is time consuming, particularly because of the steps included in proportionate pooling, a necessary next step to achieve the gain in taxa recovery for a given sequencing depth; and our efforts have shown that it may also be too difficult to implement on some types of mixed samples of invertebrates. The practical constraints on this method have led some metabarcoding researchers to abandon this method in favor of other methods for maximizing taxon identification. It may be concluded that size sorting should be utilized at the discretion of the individual researcher or laboratory, who should consider the properties of their specific samples, particularly the approximate total number of specimens and evenness of distribution of specimens between the size classes they are to be sorted into. Size sorting may be ideal when each size class contains a few hundred specimens at most, and breakage of individuals is minimal. When specimens number in the thousands or more, and there is a very skewed distribution among size classes, as was the case in our malaise trap, it would not be recommended.

*"Flying in the Dark" (metabarcoding as a reverse-taxonomic approach)*

The results of the DNA barcoding campaign of German flies and midges, which included "dark taxa", have specifically provided a significant contribution to the molecular taxonomic knowledge base of Diptera. Taxonomists working on Diptera have long been aware of the immense number of undescribed species versus approximately 160,000 named species (Borkent et al., 2018; Pape et al., 2011). Hebert et al. (2016), applying DNA barcoding to Canadian insects, propose that the actual number of species could be much higher, suggesting the possible presence of 1.8 million species in just one family, the Cecidomyiidae (gall midges) alone. Although this estimate may be high, it is very likely that this single family includes more species than are currently described for the order. Owing to a lack of specialists and the taxonomic impediment, most dipteran families have

historically been taxonomically inaccessible. But with DNA barcoding and metabarcoding, we accomplished the creation of an interim taxonomic system for half of all German Diptera. Ideally, it will provide a backbone for future taxonomic studies. As the reference libraries will continue to grow, and gaps in the species catalogue will be filled, BIN lists assembled by barcoding can gain incremental taxonomic resolution.

In addition to the reference sequences created by barcoding over 45,000 voucher specimens, the metabarcoding portion of the study contributed de novo MOTUs, which, although they do not meet the requirements for BIN assignment, can be assigned as interim taxonomic nomenclature and serve as references for putative species. When dealing with barcode sequences from unidentified specimens, it is important to be aware of their barcode gaps, or the difference between intraspecific and interspecific variation, for each taxonomic group, instead of relying on a generic similarity cutoff for all specimens. Since the barcode gaps have not yet been established for dark taxa, the ABGD tool was used to validate OTU clusters' likely concordance with predicted species, as is advisable whenever the barcoding gap of a taxon is not known.

The inverse relationship between mid-range body size and number of dark taxa within dipteran families suggests a higher incidence of cryptic diversity and overlooked species among families with the smallest body sizes--a phenomenon long suspected by taxonomists--and that the number of dipteran species in Germany is likely to be much higher than previously recognized. In some families, higher numbers of BINs were even observed than the numbers of known species in Germany. In Cecidomyiidae for example, 930 BINs were encountered in a few malaise traps, while only 836 species are known in Germany (Schumann et al., 1999). It can be concluded that the true diversity of Diptera in Germany, Europe, and the world has been seriously underestimated, a conclusion also reached in several other studies (e.g. Erwin, 1982; Hebert et al., 2016; May, 1988; Ødegaard, 2000).

Although the project aimed to develop a comprehensive DNA barcode library, resource constraints meant that only half of the specimens sorted to a family or better taxonomy could be analyzed. There is little doubt that many species and genera that are currently absent from the reference library remain within this sorted material, making the remains of the bulk samples a valuable resource for any future effort to complete the reference library. Finally, this study has highlighted the potential of DNA barcoding and metabarcoding to aid ongoing efforts to conserve the world's fauna, because these technologies substantially enhance our ability to identify--and thus conserve--biodiversity.

*Forensic entomology*

In forensic biology, estimation of the PMI, or elapsed time from the death of an individual until discovery of the body, is of paramount importance. Thus, a major research objective of forensic

entomology is to increase the accuracy of PMI estimation based on the characterization of insect communities which colonize corpses. Our pilot study succeeded in its main goal of expanding the available reference library of DNA barcoding sequences for forensically relevant arthropod species for forensic biologists around the world. We have added representatives of orders which were not yet present, and especially notable was the identification of the larvae of *Chrysomya albiceps* (Wiedemann, 1819), whose larvae were officially recorded for the first time on carrion in Bavaria, Germany (see appendix, Fig. S2 B).

The DNA barcode reference library at the ZSM was extended by 54.5% in terms of species-count through DNA extraction, PCR and Sanger Sequencing. 469 species-level molecular identifications from metabarcoding samples collected from the pig cadavers were achieved in a time frame of 4 weeks. The OTU table generated by our bioinformatic pipeline was able to be successfully used for the construction of detailed presence-absence diagrams. While diagrams displaying the succession of arthropod groups throughout the stages of decomposition of a carrion source are typical in forensic entomology, the resolution of these diagrams is in most cases limited either to the family-level or to very few single species. In our study, however, we created four different presence-absence matrices to include all possible variations between the two pig cadavers and the forensically relevant arthropod orders Coleoptera and Diptera.

As the aim of this study was to simply test the usage of high throughput sequencing on bulk samples collected from decomposing material, we have evaluated the generated data as distinct application examples without aiming at answering specific scientific questions. Hence, it is a pilot study, rather than a controlled experiment. Nonetheless, having implemented appropriate scientific rigor, we have discovered that this application can relieve the sampling and laboratory workload drastically, and provide high volumes of data, compared to traditional methods. In fact, the bulk samples generated more sequence data than could reasonably be processed, which is why we limited the scope to the two orders most important in forensic entomology: Coleoptera (beetles) and Diptera (flies). Overall, we demonstrated the usefulness of HTS to the discipline of forensic entomology, through increasing the volume and biodiversity of arthropod material able to be analyzed, in a shorter amount of time. Greater numbers of identified species can lead to increased accuracy of estimations of PMI, and help to fill in gaps in the knowledge base of species of potential forensic usefulness, where the possibility cannot be excluded that some taxa which are actually useful for this application have been previously overlooked.

*Brewing and food security*

Our collaborative application studies in the quality control of yeast starter cultures and exotic meat products were overall successes. In both cases, the main advantage that metabarcoding confers is that it is non-targeted, meaning that it can test for an almost unlimited number of species.

Indeed, we detected the presence of species other than the desired ones, which had not been targeted by standard methods (e.g. DNA-Chip, real-time PCR, DNA barcoding using Sanger sequencing). Of course, as is always the case in barcoding, the main limitation on which species can be detected is the reference library. Another limitation encountered in the yeast culture application was the discriminative ability of the particular marker in use, in this case the 26s rDNA D1/D2 region. This marker is frequently used in the barcoding of fungi, as opposed to animals, which have COI. Other potential reasons for possible false positive contaminant detections are mostly particular to yeasts, namely intraspecific hybrids, horizontal gene transfer, and syntrophic cultures. We concluded that further investigations are needed, and that future work should include additional markers, such as ITS1. If these improvements can be implemented, we may have been the first to demonstrate the application of metabarcoding as another significant technological improvement in the centuries-old industry of wine and beer brewing.

The food security application study was in some ways more straightforward to implement, being that the food was all from animal sources (meat), and therefore COI could be used, as per usual. Furthermore, all of the meat was from mammals, and Mammalia has very good coverage in the reference databases. It is easy to see here how DNA metabarcoding potentially saved researchers investigating suspected food fraud a lot of time. While it might be suspected that a manufacturer would substitute one species of camel for another, it is less obvious to test for red deer meat in sausages purported to be made exclusively from southern African artiodactyls. Thanks to our collaborative efforts, the LGL submitted formal complaints of illegal misbranding of food products. Going forward, they will customarily test the authenticity of exotic meat products sold over the internet to consumers in Germany. They even anticipate DNA metabarcoding to potentially become their primary technique of choice for this purpose.

We also performed some preliminary testing on novel food sources, i.e. insects sold as food. Ideally, the proven benefits DNA metabarcoding has brought to the food quality control industry can be implemented in the growing novel food industry as well. However, we encountered some additional difficulties in our initial testing. The primary difficulty was not inherent to Insecta or insect DNA specifically, but actually was likely due to the composition of ingredients of the finished products we tested. Exo Cricket Bars were one of the earliest widely available insect-containing food products on the market in the United States of America and are available worldwide today. Their success may be due in part to the fact that they are made with cricket “flour” (ground crickets), as opposed to whole insects, and are therefore less objectionable to much of the targeted consumer base. However, we were not able to amplify COI-5’ DNA in adequate amounts to proceed in the laboratory with barcoding. The presence of many other ingredients in the cricket bars likely contributed to the difficulty, considering that we had somewhat greater success with pure insect

“flours”. Mealworm flour yielded sequences unambiguously molecularly identified as *T. molitor.* But another insect flour sold from the same online source was not able to be identified to species level with the reference library downloaded from BOLD at the time, due to the reference library’s incomplete taxonomical classifications, but a match to the listed species was found by BLASTing against GenBank. Keeping the struggles we encountered mind for the design of future studies, we still hope to apply metabarcoding to the emerging industry of entomophagy, and that it can assist more and more people--individual consumers as well as food producers--in making this choice which is healthy for both human beings and the planet.

*General Concluding remarks and outlook*

All of the projects presented in this thesis have been based on generally accepted DNA metabarcoding techniques, which remained largely the same throughout the various application projects. Based on the successes and failures experienced when performing laboratory work on the projects involving mixed samples of invertebrates, it has become clear that homogenizing them prior to DNA extraction is generally the most efficient method, in agreement with the literature (see Carew et al., 2013; Carew et al., 2018; Hajibabaei et al., 2012). The main disadvantage of this method is that specimens are destroyed. In cases where this is acceptable, however, homogenization is time-efficient and generally reliable, and can be recommended insofar as it has had the highest record of success in our laboratory. Regarding the analysis of sequence data, we have constructed bioinformatic pipelines from free software and customized them to the needs of each project in this thesis. It requires only enough technical skills to use software packages appropriately, as determined by existing bioinformatic research. Still, much room for improvement exists, for the optimization of laboratory as well as computational methods. In order to realize the potentials of DNA barcoding and metabarcoding, many practical constraints on its implementation must be overcome. Biodiversity monitoring is an especially crucial undertaking at the current time, and in order to be carried out on a large scale, a great deal of time and effort is required. As metabarcoding has already been successfully implemented as a time-and cost-effective method, I have presented in this thesis the results of one such large scale biomonitoring effort, as well as the practical constraints and pitfalls encountered in our particular situation. By presenting our experiences, I hope to inform members of the scientific community involved in metabarcoding by making the discoveries as well as difficulties shared knowledge which can enhance and accelerate future efforts. All of the application projects presented have contributed towards the general body of biodiversity and taxonomic knowledge and/or towards the refinement of practical techniques in the implementation of DNA metabarcoding to achieve various practical purposes.

As we cannot afford to ignore any longer, the Earth is undergoing a biodiversity crisis. While it is known that species extinctions per year number in the hundreds to thousands (Chivian and

Bernstein, 2008), projecting future changes in biodiversity is difficult. Not only do we not know precisely how factors such as habitat destruction, species invasions, and climate change will interact (Bellard et al., 2012; Mantyka-Pringle et al., 2015; Segan et al., 2016), but also most of the word's extant species are still undescribed, and we cannot protect what we do not know. Robust scientific evidence is needed for lawmakers to enact policies regarding conservation and other forms of environmental protection. Therefore, there is an urgent need for biodiversity monitoring to be taken to the next level. Practical needs to conserve life on earth demand high throughputs of information, and some of the bottleneck on the amounts of information generated by traditional biodiversity monitoring has been alleviated by metabarcoding. It cannot be denied that the pressure to discover species before they become extinct is great, as is the need to monitor changes in ecosystems as early as possible so that damage may be minimized. In order to take it to the next level, a paradigm shift in biomonitoring is needed, incorporating novel methodologies to disentangle causative factors for biodiversity decline in ecosystems, in order to determine the most effective courses of action to take. This emphasizes a need to prioritize the ability to generate vast amounts of information on biodiversity quickly, even if results are not immediately as precise for every organism as they can be with the addition of morphological taxonomic analysis. Public availability of barcoding data is a significant asset, though, especially with regard to the taxonomic deficit and impediment. Optimization of collaboration between traditional and molecular taxonomists has been facilitated by recent developments in the state of the art and should be continued in pursuit of the crucial goal of characterizing and conserving life on Earth.

# References


Archaux, Frédéric. 2011. “On Methods of Biodiversity Data Collection and Monitoring.” *Revue Science Eaux & Territoires, Public Policy and Biodiversity* 2011 (N03bis): 70–75.

Armstrong, K. F., and S. L. Ball. 2005. “DNA Barcodes for Biosecurity: Invasive Species Identification.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1462): 1813–23.

Aylagas, Eva, Angel Borja, and Naiara Rodríguez-Ezpeleta. 2014. “Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI).” *PloS One* 9 (3): e90529.

Banu, Shahera, Wenbiao Hu, Yuming Guo, Cameron Hurst, and Shilu Tong. 2014. “Projecting the Impact of Climate Change on Dengue Transmission in Dhaka, Bangladesh.” *Environment International* 63 (February): 137–42.

Bellard, Céline, Cleo Bertelsmeier, Paul Leadley, Wilfried Thuiller, and Franck Courchamp. 2012. “Impacts of Climate Change on the Future of Biodiversity.” *Ecology Letters*. https://doi.org/10.1111/j.1461-0248.2011.01736.x.

Best, Troy L., Robert M. Sullivan, Joseph A. Cook, and Terry L. Yates. 1986. “Chromosomal, Genic, and Morphologic Variation in the Agile Kangaroo Rat, Dipodomys Agilis (Rodentia: Heteromyidae).” *Systematic Zoology*. https://doi.org/10.2307/2413384.

Bhat, Ashaq Hussain, Department of Biotechnology K. S. Rangasamy College of Technology, Tiruchengode, India, Department of Biotechnology K. S. Rangasamy College of Technology, Tiruchengode, India, and Puniethaa Prabhu. 2017. “OTU Clustering A Window to Analyse Uncultured Microbial World.” *International Journal of Scientific Research in Computer Science and Engineering*. https://doi.org/10.26438/ijsrcse/v5i6.6268.

Blaxter, Mark, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyualem Abebe. 2005. “Defining Operational Taxonomic Units Using DNA Barcode Data.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1462): 1935–43.

Borkent, Art, Brian V. Brown, Peter H. Adler, Dalton de Souza Amorim, Kevin Barber, Daniel Bickel, Stephanie Boucher, et al. 2018. “Remarkable Fly (Diptera) Diversity in a Patch of Costa Rican Cloud Forest: Why Inventory Is a Vital Science.” *Zootaxa* 4402 (1): 53–90.

Boyer, Frédéric, Céline Mercier, Aurélie Bonin, Yvan Le Bras, Pierre Taberlet, and Eric Coissac. 2016. “Obitools: A Unix-Inspired Software Package for DNA Metabarcoding.” *Molecular Ecology Resources* 16 (1): 176–82.

Brannock, Pamela M., and Kenneth M. Halanych. 2015. “Meiofaunal Community Analysis by High-Throughput Sequencing: Comparison of Extraction, Quality Filtering, and Clustering Methods.” *Marine Genomics* 23 (October): 67–75.

Brix, Saskia, Florian Leese, Torben Riehl, and Terue Cristina Kihara. 2015. “A New Genus and New Species of Desmosomatidae Sars, 1897 (Isopoda) from the Eastern South Atlantic Abyss Described by Means of Integrative Taxonomy.” *Marine Biodiversity*. https://doi.org/10.1007/s12526-014-0218-3.

Buckley, Ralf C. 2015. “Grand Challenges in Conservation Research.” *Frontiers in Ecology and Evolution*. https://doi.org/10.3389/fevo.2015.00128.

Burkett-Cadena, Nathan D., and Amy Y. Vittor. 2018. “Deforestation and Vector-Borne Disease: Forest Conversion Favors Important Mosquito Vectors of Human Pathogens.” *Basic and Applied Ecology*. https://doi.org/10.1016/j.baae.2017.09.012.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods* 7 (5): 335–36.

Carew, Melissa E., Rhys A. Coleman, and Ary A. Hoffmann. 2018. “Can Non-Destructive DNA Extraction of Bulk Invertebrate Samples Be Used for Metabarcoding?” *PeerJ*. https://doi.org/10.7717/peerj.4980.

Carew, Melissa E., Vincent J. Pettigrove, Leon Metzeling, and Ary A. Hoffmann. 2013.

"Environmental Monitoring Using Next Generation Sequencing: Rapid Identification of Macroinvertebrate Bioindicator Species." *Frontiers in Zoology* 10 (1): 45.
Carvalho, Marcelo R. de, Flávio A. Bockmann, Dalton S. Amorim, Mário de Vivo, Mônica de Toledo-Piza, Naércio A. Menezes, José L. de Figueiredo, et al. 2005. "Revisiting the Taxonomic Impediment." *Science*.
Chen, Wei, Clarence K. Zhang, Yongmei Cheng, Shaowu Zhang, and Hongyu Zhao. 2013. "A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0070837.
Chimeno, Caroline, Jérôme Morinière, Jana Podhorna, Laura Hardulak, Axel Hausmann, Frank Reckel, Jan E. Grunwald, Randolph Penning, and Gerhard Haszprunar. 2018. "DNA Barcoding in Forensic Entomology - Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Species." *Journal of Forensic Sciences* 64 (2): 593–601.
Chivian, Eric, and Aaron Bernstein. 2008. *Sustaining Life: How Human Health Depends on Biodiversity*. Oxford University Press.
Cock, Matthew J. W. 2003. "Biosecurity and Forests: An Introduction with particular emphasis on forest pests." *Forest Resources Development Service*, Forest Resources Division FAO, Rome, Italy. Working Paper FBS/2E.
Coleman, Charles Oliver. 2015. "Taxonomy in Times of the Taxonomic Impediment – Examples from the Community of Experts on Amphipod Crustaceans." *Journal of Crustacean Biology*. https://doi.org/10.1163/1937240x-00002381.
Collins, R. A., and R. H. Cruickshank. 2013. "The Seven Deadly Sins of DNA Barcoding." *Molecular Ecology Resources* 13 (6): 969–75.
Cox, Andrea J., and Paul D. N. Hebert. 2001. "Colonization, Extinction, and Phylogeographic Patterning in a Freshwater Crustacean." *Molecular Ecology*. https://doi.org/10.1046/j.1365-294x.2001.01188.x.
Crampton-Platt, Alex, Douglas W. Yu, Xin Zhou, and Alfried P. Vogler. 2016. "Mitochondrial Metagenomics: Letting the Genes out of the Bottle." *GigaScience* 5 (March): 15.
Cruaud, Perrine, Jean-Yves Rasplus, Lillian Jennifer Rodriguez, and Astrid Cruaud. 2017. "High-Throughput Sequencing of Multiple Amplicons for Barcoding and Integrative Taxonomy." *Scientific Reports* 7 (February): 41948.
Davies, Thomas, Andrew Cowley, Jon Bennie, Catherine Leyshon, Richard Inger, Hazel Carter, Beth Robinson, et al. 2018. "Popular Interest in Vertebrates Does Not Reflect Extinction Risk and Is Associated with Bias in Conservation Investment." *PLOS ONE*. https://doi.org/10.1371/journal.pone.0203694.
Deagle, Bruce E., Simon N. Jarman, Eric Coissac, François Pompanon, and Pierre Taberlet. 2014. "DNA Metabarcoding and the Cytochrome c Oxidase Subunit I Marker: Not a Perfect Match." *Biology Letters* 10 (9). https://doi.org/10.1098/rsbl.2014.0562.
DeFoliart, Gene R. 1992. "Insects as Human Food." *Crop Protection*. https://doi.org/10.1016/0261-2194(92)90020-6.
Doyle, J. J., and B. S. Gaut. 2000. "Evolution of Genes and Taxa: A Primer." *Plant Molecular Biology* 42 (1): 1–23.
Duellman, William E., and Pablo Venegas. 2005. "MARSUPIAL FROGS (ANURA: HYLIDAE: GASTROTHECA) FROM THE ANDES OF NORTHERN PERU WITH DESCRIPTIONS OF TWO NEW SPECIES." *Herpetologica*. https://doi.org/10.1655/04-105.1.
Ebach, Malte C., and Craig Holdrege. 2005. "DNA Barcoding Is No Substitute for Taxonomy." *Nature*. https://doi.org/10.1038/434697b.
Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61.
Elbrecht, Vasco, and Florian Leese. 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.
Elbrecht, Vasco and Florian Leese. 2017. "Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment." *Frontiers in Environmental*

*Science*, *5*, 11.
Elbrecht, Vasco, Ecaterina Edith Vamos, Kristian Meissner, Jukka Aroviita, and Florian Leese. 2017(a). “Assessing Strengths and Weaknesses of DNA Metabarcoding-Based Macroinvertebrate Identification for Routine Stream Monitoring.” *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210x.12789.
Elbrecht, Vasco, Bianca Peinert, and Florian Leese. 2017(b). “Sorting Things out: Assessing Effects of Unequal Specimen Biomass on DNA Metabarcoding.” *Ecology and Evolution* 7 (17): 6918–26.
Ferri, Gianmarco, Milena Alù, Beatrice Corradini, Manuela Licata, and Giovanni Beduschi. 2009. “Species Identification through DNA ‘Barcodes.’” *Genetic Testing and Molecular Biomarkers* 13 (3): 421–26.
Fischer, Bob. 2016. “Bugging the Strict Vegan.” *Journal of Agricultural and Environmental Ethics*. https://doi.org/10.1007/s10806-015-9599-y.
Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. “DNA Primers for Amplification of Mitochondrial Cytochrome c Oxidase Subunit I from Diverse Metazoan Invertebrates.” *Molecular Marine Biology and Biotechnology* 3 (5): 294–99.
Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52.
Garnett, Stephen T., and Les Christidis. 2017. “Taxonomy Anarchy Hampers Conservation.” *Nature* 546 (7656): 25–27.
Geiger, M. F., J. J. Astrin, T. Borsch, U. Burkhardt, P. Grobe, R. Hand, A. Hausmann, et al. 2016. “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*. https://doi.org/10.1139/gen-2015-0185.
Geldmann, Jonas, Jacob Heilmann-Clausen, Thomas E. Holm, Irina Levinsky, Bo Markussen, Kent Olsen, Carsten Rahbek, and Anders P. Tøttrup. 2016. “What Determines Spatial Bias in Citizen Science? Exploring Four Recording Schemes with Different Proficiency Requirements.” *Diversity and Distributions*. https://doi.org/10.1111/ddi.12477.
Goecks, Jeremy, Anton Nekrutenko, James Taylor, and The Galaxy Team. 2010. “Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences.” *Genome Biology*. https://doi.org/10.1186/gb-2010-11-8-r86.
Gottwalt, Allison. 2014. "Impacts of Deforestation on Vector-borne Disease Incidence." *Global journal of health science* 3(2):16-19.
Grada, Ayman, and Kate Weinbrecht. 2013. “Next-Generation Sequencing: Methodology and Application.” *The Journal of Investigative Dermatology* 133 (8): e11.
Habel, Jan Christian, Andreas Segerer, Werner Ulrich, Olena Torchyk, Wolfgang W. Weisser, and Thomas Schmitt. 2016. “Butterfly Community Shifts over Two Centuries.” *Conservation Biology: The Journal of the Society for Conservation Biology* 30 (4): 754–62.
Hajibabaei, Mehrdad, Shadi Shokralla, Xin Zhou, Gregory A. C. Singer, and Donald J. Baird. 2011. “Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos.” *PloS One* 6 (4): e17497.
Hajibabaei, Mehrdad, Jennifer L. Spall, Shadi Shokralla, and Steven van Konynenburg. 2012. “Assessing Biodiversity of a Freshwater Benthic Macroinvertebrate Community through Non-Destructive Environmental Barcoding of DNA from Preservative Ethanol.” *BMC Ecology*. https://doi.org/10.1186/1472-6785-12-28.
Hallmann, Caspar A., Martin Sorg, Eelke Jongejans, Henk Siepel, Nick Hofland, Heinz Schwan, Werner Stenmans, et al. 2017. “More than 75 Percent Decline over 27 Years in Total Flying Insect Biomass in Protected Areas.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0185809.
Hao, Xiaolin, Rui Jiang, and Ting Chen. 2011. “Clustering 16S rRNA for OTU Prediction: A Method of Unsupervised Bayesian Clustering.” *Bioinformatics* 27 (5): 611–18.
Hartl, D. L., Clark, A. G., and Clark, A. G. 1997. *Principles of population genetics* (Vol. 116).

Sunderland, MA: Sinauer associates.
Hausmann, Axel, Andreas H. Segerer, Thomas Greifenstein, Johannes Knubben, Jerôme Morinière, Vedran Bozicevic, Dieter Doczkal, Armin Günter, Werner Ulrich, and Jan Christian Habel. 2020. “Toward a Standardized Quantitative and Qualitative Insect Monitoring Scheme.” *Ecology and Evolution*. https://doi.org/10.1002/ece3.6166.
Hebert, Paul D. N., Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. 2003. “Biological Identifications through DNA Barcodes.” *Proceedings of the Royal Society of London. Series B: Biological Sciences*. https://doi.org/10.1098/rspb.2002.2218.
Hebert, Paul D. N., and T. Ryan Gregory. 2005. “The Promise of DNA Barcoding for Taxonomy.” *Systematic Biology* 54 (5): 852–59.
Hebert, Paul D. N., Sujeevan Ratnasingham, Evgeny V. Zakharov, Angela C. Telfer, Valerie Levesque-Beaudin, Megan A. Milton, Stephanie Pedersen, Paul Jannetta, and Jeremy R. deWaard. 2016. “Counting Animal Species with DNA Barcodes: Canadian Insects.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1702). https://doi.org/10.1098/rstb.2015.0333.
Hinlo, Rheyda, Dianne Gleeson, Mark Lintermans, and Elise Furlan. 2017. “Methods to Maximise Recovery of Environmental DNA from Water Samples.” *PloS One* 12 (6): e0179251.
Iwobi, A., D. Sebah, G. Spielmann, M. Maggipinto, M. Schrempp, I. Kraemer, L. Gerdes, U. Busch, and I. Huber. 2017. “A Multiplex Real-Time PCR Method for the Quantitative Determination of Equine (horse) Fractions in Meat Products.” *Food Control*. https://doi.org/10.1016/j.foodcont.2016.11.035.
Jian, Chen, Qiu Deyi, Yue Qiaoyun, Hu Jia, Liu Dexing, Wei Xiaoya, and Zheng Leiqing. 2014. “A Successful Case of DNA Barcoding Used in an International Trade Dispute.” *DNA Barcodes*. https://doi.org/10.2478/dna-2014-0004.
Karvonen, Anssi, Päivi Rintamäki, Jukka Jokela, and E. Tellervo Valtonen. 2010. “Increasing Water Temperature and Disease Risks in Aquatic Systems: Climate Change Increases the Risk of Some, but Not All, Diseases.” *International Journal for Parasitology*. https://doi.org/10.1016/j.ijpara.2010.04.015.
Klausnitzer, B. 2006. “Stiefkinder der Entomologie in Mitteleuropa.” *Beiträge zur Entomologie* 56: 360–368.
Knowlton, Nancy, and Lee A. Weigt. 1998. “New Dates and New Rates for Divergence across the Isthmus of Panama.” *Proceedings of the Royal Society of London. Series B: Biological Sciences*. https://doi.org/10.1098/rspb.1998.0568.
Kopylova, Evguenia, Jose A. Navas-Molina, Céline Mercier, Zhenjiang Zech Xu, Frédéric Mahé, Yan He, Hong-Wei Zhou, Torbjørn Rognes, J. Gregory Caporaso, and Rob Knight. 2016. “Open-Source Sequence Clustering Methods Improve the State of the Art.” *mSystems* 1 (1). https://doi.org/10.1128/mSystems.00003-15.
Krehenwinkel, Henrik, Madeline Wolf, Jun Ying Lim, Andrew J. Rominger, Warren B. Simison, and Rosemary G. Gillespie. 2017. “Estimating and Mitigating Amplification Bias in Qualitative and Quantitative Arthropod Metabarcoding.” *Scientific Reports* 7 (1): 17668.
Larsen, Brendan B., Elizabeth C. Miller, Matthew K. Rhodes, and John J. Wiens. 2017. “Inordinate Fondness Multiplied and Redistributed: The Number of Species on Earth and the New Pie of Life.” *The Quarterly Review of Biology*. https://doi.org/10.1086/693564.
Leray, Matthieu, and Nancy Knowlton. 2015. “DNA Barcoding and Metabarcoding of Standardized Samples Reveal Patterns of Marine Benthic Diversity.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (7): 2076–81.
Lister, Bradford C., and Andres Garcia. 2018. “Climate-Driven Declines in Arthropod Abundance Restructure a Rainforest Food Web.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (44): E10397–406.
Liu, Shanlin, Xin Wang, Lin Xie, Meihua Tan, Zhenyu Li, Xu Su, Hao Zhang, et al. 2016. “Mitochondrial Capture Enriches Mito-DNA 100-Fold, Enabling PCR-Free Mitogenomics Biodiversity Analysis.” *Molecular Ecology Resources* 16 (2): 470–79.
Mace, Georgina M. 2004. “The Role of Taxonomy in Species Conservation.” *Philosophical*

*Transactions of the Royal Society of London. Series B, Biological Sciences* 359 (1444): 711–19.
Mantyka-Pringle, Chrystal S., Tara G. Martin, and Jonathan R. Rhodes. 2013. “Interactions between Climate and Habitat Loss Effects on Biodiversity: A Systematic Review and Meta-Analysis.” *Global Change Biology*. https://doi.org/10.1111/gcb.12148.
Mayr, E. 1942. "Systematics and the Origin of Species." *Columbia Univ. Press, New York*.
Meier, Rudolf, Winghing Wong, Amrita Srivathsan, and Maosheng Foo. 2016. “$1 DNA Barcodes for Reconstructing Complex Phenomes and Finding Rare Species in Specimen-Rich Samples.” *Cladistics*. https://doi.org/10.1111/cla.12115.
Meusnier, Isabelle, Gregory A. C. Singer, Jean-François Landry, Donal A. Hickey, Paul D. N. Hebert, and Mehrdad Hajibabaei. 2008. “A Universal DNA Mini-Barcode for Biodiversity Analysis.” *BMC Genomics* 9 (May): 214.
Michel, M., Hardulak, L. A., Meier-Dörnberg, T., Morinière, J., Hausmann, A., Back, W., Haszprunar, G., Jacob, F., and Hutzler, M. (2019). “High throughput sequencing as a novel quality control method for industrial yeast starter cultures.” *BrewingScience* 72 (March/April): 63-68.
Miyamoto, Michael M. 1981. “Congruence Among Character Sets in Phylogenetic Studies of the Frog Genus Leptodactylus.” *Systematic Zoology*. https://doi.org/10.2307/2413250.
Mora, Camilo, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. 2011. “How Many Species Are There on Earth and in the Ocean?” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1001127.
Morinière, Jérôme, Michael Balke, Dieter Doczkal, Matthias F. Geiger, Laura A. Hardulak, Gerhard Haszprunar, Axel Hausmann, et al. 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.
Morinière, Jérôme, Bruno Cancian de Araujo, Athena Wai Lam, Axel Hausmann, Michael Balke, Stefan Schmidt, Lars Hendrich, et al. 2016. “Species Identification in Malaise Trap Samples by DNA Barcoding Based on NGS Technologies and a Scoring Matrix.” *PloS One* 11 (5): e0155497.
Ødegaard, Frode. 2000. “How Many Species of Arthropods? Erwin’s Estimate Revised.” *Biological Journal of the Linnean Society*. https://doi.org/10.1111/j.1095-8312.2000.tb01279.x.
Ostfeld, R. S. 2009. “Biodiversity Loss and the Rise of Zoonotic Pathogens.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 15 Suppl 1 (January): 40–43.
Page, Roderic D. M. 2016. “DNA Barcoding and Taxonomy: Dark Taxa and Dark Texts.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1702). https://doi.org/10.1098/rstb.2015.0334.
Pante, E., C. Schoelinck, and N. Puillandre. 2015. “From Integrative Taxonomy to Species Description: One Step beyond.” *Systematic Biology* 64 (1): 152–60.
Pape, Thomas, Vladimir Blagoderov, and Mikhail B. Mostovski. 2011. “Order Diptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal Biodiversity: An Outline of Higher-Level Classification and Survey of Taxonomic Richness.” *Zootaxa*. https://doi.org/10.11646/zootaxa.3148.1.42.
Piñol, J., G. Mir, P. Gomez-Polo, and N. Agustí. 2015. “Universal and Blocking Primer Mismatches Limit the Use of High-Throughput DNA Sequencing for the Quantitative Metabarcoding of Arthropods.” *Molecular Ecology Resources* 15 (4): 819–30.
Potts, Simon G., Jacobus C. Biesmeijer, Claire Kremen, Peter Neumann, Oliver Schweiger, and William E. Kunin. 2010. “Global Pollinator Declines: Trends, Impacts and Drivers.” *Trends in Ecology & Evolution* 25 (6): 345–53.
Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012. “ABGD, Automatic Barcode Gap Discovery for Primary Species Delimitation.” *Molecular Ecology*. https://doi.org/10.1111/j.1365-294x.2011.05239.x.
Raclariu, Ancuta Cristina, Michael Heinrich, Mihael Cristin Ichim, and Hugo de Boer. 2018. “Benefits and Limitations of DNA Barcoding and Metabarcoding in Herbal Product Authentication.” *Phytochemical Analysis: PCA* 29 (2): 123–28.

Ratnasingham, Sujeevan, and Paul D. N. Hebert. 2007. “Bold: The Barcode of Life Data System (http://www.barcodinglife.org).” *Molecular Ecology Notes* 7 (3): 355–64.
Ratnasingham, Sujeevan, and Paul D. N. Hebert. 2013. “A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System.” *PloS One* 8 (7): e66213.
Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder. 2015. “High-Throughput Sequencing Technologies.” *Molecular Cell* 58 (4): 586–97.
Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89.
Riedel, Alexander, Katayo Sagata, Yayuk R. Suhardjono, Rene Tänzler, and Michael Balke. 2013. “Integrative Taxonomy on the Fast Track - towards More Sustainability in Biodiversity Research.” *Frontiers in Zoology* 10 (1): 15.
“Robert Edgar on ‘Usearch.’” 2010. *SciVee*. https://doi.org/10.4016/19198.01.
Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. “VSEARCH: A Versatile Open Source Tool for Metagenomics.” *PeerJ*. https://doi.org/10.7717/peerj.2584.
Rothberg, Jonathan M., and John H. Leamon. 2008. “The Development and Impact of 454 Sequencing.” *Nature Biotechnology* 26 (10): 1117–24.
Santos, Daubian, Stephanie Sampronha, and Charles Morphy Dias Santos. 2017. “Advances on Dipterology in the 21st Century and Extinction Rates.” *Papéis Avulsos de Zoologia (São Paulo)*. https://doi.org/10.11606/0031-1049.2017.57.33.
Savolainen, Vincent, Robyn S. Cowan, Alfried P. Vogler, George K. Roderick, and Richard Lane. 2005. “Towards Writing the Encyclopaedia of Life: An Introduction to DNA Barcoding.” *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2005.1730.
Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23): 7537–41.
Schumann, H., Bährmann, R. & Stark, A. (Hrsg.) 1999: *Entomofauna Germanica 2.* Checkliste der Dipteren Deutschlands. In: *Studia Dipterologica Supplement.* Bd. 2, Ampyx-Verlag, Halle (Saale), ISBN 3-932795-01-6, ISSN 1433-4968, S. 1-354.
Segan, Daniel B., Kris A. Murray, and James E. M. Watson. 2016. “A Global Assessment of Current and Future Biodiversity Vulnerability to Habitat Loss–climate Change Interactions.” *Global Ecology and Conservation*. https://doi.org/10.1016/j.gecco.2015.11.002.
Shaklee, J. B., and C. S. Tamaru. 1981. “Biochemical and Morphological Evolution of Hawaiian Bonefishes (Albula).” *Systematic Biology*. https://doi.org/10.1093/sysbio/30.2.125.
Sharma, Pranay, and Tsuyoshi Kobayashi. 2014. “Are ‘universal’ DNA Primers Really Universal?” *Journal of Applied Genetics*. https://doi.org/10.1007/s13353-014-0218-9.
Shockley, Marianne, and Aaron T. Dossey. 2014. “Insects for Human Consumption.” *Mass Production of Beneficial Organisms*. https://doi.org/10.1016/b978-0-12-391453-8.00018-2.
Shokralla, Shadi, Joel F. Gibson, Ian King, Donald J. Baird, Daniel H. Janzen, Winnie Hallwachs, and Mehrdad Hajibabaei. 2016. “Environmental DNA Barcode Sequence Capture: Targeted, PCR-Free Sequence Capture for Biodiversity Analysis from Bulk Environmental Samples.” https://doi.org/10.1101/087437.
Siddig, Ahmed A. H., Aaron M. Ellison, Alison Ochs, Claudia Villar-Leeman, and Matthew K. Lau. 2016. “How Do Ecologists Select and Use Indicator Species to Monitor Ecological Change? Insights from 14 Years of Publication in Ecological Indicators.” *Ecological Indicators*. https://doi.org/10.1016/j.ecolind.2015.06.036.
Sorg, M., H. Schwan, W. Stenmans, and A. Müller. 2013. “Ermittlung der Biomassen flugaktiver Insekten im Naturschutzgebiet Orbroicher Bruch mit Malaise Fallen in den Jahren 1989 und 2013.” *Mitteilungen Entomologischer Verein Krefeld* 1: 1-5.
Sperling, Felix A. H., Gail S. Anderson, and Donal A. Hickey. 1994. “A DNA-Based Approach to the Identification of Insect Species Used for Postmortem Interval Estimation.” *Journal of Forensic*

*Sciences*. https://doi.org/10.1520/jfs13613j.
Stamer, Andreas. 2015. “Insect Proteins-a New Source for Animal Feed: The Use of Insect Larvae to Recycle Food Waste in High-Quality Protein for Livestock and Aquaculture Feeds Is Held Back Largely Owing to Regulatory Hurdles.” *EMBO Reports* 16 (6): 676–80.
Tedersoo, Leho, Sten Anslan, Mohammad Bahram, Sergei Põlme, Taavi Riit, Ingrid Liiv, Urmas Kõljalg, et al. 2015. “Shotgun Metagenomes and Multiple Primer Pair-Barcode Combinations of Amplicons Reveal Biases in Metabarcoding Analyses of Fungi.” *MycoKeys*. https://doi.org/10.3897/mycokeys.10.4852.
Thomson, Scott A., Richard L. Pyle, Shane T. Ahyong, Miguel Alonso-Zarazaga, Joe Ammirati, Juan Francisco Araya, John S. Ascher, et al. 2018. “Taxonomy Based on Science Is Necessary for Global Conservation.” *PLoS Biology*.
Trontelj, Peter, Yoichi Machino, and Boris Sket. 2005. “Phylogenetic and Phylogeographic Relationships in the Crayfish Genus Austropotamobius Inferred from Mitochondrial COI Gene Sequences.” *Molecular Phylogenetics and Evolution* 34 (1): 212–26.
Troudet, Julien, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, and Frédéric Legendre. 2017. “Taxonomic Bias in Biodiversity Data and Societal Preferences.” *Scientific Reports* 7 (1): 9132.
Unno, Tasuya. 2015. “Bioinformatic Suggestions on MiSeq-Based Microbial Community Analysis.” *Journal of Microbiology and Biotechnology* 25 (6): 765–70.
Vanbergen, Adam J., and the Insect Pollinators Initiative. 2013. “Threats to an Ecosystem Service: Pressures on Pollinators.” *Frontiers in Ecology and the Environment*. https://doi.org/10.1890/120126.
Walsh, J. F., D. H. Molyneux, and M. H. Birley. 1993. “Deforestation: Effects on Vector-Borne Disease.” *Parasitology* 106 Suppl: S55–75.
Wheeler, Q. D. 2004. “Taxonomy: Impediment or Expedient?” *Science*. https://doi.org/10.1126/science.303.5656.285.
Willette, Demian A., Sara E. Simmonds, Samantha H. Cheng, Sofia Esteves, Tonya L. Kane, Hayley Nuetzel, Nicholas Pilaud, Rita Rachmawati, and Paul H. Barber. 2017. “Using DNA Barcoding to Track Seafood Mislabeling in Los Angeles Restaurants.” *Conservation Biology: The Journal of the Society for Conservation Biology* 31 (5): 1076–85.
Will, Kipling W., and Daniel Rubinoff. 2004. “Myth of the Molecule: DNA Barcodes for Species Cannot Replace Morphology for Identification and Classification.” *Cladistics*. https://doi.org/10.1111/j.1096-0031.2003.00008.x.
Wong, Wing Hing, Ywee Chieh Tay, Jayanthi Puniamoorthy, Michael Balke, Peter S. Cranston, and Rudolf Meier. 2014. “‘Direct PCR’ Optimization Yields a Rapid, Cost-Effective, Nondestructive and Efficient Method for Obtaining DNA Barcodes without DNA Extraction.” *Molecular Ecology Resources* 14 (6): 1271–80.
World Bank, World Development Indicators. 2016. “Forest area (sq. km)” https://data.worldbank.org/indicator/AG.LND.FRST.K2. Accessed 04.06.2020.
Yu, Douglas W., Yinqiu Ji, Brent C. Emerson, Xiaoyang Wang, Chengxi Ye, Chunyan Yang, and Zhaoli Ding. 2012. “Biodiversity Soup: Metabarcoding of Arthropods for Rapid Biodiversity Assessment and Biomonitoring.” *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210x.2012.00198.x.
Zell, Roland. 2004. “Global Climate Change and the Emergence/re-Emergence of Infectious Diseases.” *International Journal of Medical Microbiology: IJMM* 293 Suppl 37 (April): 16–26.
Zizka, Vera M. A., Florian Leese, Bianca Peinert, and Matthias F. Geiger. 2019. “DNA Metabarcoding from Sample Fixative as a Quick and Voucher-Preserving Biodiversity Assessment Method.” *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada* 62 (3): 122–36.

# Appendices

## Publication I - DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species

Accepted for publication June 9th, 2020.

Citation: Hardulak, L.A., Morinière, J., Hausmann, A., Hendrich, L., Schmidt, S., Doczkal, D., Müller, J., Hebert, P.D.N., and Haszprunar, G. “DNA metabarcoding for biodiversity monitoring in a national park: Screening for invasive and pest species.“ *Molecular Ecology Resources* (2020 Jun 19). https://doi.org/10.1111/1755-0998.13212

Received: 24 September 2019 | Revised: 3 June 2020 | Accepted: 9 June 2020

DOI: 10.1111/1755-0998.13212

Check for updates

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

# DNA metabarcoding for biodiversity monitoring in a national park: Screening for invasive and pest species

Laura A. Hardulak[1,2] | Jérôme Morinière[1] | Axel Hausmann[1] | Lars Hendrich[1] | Stefan Schmidt[1] | Dieter Doczkal[1] | Jörg Müller[3,4] | Paul D. N. Hebert[5] | Gerhard Haszprunar[1]

[1]SNSB-Zoologische Staatssammlung München, Munich, Germany

[2]Ludwig-Maximilians-Universität München, Munich, Germany

[3]National Park Bavarian Forest, Grafenau, Germany

[4]Field Station Fabrikschleichach, Department of Animal Ecology and Tropical Biology, University of Würzburg, Biocenter, Rauhenebrach, Germany

[5]Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada

**Correspondence**
Laura A. Hardulak, SNSB-Zoologische Staatssammlung München, Münchhausenstraße 21, 81247 Munich, Germany.
Email: hardulak@snsb.de

**Present address**
Jérôme Morinière, AIM – Advanced Identification Methods GmbH, München, Germany

**Funding information**
Bundesministerium für Bildung und Forschung, Grant/Award Number: GBOL: BMBF FKZ 01LI1101 and 01LI1501; Bayerisches Staatsministerium für Wissenschaft, Forschung und Kunst

## Abstract

DNA metabarcoding was utilized for a large-scale, multiyear assessment of biodiversity in Malaise trap collections from the Bavarian Forest National Park (Germany, Bavaria). Principal component analysis of read count-based biodiversities revealed clustering in concordance with whether collection sites were located inside or outside of the National Park. Jaccard distance matrices of the presences of barcode index numbers (BINs) at collection sites in the two survey years (2016 and 2018) were significantly correlated. Overall similar patterns in the presence of total arthropod BINs, as well as BINs belonging to four major arthropod orders across the study area, were observed in both survey years, and are also comparable with results of a previous study based on DNA barcoding of Sanger-sequenced specimens. A custom reference sequence library was assembled from publicly available data to screen for pest or invasive arthropods among the specimens or from the preservative ethanol. A single 98.6% match to the invasive bark beetle *Ips duplicatus* was detected in an ethanol sample. This species has not previously been detected in the National Park.

**KEYWORDS**

biodiversity, DNA barcoding, invasive species, metabarcoding, monitoring, pest species

## 1 | INTRODUCTION

The worldwide decline in biodiversity currently presents an urgent challenge facing humanity, and slowing down or halting this decline is an objective of broad international political agreement (Thomsen & Willerslev, 2015). A major barrier to achieving this objective is the lack of knowledge of biodiversity states and patterns on a global scale (Geijzendorffer et al., 2016; Lindenmayer et al., 2012). Hundreds or possibly thousands of species become extinct each year (Ceballos & Ehrlich, 2018; Chivian & Bernstein, 2008), and conservation of biodiversity depends upon ongoing monitoring efforts which can elucidate patterns of distribution and abundances of species

[Correction added on 20-October-2020, after first online publication: Ludwig-Maximilians-Universität München was added for Laura A. Hardulak.]

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. *Molecular Ecology* Resources published by John Wiley & Sons Ltd

and populations (Geijzendorffer et al., 2016; Honrado, Pereira, & Guisan, 2016; Schmeller et al., 2015; Thomsen & Willerslev, 2015). A well-designed monitoring effort should provide an early warning of changes in the ecosystem which could otherwise become problems that are difficult or impossible to remediate (Bohmann et al., 2014; Lindenmayer et al., 2012). One such change is the introduction of animal and plant species to non-native geographical areas. With the globalization of trade, reduced travel time and immense passenger travel, species invasions have recently intensified (Keller, Geist, Jeschke, & Kühn, 2011; Sala et al., 2000), and are now one of the major recognized causes of biodiversity loss (Bellard, Cassey, & Blackburn, 2016; Ehrenfeld, 2010).

Accurate, rapid identifications of invasive species are needed to better manage the risks associated with alien species. An estimated 1% of all neozoans and neophytes become invasive with serious economic impacts (Meyerson & Reaser, 2002; Williamson, 1996). Some taxa which are innocuous or only minor pests in their native regions have unforeseen consequences after arriving in new areas lacking microbial control, competition or predators. For example, of the six most serious forestry pests introduced in North America, only the European gypsy moth had pest status in its indigenous range (Cock, 2003). In New Zealand, the introduced painted apple moth, *Orgyia anartoides* (Walker, 1855), from Australia was predicted to cause €33–205 million in damage if it was not eradicated (Armstrong & Ball, 2005).

Traditional biodiversity monitoring has relied on visual observation and identification of species and counting of individuals. These efforts may be hampered by a lack of available taxonomic expertise for morphological identifications, as well as nonstandard sampling techniques (Beng et al., 2016; Corlett, 2017; Ji et al., 2013; Thomsen & Willerslev, 2015). Towards the aim of fulfilling an urgent need for accurate large-scale biodiversity monitoring, molecular methods have been applied in recent years, particularly since the advent of DNA barcoding (Hebert, Ratnasingham, & de Waard, 2003). DNA barcoding (Hebert et al., 2003), the characterization of sequence variation in a standard DNA fragment, is a broadly applicable and objective method, which increases the speed and taxonomic resolution of specimen identification as well as reducing costs. In this way, DNA barcoding and, more recently, metabarcoding (Hajibabaei, Shokralla, Zhou, Singer, & Baird, 2011)—a process by which genetic material is extracted from mixed or bulk samples, amplified, sequenced by high-throughput sequencing (HTS) and analysed holistically—assist in augmenting biodiversity monitoring efforts (Ji et al., 2013). In its first few years, metabarcoding was shown to recover significant portions of existing biodiversity (Aylagas, Borja, & Rodríguez-Ezpeleta, 2014; Yu et al., 2012) and to reveal unknown patterns of biodiversity (Leray & Knowlton, 2015), and it has been successfully applied to large-scale biodiversity assessments (e.g. Elbrecht, Peinert, & Leese, 2017; Epp et al., 2012; Ji et al., 2013; Morinière et al., 2016; Shokralla, Spall, Gibson, & Hajibabaei, 2012; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012; Yu et al., 2012). DNA barcoding and metabarcoding also permit species-level identifications when only eggs, larvae or parts of specimens are available for analysis. These may be intercepted at borders (e.g. wooden pallets at airports, ports, railway stations) as they are transported by vectors or accidentally by humans, such as in the ballast waters of ships, or with animals and plants in the food trade (Borrell, Miralles, Do Huu, Mohammed-Geba, & Garcia-Vazquez, 2017). For these reasons, HTS has been considered the ideal method for early warning of invasive species (Comtet, Sandionigi, Viard, & Casiraghi, 2015).

In terrestrial ecosystems, macroinvertebrates are often stored directly in ethanol following their collection. DNA can subsequently be harvested either directly from the specimens or from the preservative. Maceration of the specimens followed by subsequent extraction of DNA from a subsample of the homogenate is commonly practised (Yu et al., 2012), and it is probably both the simplest and the most effective way of securing a representative DNA extract from a bulk sample for subsequent metabarcoding (Elbrecht et al., 2017). However, there is a growing need to integrate sequence-based with morphological research (Silva-Santos, Ramirez, Galetti, & Freitas, 2018), and requirements to keep specimens intact for subsequent morphological control sometimes exist. Therefore, the efficiency and effectiveness of various nondestructive methods of sample preparation and DNA extraction of mixed samples for metabarcoding is a subject of ongoing research.

Additionally, an issue impacting the ability of metabarcoding to recover sequences representing the total biodiversity of a holistically homogenized sample is the bias in primer competition due to unequal specimen size (Elbrecht & Leese, 2015; Elbrecht et al., 2017; Leray & Knowlton, 2015). Larger specimens have more biomass and thus more DNA to contribute to lysed tissue pools. Therefore, larger individuals become overrepresented in sequencing results, and smaller ones underrepresented, increasing the risk of failure to detect taxa with small body sizes. Nondestructive ethanol-based DNA extraction methods have been recommended for their potential to provide solutions to sampling and vouchering challenges of metabarcoding (Hajibabaei, Spall, Shokralla, & van Konynenburg, 2012); and specifically, an ethanol filtration method has been shown to exhibit weak or even no correlation between specimen biomass and read numbers (Zizka, Leese, Peinert, & Geiger, 2019), thus potentially remediating the size-bias problem. As an objective of the present study is qualitative biodiversity analysis of mixed samples of invertebrates, we decided to supplement the standard homogenized tissue DNA extraction method with ethanol-based methods in 2018, in order to improve taxon recovery rates. The aims of the present study are to (a) perform biodiversity analysis comparing collection sites in and around the Bavarian Forest National Park (Nationalpark Bayerischer Wald, NPBW) and in two study years; and (b) construct a custom database of potential pest and invasive arthropod species in Germany based on public data sets and literature, and use it to screen our samples for these taxa.

The results reported in this study derive from two major DNA barcoding campaigns: "Barcoding Fauna Bavarica" (BFB, www.faunabavarica.de, Haszprunar, 2009) and the "German Barcode of Life" project (GBOL, www.bolgermany.de, Geiger, Astrin, et al., 2016), which aim to establish a DNA barcode reference library for all German species. Since their initiation in 2009, DNA barcodes for more than 23,000 metazoan species in Germany have been

assembled. Through the analysis of more than 250,000 specimens, the SNSB – Bavarian State Collection of Zoology (ZSM, see www.barcoding-zsm.de) has made a major contribution to parameterization of the global DNA barcode library maintained in the Barcode of Life Data System (BOLD, www.boldsystems.org, Ratnasingham & Hebert, 2007). Currently, the DNA barcode library created by researchers at the ZSM represents the second-most comprehensive library of any nation, with good coverage for Coleoptera, Diptera, Heteroptera, Hymenoptera, Lepidoptera, Neuroptera, Orthoptera, Araneae and Opiliones, and Myriapoda (see Table 1).

## 2 | MATERIALS AND METHODS

### 2.1 | Sample collection

Nine Malaise traps were deployed around the perimeter of the Bavarian Forest National Park from May to September in 2016 and in 2018 (Figure 1; Table 2). Traps were emptied twice a month, producing 10 samples for each trap year (collection periods designated 1 May to 2 September), for a total of 90 samples annually. In 2016, the original preservative ethanol was changed prior to transportation to the laboratory. Samples were stored in 80% ethanol at room temperature until laboratory analysis. 2016 samples were processed in the laboratory in November 2016. The first 54 samples of 2018 were processed in the laboratory in August 2018, and the latter 36 were processed in November 2018; the original preservative ethanol was processed in December 2018.

### 2.2 | DNA extraction

#### 2.2.1 | Destructive methods

Preservative ethanol was removed, and specimens were transferred to 500-ml PET bottles, dried at 70°C for at least 3 hr and then left at room temperature overnight if necessary, to evaporate off the

**TABLE 1** Major arthropod orders and respective species and specimen numbers represented by DNA barcode sequences from the ZSM

| Order | Number of barcoded individuals | Number of species | Reference |
|---|---|---|---|
| Coleoptera | 15,948 | 3,514 | Hendrich et al. (2015) |
| | 819 | 78 | Raupach, Hannig, Morinière, and Hendrich (2016) |
| | 690 | 47 | Raupach, Hannig, Morinière, and Hendrich (2018) |
| | 13,516 | 2,846 | Rulik et al. (2017) |
| Diptera | 45,040 | 2,453 | Morinière et al. (2019) |
| Ephemeroptera, Plecoptera and Trichoptera | 2,613 | 363 | Morinière et al. (2017) |
| Heteroptera | 1,742 | 457 | Raupach et al. (2014) |
| | 712 | 67 | Havemann et al. (2018) |
| Hymenoptera | 4,118 | 561 | Schmidt, Schmid-Egger, Morinière, Haszprunar, and Hebert (2015) |
| | 4,362 | 1,037 | Schmidt et al. (2017) |
| | 3,695 | 661 | Schmid-Egger et al. (2019) |
| Lepidoptera | 1,395 | 331 | Hausmann, Haszprunar, and Hebert (2011) |
| | 3,467 | 957 | Hausmann, Haszprunar, Segerer, et al. (2011) |
| | 2,130 | 219 | Hausmann et al. (2013) |
| Neuroptera | 237 | 83 | Morinière et al. (2014) |
| Orthoptera | 748 | 122 | Hawlitschek et al. (2017) |
| Araneae and Opiliones | 3,537 | 598 | Astrin et al. (2016) |
| Myriapoda | 320 | 122 | Spelda, Reip, Oliveira Biener, and Melzer (2011) |


**FIGURE 1** Overview of the Malaise trap sample sites in the Bavarian Forest National Park (left). Example image of a Malaise trap setup in the National Park (right)

**TABLE 2** Locations of the nine Malaise traps deployed in this study in 2016 and 2018

| Plot | Location | Latitude (deg.) | Longitude (deg.) | Altitude (m a.s.l.) | In the NPBW? |
|---|---|---|---|---|---|
| Igg35 | Iggensbach | 48.73 | 13.10 | 379 | N |
| Jos21 | Assmann | 48.52 | 13.72 | 364 | N |
| Sal25 | Saldenburg | 48.80 | 13.35 | 505 | N |
| T1_2 | Plattenhausen_1 | 48.92 | 13.40 | 740 | Y |
| T1_34 | Plattenhausen_1 | 48.94 | 13.42 | 819 | Y |
| T1_52 | Plattenhausen_1 | 48.95 | 13.44 | 945 | Y |
| T1_63 | Plattenhausen_1 | 48.96 | 13.45 | 1,287 | Y |
| T3_50 | Scheuereck_3 | 49.10 | 13.32 | 1,182 | Y |
| T4_64 | Lackenberg_4 | 49.10 | 13.28 | 1,137 | Y |

[a]Abbreviation: NPBW, nationalpark bayerischer wald.

residual ethanol. In 2016, dried specimens were ground with a sterilized pestle to homogenize the tissue. Samples from 2018 were homogenized in 500-ml PET bottles with 5–10 sterile steel balls using a FastPrep 96 (MP Biomedicals). Because the specimens were not quantified (e.g. by weighing or counting them) prior to homogenization, a 9:1 mixture of insect lysis buffer and Proteinase K was added in sufficient amounts to cover the ground specimens. Lysis was performed overnight at 56°C. Lysates were then allowed to cool to 20°C and 200-µl aliquots were used for DNA extraction using a Qiagen DNEasy Blood & Tissue Kit (Qiagen) following the manufacturer's instructions.

### 2.2.2 | Nondestructive methods

*DNA extraction from preservative ethanol*

For extraction of DNA from the preservative ethanol, we followed protocols employed by Hajibabaei et al. (2012). This evaporative ethanol technique was performed on five samples (1 May to 1 July) from each of the nine traps in 2018. A 50-ml aliquot of preservative ethanol was taken from each bottle. From this, two 1-ml aliquots were placed into Eppendorf tubes and allowed to dry overnight at 56°C. Fifty microlitres of molecular water was added the next morning, and the tubes were vortexed. Afterwards, DNA extraction was performed on the entire 50-µl sample using the DNeasy Blood and Tissue kit.

For another five samples (trap T3-50B 2018; 2 July, 1 August, 2 August, 1 September, 2 September II) a 50-ml aliquot of ethanol was used for filtration of DNA and tissue residuals using analytical test filter funnels (0.45 µm, Fisher Scientific) equipped with a water jet pump. After ethanol was filtered, the filter funnels were lysed overnight at 56°C. DNA extraction was performed using the DNeasy Blood and Tissue kit following the manufacturer's instructions and eluted into 50 µl of molecular-grade water.

*Semilysis of bulk samples*

Five bulk samples of 2018 (Sal-25, 2 July; T1-02, 2 July; T1-52, 2 July; T1-34, 2 July; and T3-50, 1 July) were used for semilysis and

subsequent DNA extraction. PET bottles (500 ml) were filled with sufficient amounts of lysis mixture (9:1 insect lysis buffer/Proteinase K) and incubated overnight at 56°C. For DNA extraction, 1 ml of the lysate was used following the above-mentioned methods using the DNeasy Blood and Tissue kit. The remaining bulk sample was then dried, and the residual insect lysis buffer was discarded. Samples were then homogenized as described in the Section 2.2.1 above.

## 2.3 | Amplification of the CO1 barcode fragment

From each sample, 5 µl of extracted genomic DNA was used, along with 20 µl of the following mixture: 1.5 µl Mango TAQ (Bioline), 5 µl forward and 5 µl reverse HTS-adapted minibarcode primers of Leray et al. (2013), 6.25 µl $MgCl_2$, 10 µl dNTPs, 25 µl Mango Buffer and 62.5 µl molecular-grade water. DNA extractions from preservative ethanol were amplified using a MyTaq Plant-PCR Kit (Bioline). PCR conditions were as follows: 2 min at 96°C; three cycles of 15 s at 96°C, 30 s at 48°C and 90 s at 65°C; 30 cycles of 15 s at 96°C, 30 s at 55°C and 90 s at 65°C; 10 min at 72°C (see Morinière et al., 2016). Amplification success and fragment lengths (~350 bp) were observed using gel electrophoresis on a 1% agarose gel.

## 2.4 | Purification and next generation sequencing

Amplified DNA was cleaned up by centrifugation of each sample with a 1:10 mixture of 3 m sodium acetate and ice cold 100% ethanol and resuspended in 50 µl molecular-grade water before proceeding. Illumina Nextera XT (Illumina Inc.) indices were ligated to the samples by PCR, and ligation success was confirmed by gel electrophoresis (as described in Morinière et al., 2019). DNA concentrations were measured using a Qubit fluorometer (Life Technologies), and samples were combined into 40-µl pools containing equimolar concentrations of 100 ng each. Pools were loaded into a 1.5% agarose gel, run at 90 V for 45 min, and bands of target amplicons were excised with sterilized razor blades, and purified with a GeneJet Gel Extraction kit (Life Technologies), following the manufacturer's instructions. A final elution volume of 20 µl was used. Sequencing runs were performed on an Illumina MiSeq using V2 chemistry (2 × 250 bp, 500 cycles, 20 million paired-end reads maximum).

## 2.5 | Pre-processing and clustering of sequence data

All FASTQ files generated were combined although they were sequenced on separate runs throughout the study period. Sequence processing was performed with the VSEARCH version 2.4.3 suite (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) and CUTADAPT version 1.14 (Martin, 2011). Because some runs did not yield reverse reads of sufficiently high quality to enable paired-end merging, only forward reads were utilized. Forward primers were removed with CUTADAPT. Quality filtering was with the fastq_filter program of VSEARCH, fastq_maxee 2, with a minimum length of 100 bp. Sequences were dereplicated with derep_fulllength, first at the sample level, and then concatenated into one FASTA file, which was then dereplicated. Chimeric sequences were removed from the FASTA file using uchime_denovo. Remaining sequences were clustered into operational taxonomic units (OTUs) at 97% identity with cluster_size, and an OTU table was created with usearch_global. To reduce probable false positives, a cleaning step was employed which excluded read counts in the OTU table that represented less than 0.01% of the total read count for their respective sample (see Elbrecht & Steinke, 2019).

## 2.6 | Construction of reference databases and sequence identification

### 2.6.1 | BIN-based reference library

All arthropod sequences on BOLD were downloaded (FASTA format, including private and public data) to create a general reference database containing hierarchical taxonomic information and barcode index numbers (BINs). To create this database, downloaded FASTA files were concatenated and imported into GENEIOUS (version 10 Biomatters) (Kearse et al., 2012). To aid the monitoring of species of interest, a broad list of potentially relevant arthropod species was compiled from the following literature sources: Index of Economically Important Lepidoptera (Zhang, 1994), and *Die Forstschädlinge Europas* ("The Forest Pests of Europe") (Pschorn-Walcher & Schwenke, 1982). Of the Index of Economically Important Lepidoptera, 2,684 species names were found on BOLD. Of the Forest Pests of Europe, 294 species names were found on BOLD. About two-thirds (1,962/2,978) of these species had BINs. OTUs were BLASTed (MEGABLAST, default parameters) against the downloaded database. The result was joined to the OTU table in LIBREOFFICE, where the spreadsheet of pest names and BINs was used to cross-check with the BLAST results. All of these BINs and species names available on BOLD were added to a publicly available data set named "Dataset – DS-BWPST Database of Pest Species of Insects in Germany" (data set https://doi.org/10.5883/DS-BWPST).

### 2.6.2 | Pest and invasive species custom reference libraries

Reference sequences for species from the following sources were compiled into a list of 1,017 names: Nature protection warning list of the German Federal Office for Nature Conservation in Bonn ("Erstellung einer Warnliste in Deutschland noch nicht vorkommender invasiver Tiere und Pflanzen") (Rabitsch, Gollasch, Isermann, Starfinger, & Nehring, 2013), terrestrial arthropods only; "Die invasiven gebietsfremden Arten der Unionsliste der Verordnung (EU) Nr.1143/2014 -Erste Fortschreibung

2017" (Nehring and Skowronek); The International Union for Conservation of Nature's Red List of Threatened Species (IUCN, 2019), accessed online, https://www.iucnredlist.org, filter criteria of phylum = Arthropoda, land regions = Europe, Geographical scale = global, Red List Category = Critically Endangered, Endangered, Extinct in the wild, Lower risk/Conservation dependent, near threatened, or vulnerable; the European Plant Protection Global Database (https://gd.eppo.int/country/DE), filter criteria of "Germany"; as well as the following 28 widely known invasive species (with one synonym), if not already listed: *Periplaneta americana* (Linnaeus, 1758), *Harmonia axyridis* (Pallas, 1773), *Stictocephala bisonia* (Kopp and Yonke, 1977), *Anoplophora chinensis* (Forster, 1771), *Corythucha ciliata* (Say, 1832), *Rhagoletis completa* (Cresson, 1929), *Sceliphron curvatum* (Smith, 1870), *Leptinotarsa decemlineata* (Say 1824), *Reticulitermes flavipes* (Kollar, 1837), *Anoplophora glabripennis* (Motschulsky, 1853), *Hulecoeteomyia japonica* (Theobald, 1901), *Aedes japonicus* (Theobald, 1901), *Aedes koreicus* (Edwards, 1917), *Dryocosmus kuriphilus* (Yasumatsu, 1951), *Aproceros leucopoda* (Takeuchi, 1939), *Cacyreus marshalli* (Butler, 1898), *Dreyfusia nordmannianae* (Eckstein, 1890), *Frankliniella occidentalis* (Pergande, 1895), *Leptoglossus occidentalis* (Heidemann, 1910), *Cameraria ohridella* (Deschka and Dimic, 1986), *Cydalima perspectalis* (Walker, 1859), *Monomorium pharaonis* (Linnaeus, 1758), *Hypoponera punctatissima* (Roger, 1859), *Phyllonorycter robiniella* (Clemens, 1859), *Drosophila suzukii* (Matsumura, 1931), *Trialeurodes vaporariorum* (Westwood, 1856), *Diabrotica virgifera* (J.L. LeConte, 1868), *Viteus vitifoliae* (Fitch, 1855) and *Ectobius vittiventris* (Costa, 1847).

Sequences were downloaded using the R (R Core Team, 2019) package *BOLD* (Chamberlain, 2018). Of the 1,004 total species names, 361 were found in BOLD. These were exported as a tab-separated file and processed into FASTA format with Linux command lines. The remaining species were searched for on NCBI GenBank (advanced search, criteria including ["COI" OR "CO1" OR "COXI" OR "COX1"]). Forty-one of the species names were found and downloaded as FASTA files. To combine the sequences from both sources into a single database and BLAST, we used BOLD_NCBI_Merger (Macher, Macher, & Leese, 2017). The highest scoring pair of the top hit (NCBI BLAST+, outfmt 6) for each OTU was imported into LIBREOFFICE, joined with the OTU table, and filtered. A taxonomic neighbour-joining tree was constructed using the BOLD website. All arthropod species and corresponding BINs on the list that were available on BOLD were added to a publicly available data set named "Dataset – DS-BFNWARN Bundesamt für Naturschutz Warnliste, Arthropoden" (data set https://doi.org/10.5883/DS-BFNWARN).

### 2.7 | Biodiversity analysis

As DNA metabarcoding is not quantitative (Krehenwinkel et al., 2017; Piñol, Senar, & Symondson, 2019) we utilized presence–absence data of BINs recovered at ≥97% identity over geographical areas represented by Malaise trap locations to calculate many of the biodiversity metrics. The OTU table indicates which BINs (or higher corresponding taxa) were detected in each collection event. To calculate detection frequencies, all counts in the table greater than zero were set to one. In this way, row sums across the table indicate the number of samples from which a particular taxon was recovered, while column sums indicate the total numbers of taxa recovered from a sample. Presence–absence data for the homogenized samples for all traps from 2016 and 2018 were also analysed together with a data set from the Global Malaise Trap Program (GMTP) downloaded from BOLD, project "GMTPE Germany Malaise 2012" (see Geiger, Moriniere, et al., 2016). Frequencies of BIN detection throughout the growing seasons could then be compared for each of the three years. Bar and line charts were created with GGPLOT2 (Wickham, 2016) or base R.

The presence of BINs in the 2016 and 2018 samples was used to calculate Jaccard distances and dissimilarity matrices for traps inside and outside the National Park, with the R packages VEGAN (Dixon, 2003) and BETAPART (Baselga & Orme, 2012). A Mantel test was performed to compare the study years in terms of their dissimilarities among trap sites, utilizing the R packages GEOSPHERE (Hijmans, Williams, Vennes, & Hijmans, 2017) and ADE4 (Dray & Dufour, 2007). Analysis of similarities (ANOSIM) tests to compare BIN compositions of trap sites inside and outside of the park were performed with the anosim function of VEGAN: *Community Ecology Package* (Oksanen et al., 2010). Additionally, principal component analyses for the 2016 and 2018 taxonomic composition data for each trap site were performed based on seven-level taxonomic identifications of OTUs and their read counts, with the R package AMPVIS2 (Andersen, Kirkegaard, Karst, & Albertsen, 2018), amp_ordinate function, Hellinger transform.

## 3 | RESULTS

### 3.1 | Biodiversity analysis (BOLD BIN-based database)

A total of 19,727 OTUs were produced by the pipeline. Of these, 12,513 matched at ≥73% identity to the database downloaded from BOLD. After filtering for alignment lengths of ≥100 bp, *E*-value of 10e-6 and ≥97% identity to the reference sequences, 5,782 matches remained. The majority of matches belonged to Arthropoda, with the majority of those belonging to Diptera (3,169), Hymenoptera (1,173), Lepidoptera (527) and Coleoptera (411). Table 3 lists total BIN detections broken down by order in 2016 and 2018, and the proportion of BINs which were recovered in both years (percentage overlap). Total read numbers produced per sample are given in Table S1, and rarefaction curves for BINs detected are in Figure S1.

Of the BOLD BIN-based database records to which OTUs matched at ≥97%, roughly half (2,918) had species-level taxonomic classifications in BOLD. The rest of the records to which OTUs

**TABLE 3** Comparison of total BIN detections within Malaise trap surveys in 2016 and 2018. The overlap indicates the number of identical BINs detected in both survey years

| Class | Order | 2016 | 2018 | Overlap (%) |
|---|---|---|---|---|
| Arachnida | Araneae | 67 | 42 | 42 |
| | Mesostigmata | 2 | 3 | 25 |
| | Opiliones | 2 | 4 | 50 |
| | Sarcoptiformes | 2 | 2 | 100 |
| Collembola | Entomobryomorpha | 6 | 6 | 71 |
| | Symphypleona | 4 | 2 | 50 |
| Insecta | Blattodea | 2 | 3 | 67 |
| | Coleoptera | 268 | 234 | 40 |
| | Dermaptera | 3 | 3 | 100 |
| | Diptera | 2,119 | 1,900 | 61 |
| | Ephemeroptera | 2 | 2 | 0 |
| | Hemiptera | 94 | 92 | 46 |
| | Hymenoptera | 731 | 709 | 45 |
| | Lepidoptera | 328 | 351 | 44 |
| | Mecoptera | 3 | 3 | 100 |
| | Neuroptera | 19 | 17 | 44 |
| | Odonata | 0 | 14 | 0 |
| | Orthoptera | 13 | 17 | 50 |
| | Plecoptera | 16 | 10 | 53 |
| | Psocodea | 9 | 9 | 100 |
| | Raphidioptera | 4 | 3 | 75 |
| | Thysanoptera | 1 | 1 | 0 |
| | Trichoptera | 24 | 19 | 59 |
| Malacostraca | Isopoda | 0 | 3 | 0 |
| Gastropoda | Stylommatophora | 1 | 3 | 0 |

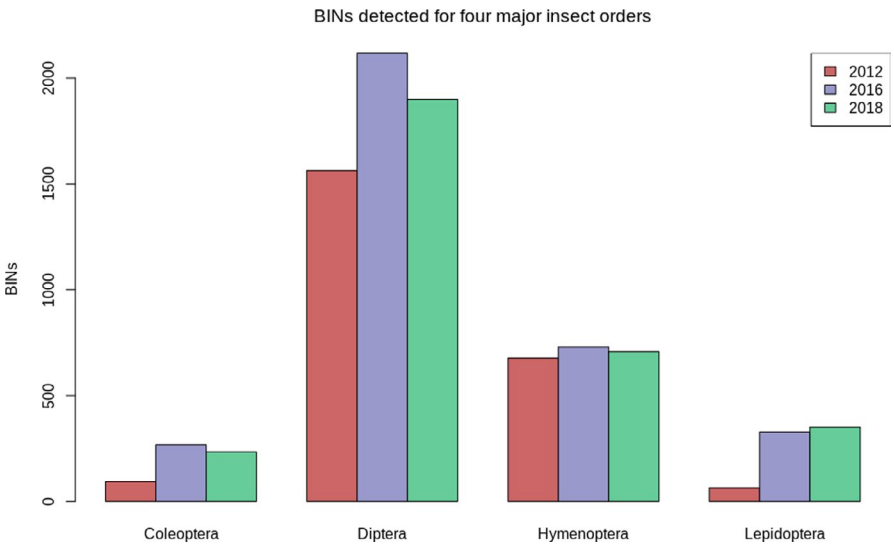


**FIGURE 2** Detected BINs belonging to the orders Coleoptera, Diptera, Hymenoptera, and Lepidoptera within study years 2016 and 2018

matched were classified to lower levels. This is a consequence of the BIN system assigning BINs to sequence clusters algorithmically, whereas taxonomic classifications must be assigned by taxonomists to voucher specimens from which barcode sequences are obtained, a process which requires more time. At the time of writing, an effort is underway to provide taxonomic classifications for all records in BOLD with BINs, with particular emphasis on Diptera and Hymenoptera.

In 2016, 3,430 total BIN matches were detected from all tissue-based (homogenized) samples, and 2,957 in 2018 (counts include BINs belonging to classes Arachnida, Chilopoda, Clitellata, Collembola, Diplopoda, Gastropoda, Insecta and Malacostraca).

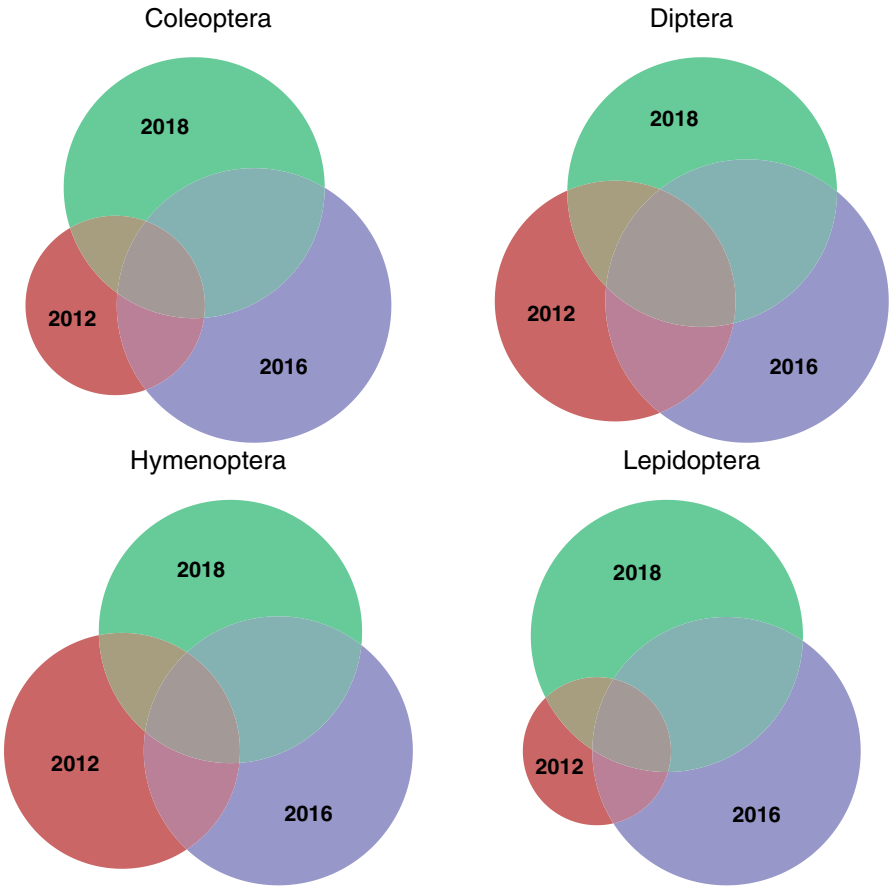


**FIGURE 3** Venn diagrams depict the overlaps in BINs belonging to Coleoptera, Diptera, Hymenoptera, and Lepidoptera which were detected in the 2012 GMTP dataset and the two study years

Figure 2 compares BIN detections within four major insect orders for the two study years and for the 2012 GMTP data set. Figure 3 depicts proportions of shared BINs between the three years for the same orders. BIN recoveries tended to peak in June or early July of each year (Figure 4). Counts of shared BINs between 2016 and 2018 for the four orders are shown in Figure 4 as black lines; for comparison, coloured dotted lines represent counts of individual BINs (presence–absence data for each collection period) for each year. Coloured solid lines take into account how many times BINs were detected in each collection period (total BIN detections). Diptera was the largest order by BIN count. In this order, 2,119 BINs were detected in 2016 (homogenized tissue), 1,900 in 2018 (homogenized tissue) and 2,021 in 2018 (all extraction methods in total).

Based on presence and absence of BINs, a Mantel test revealed a significant correlation between matrices of the mean Jaccard distances by trap sites in 2016 with those of 2018 ($r$ = .4995, $p$ = .005). Based on read abundances, biodiversity analyses of taxa in each trap for 2016 and 2018 are shown as principal component analyses in Figure 5. Malaise traps Igg35, Jos21 and Sal25, which are outside of the National Park, can be observed here clustering the furthest along PC2 in 2016 and PC1 in 2018, compared to all other traps, which are within the park. Additionally, ANOSIM tests showed significant differences between BIN detections in traps inside versus outside of the park in both years (2016 $r$ = .2, $p$ = 2e-04; 2018 $r$ = .239, $p$ = 1e-04).

## 3.2 | Economically important terrestrial arthropods and other species of interest

A total of 83 species names and 118 BINs from the list compiled from *Economically Important Lepidoptera* and *Forest Pests of Europe* matched in the BOLD database BLAST results for all samples (≥97% sequence similarity, *E*-value ≤ 10e-6, highest scoring pairs). We chose two cases of detected species of interest from this list: the noctuid *Lymantria dispar* (Linnaeus, 1758), a common forest pest, and the tortricid *Epinotia tedella* (Clerck, 1759), the presence of which relates to that of a potential regulatory parasitoid species of ichneumonid wasps (*Lissonota dubia* Holmgren, 1856). Total numbers of collection events in which these species of interest were detected in each year are shown in Figures 6 and 7. *Lymatria dispar* is an invasive lepidopteran listed in the *Index of Economically Important Lepidoptera* (Zhang, 1994). Eurasian in origin, it was introduced to the USA in the 19th century. We detected its sequences at 100% match to the database in Malaise trap Jos21 in May and the second collection of July 2016; in 2018 it was found in the same trap but more frequently: in every collection through August, and also in trap T1-34 in the first collection of June (Figure 6). Interestingly, we also observed similar patterns of presence/absence for *E. tedella* and its parasite, *Lissonota dubia* (Figure 7).

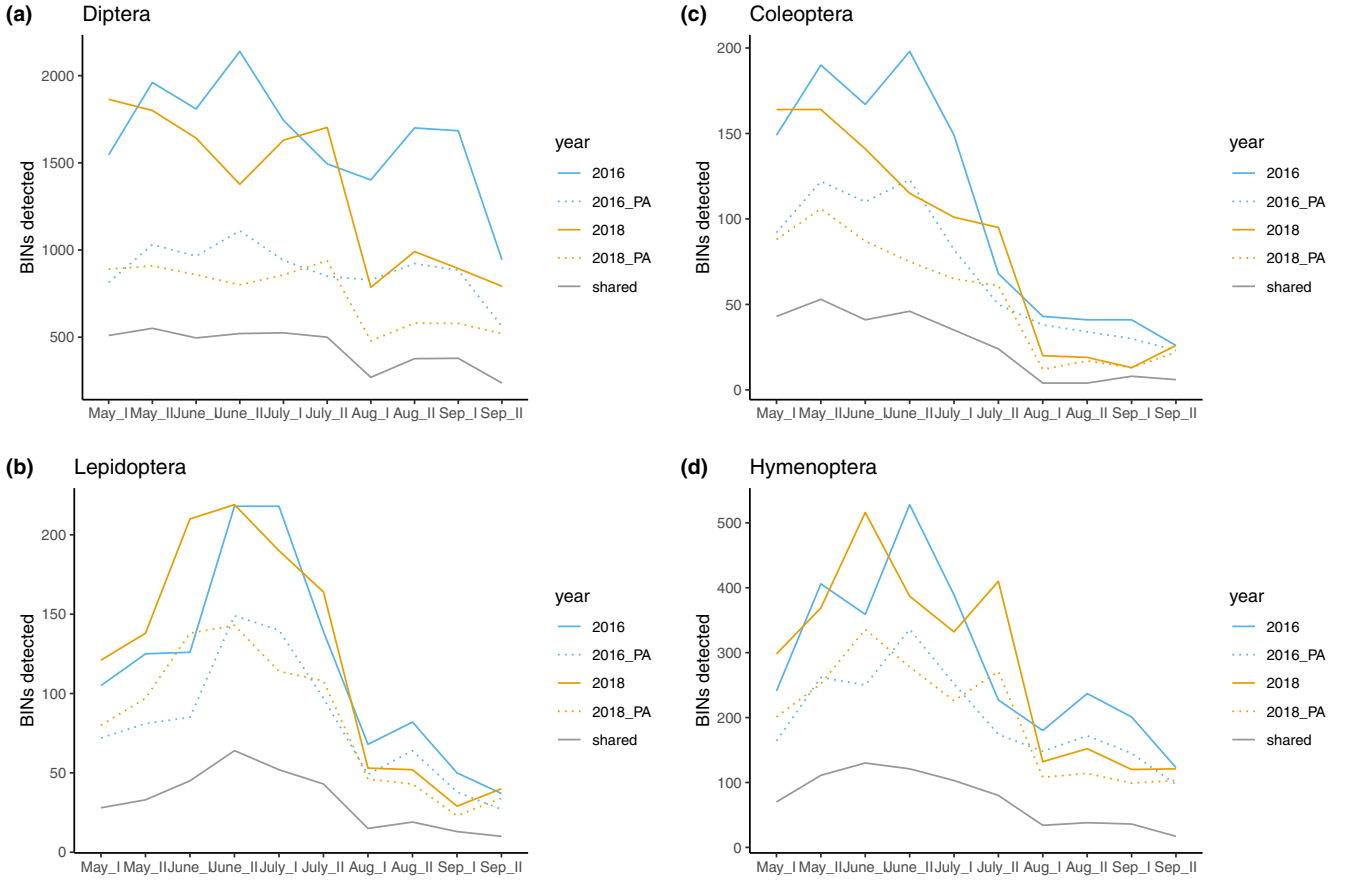


**FIGURE 4** Breakdown of BINs detected in the two survey years by the four Orders (Diptera [a], Coleoptera [b], Lepidoptera [c], and Hymenoptera [d]). Colored solid lines take into account how many times BINs were detected in each collection period. "PA" denotes presence-absence BIN counts. Black lines indicate counts of BINs shared between both years

### 3.3 | Species of interest custom database

Two species from our species of interest database matched to the samples' OTUs by BLAST at ≥97%: the lasiocampid moth *Dendrolimus superans* (Butler, 1877) and the bark beetle *Ips duplicatus* (Sahlberg, 1836) (Table 4), both from the warning list of the German Federal Office for Nature Conservation (Rabitsch et al., 2013). *D. superans* (BOLD: AAB6845) matched at 99.55% identity in Malaise trap sample T1-52 (inside the National Park), collection 1 September 2016.

*Dendrolimus superans* is endemic to Siberia and is a pest of over 20 species of coniferous plants. It has not yet been observed in Germany (Rabitsch et al., 2013). It also shares the BIN BOLD: AAB6845 with *Dendrolimus pini* (Linnaeus, 1758), which is known throughout most of Europe, including Germany. This result illustrates that, because a small custom database was used for this task, consisting of *only* species of interest, hits must be investigated further when the possibility exists that a specimen actually belongs to a closely related species not in this database.

Figure 8 presents a section of a neighbour-joining tree from barcode sequences on BOLD showing representatives of these species clustering together, also with the OTU sequence in question ("Unknown Specimen"). As observed by the BLAST against the general BOLD database, *D. pini* was also detected at a similar identity (99.5%) in the same trap in the BLAST against the BOLD BIN-based database. Therefore, it is probable that the latter was the species which was collected. Further integrative taxonomic study is needed to examine whether *superans* may better be downgraded to subspecies rank or synonymy of *pini*.

*Ips duplicatus* (BOLD: ACD5566) matched at 98.64% identity to the database in Malaise trap T3-50 (inside the National Park), collection 2 July 2018, filtered ethanol sample (Table 4). *I. duplicatus* is endemic to northern Europe, where it is a pest of pine trees (*Pinus* spp.), whereas it is unknown if it additionally poses a threat to biodiversity. The species was unknown in Germany at the time of publication of the warning list, but has recently been spreading southward, through central, eastern and southern Europe (Fiala & Holuša, 2019). Although another congeneric species, *Ips typographus* (Linnaeus, 1758) (BOLD: ACT0826), a keystone pest species in the Bavarian Forest National Park (Müller, Bußler, Goßner, Rettelbach, & Duelli, 2008), was also detected in the same trap at 100% identity, these two species' barcode sequences cluster less closely together, and they do not share a BIN. The present result is therefore likely to be a case of correct molecular identification of *I. duplicatus*, and to represent the first detection of this invasive saproxylic beetle in the National Park.

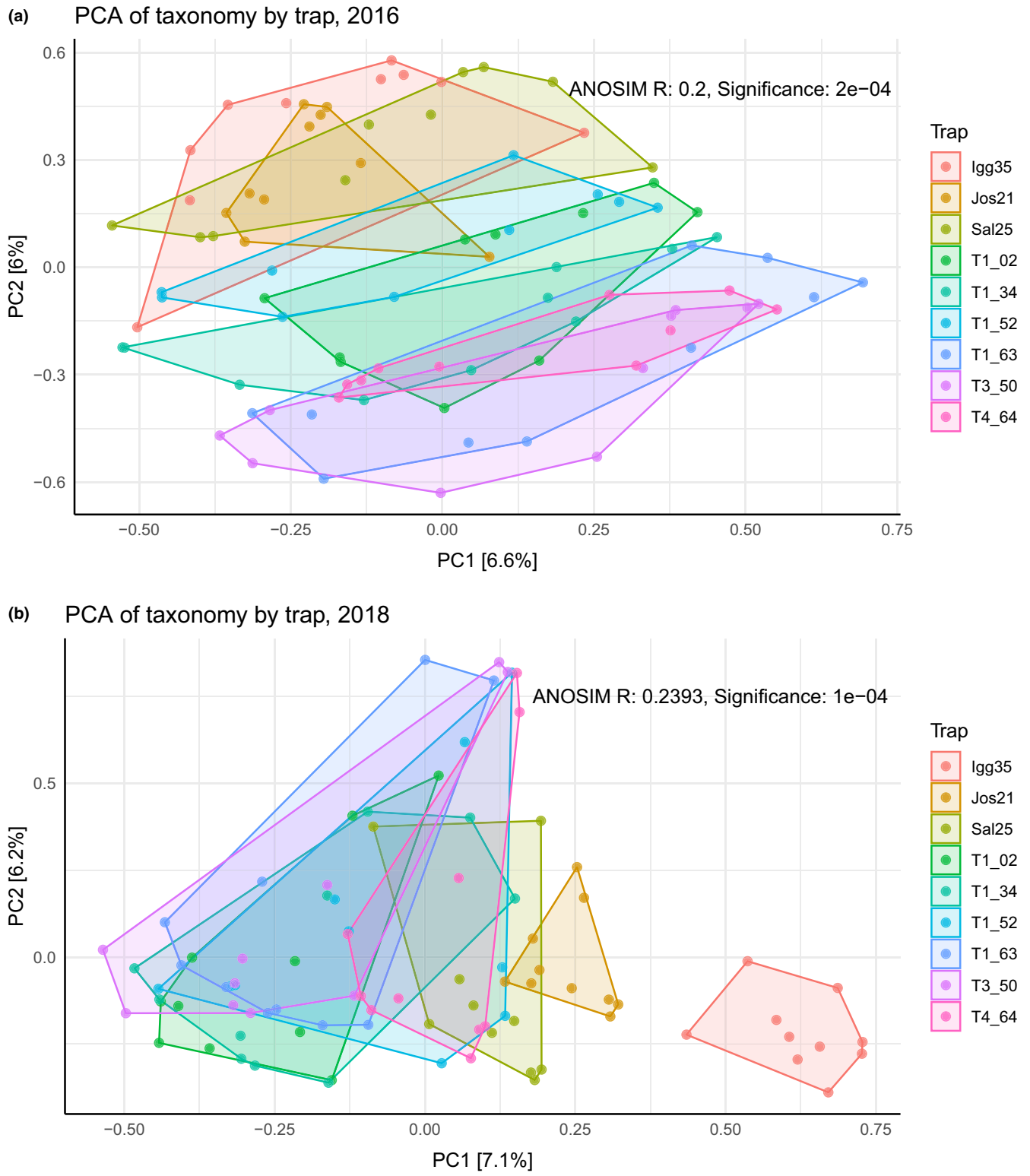


**FIGURE 5** Principal component analyses of read abundances and 7-level taxonomic assignments of OTUs, for survey years 2016 (a) and 2018 (b)

## 4 | DISCUSSION

In the present study, we have been able to accomplish large-scale biomonitoring of the largest national park in Europe, using DNA metabarcoding. By way of presence–absence and read count-based biodiversity analyses, we observed trends in frequencies of observations of taxa throughout two years, utilizing bulk samples from Malaise traps at sites inside and outside of the park, de novo OTU generation and existing reference libraries. Analysing the data from homogenized samples from 2016 and 2018 together with data from a GMTP

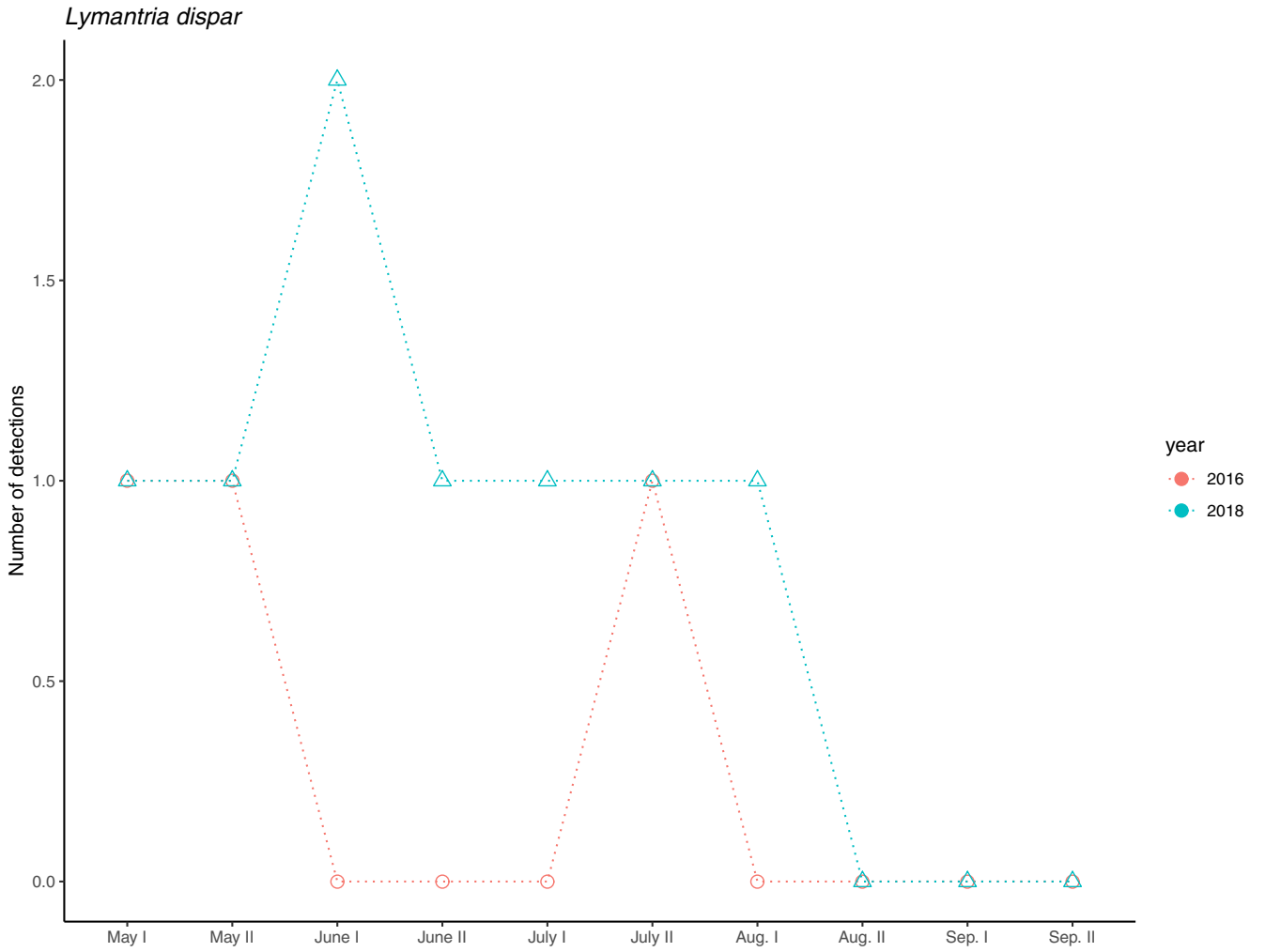


**FIGURE 6** Patterns of detections of *Lymantria dispar* in 2016 and 2018

voucher-based DNA barcoding survey in the Bavarian Forest National Park during 2012 (see Geiger, Moriniere, et al., 2016), we have examined patterns in biodiversity over time. Comparison with the DNA barcoding reference library offers an interesting opportunity to compare local ecosystems with digitized voucher animals over a longer period.

For survey years 2016 and 2018, as well as from the GMTP data, yearly trends in BIN detection overall, as well as on a per-site basis, followed a similar pattern, peaking in June or July, and gradually declining again throughout the remainder of the growing season. Although the samples in the GMTP were screened by morphotype species and DNA barcoded individually, BIN detection for major insect orders was similar to that of both years of the present study (Figure 2). In particular among the dipteran families Cecidomyiidae and Chironomidae, and in the hymenopteran families Braconidae and Ichneumonidae, a BOLD BLAST of our metabarcoding sequences yielded many matches to sequences which had been uploaded to BOLD from voucher specimens collected as part of the GMTP at the very same sites within the Bavarian Forest National Park utilized in the present study. This observation provides support for the exactness and efficacy of metabarcoding for the re-detection of local species.

Detection frequencies of species of interest could also be examined. Same-time detection of host and parasite species was observed, in *Epinotia tedella* and *Lissonota dubia*, in both study years (Figures 6 and 7). These results provide support for the use of metabarcoding as a reliable method for informing phenologies of individual species. It is noteworthy, too, that detection patterns of *Lymantria dispar*, a known pest, potentially suggest an increase in its abundance throughout the National Park. Efforts to track the spread of pest and invasive arthropods should be continued, and metabarcoding represents a viable time- and cost-efficient method of their early detection. We think that implementation of biodiversity data from various sources—such as bulk data on BOLD—will be valuable for ongoing monitoring efforts. Spatial biodiversity analysis revealed a strong correlation of similarity indices of collection sites between the two study years based on presence–absence data of BINs. Furthermore, principal component analysis revealed clustering patterns of abundance-based biodiversities by collection site in each year; and ANOSIM tests showed significant differences in BIN detection between groups of traps located inside and outside of the park (Figure 5). These results provide evidence in support of multi-year repeatability of the methods.

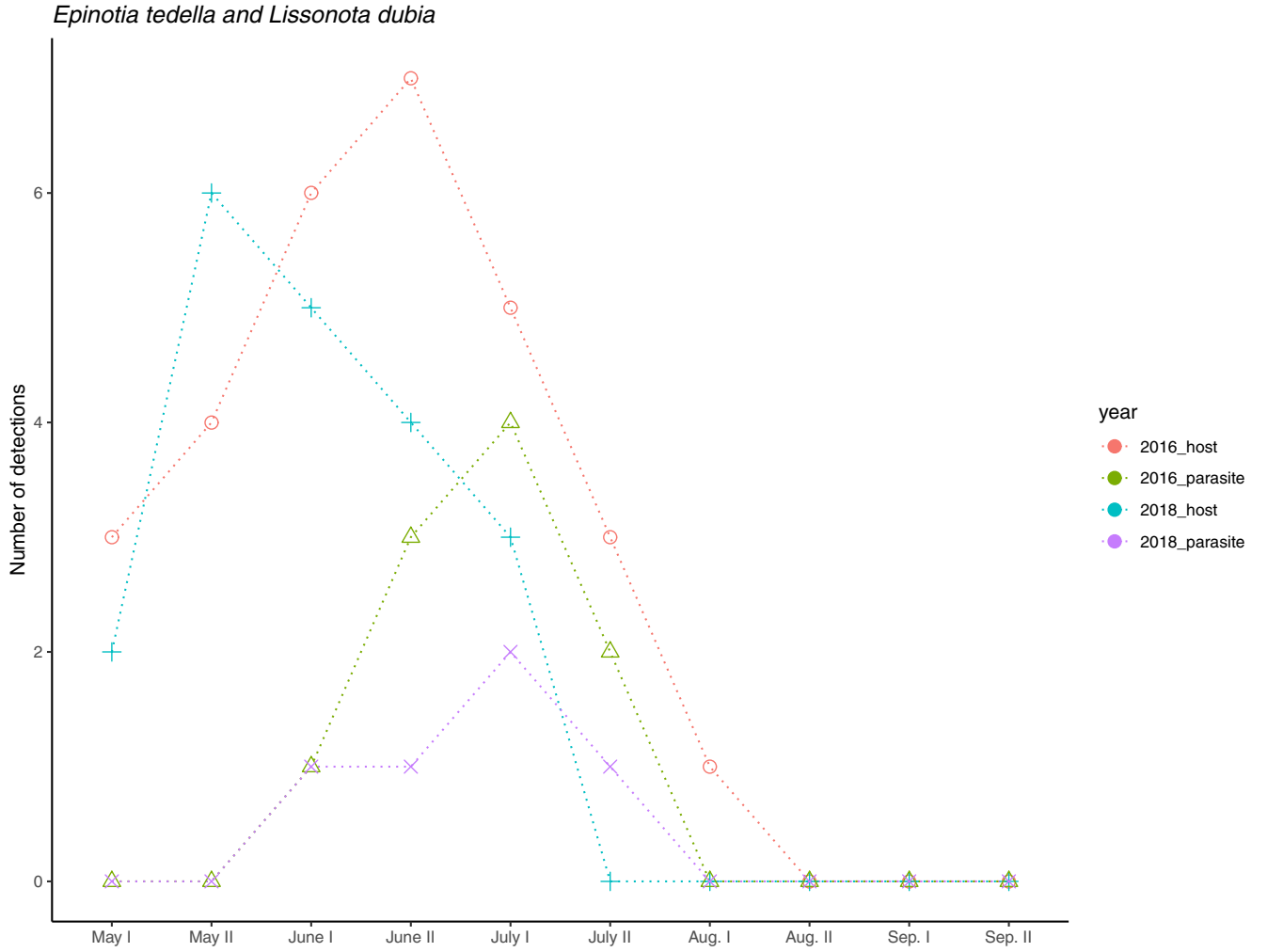


**FIGURE 7** Patterns of detections of *Epinotia tedella* (host) and *Lissonota dubia* (parasite) in both survey years

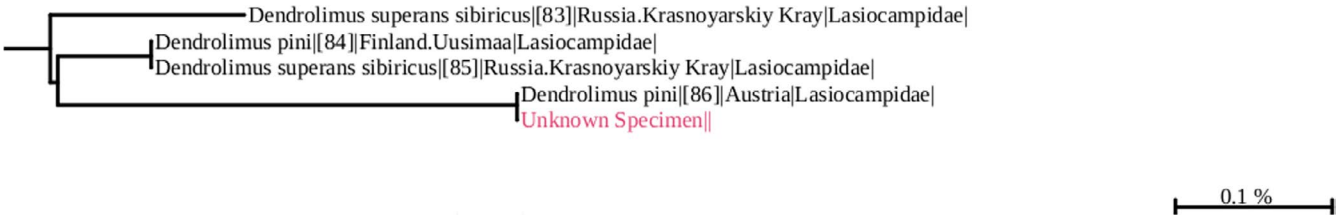


**FIGURE 8** Neighbor Joining tree from BOLD shows barcode sequences of *Dendrolimus superans* and *Dendrolimus pini* specimens clustering together

A comprehensive, well-curated reference sequence library is necessary to realize the full potential of metabarcoding. Barcode databases, most notably NCBI GenBank (Benson et al., 2017) and BOLD (Ratnasingham & Hebert, 2007), now contain millions of reference sequences, especially for the 5′ segment of the mitochondrial cytochrome *c* oxidase I (*COI*) gene (see Porter & Hajibabaei, 2018), designated as the barcode region (Hebert et al., 2003). As OTUs from metabarcoding reads are generally employed for comparison by algorithms such as BLAST, reference sequences should ideally represent intraspecific variation in all taxa. As downloading or comparing against all sources is generally impractical due to their size, the standard approach in metabarcoding is to download only taxa of interest and format them into a local database for comparison by, for example, BLAST. Studies have shown, however, that combining multiple databases provides increased taxonomic coverage and reliability of results (Macher et al., 2017).

In the present study, we have utilized species lists from the literature and publicly available gene banks to create a custom reference database for taxa of potential interest as pests or invasive species, using multiple sources of reference sequences. The

**TABLE 4** Selection from the OTU and BLAST result table showing invasive terrestrial arthropods from the German national institute for nature conservation warning list (2013) detected in samples

| Hit description | Percentage identity | Alignment length | X2016_T1_34B_2LMai | X2016_T1_34B_2LSep | X2016_T1_52B_1LSep | X2016_T1_52B_2LAug | X2016_T1_52B_2LSep | X2018_T3_50B_1LAug | X2018_T3_50_2LAug_filter | X2018_T3_50_2LJuli_filter |
|---|---|---|---|---|---|---|---|---|---|---|
| Dendrolimus_superans BOLD:AAB6845 | 99.548 | 221 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 |
| Ips_duplicatus BOLD:ACD5566 | 98.636 | 220 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |

database, however, could have been even more comprehensive if *COI* reference sequences for more of the species of interest were publicly available, underlining the ongoing need for comprehensive reference libraries for DNA metabarcoding. In conjunction with the application of multiple methods of DNA extraction, this database enabled us to find a match to a warning-list species in our samples (Table 4). Of two potential matches above 97% identity to database sequences, one was a participant in BIN-sharing, clustering together with an endemic species. Therefore, *Ips duplicatus* was the only molecular identification from the warning list. The unambiguous molecular identification of the heavily invasive pest *I. duplicatus* represents a new record of this pest in the Bavarian Forest National Park. Bark beetles of the genus *Ips* are of interest to biologists for the roles they play in the decomposition of pine and spruce trees in forest ecosystems. Although this species was detected in only one sample with one extraction method (filtered ethanol) with low read numbers (11), it nevertheless remained in the OTU table after applying our cleaning steps; and although the possibility of a false positive (e.g. from contamination) cannot be definitively excluded, it may have been a result of traces of this species in the environment, especially in light of its invasive patterns observed recently (Fiala & Holuša, 2019). One possibility is regurgitated gut contents from a predator species in the trap (see Zizka et al., 2019). Detection of this pest may suggest a need for follow-up monitoring with particular attention to this species. If this result is indeed an early detection of a pest species at its invasive front, it may assist in the implementation of timely measures to reduce the risk of damage to the ecosystem. Additionally, the fact that this species was detected exclusively by ethanol filtration provides further support for our recommendation of the use of multiple methods of DNA extraction in conjunction for metabarcoding efforts, whenever possible.

With the rapidly growing demand for large-scale biodiversity data, metabarcoding has gained popularity as the method of choice for any major biomonitoring initiative. Our study shows that the method qualifies as a cost- and time-efficient alternative to traditional approaches. However, despite its apparent advantages, more research is needed to overcome its current limitations in both the laboratory and informatic areas. We encourage further studies towards this aim, to investigate patterns of biodiversity across all varieties and scales of ecosystems and environments, in order to increase the ability of scientists to effectively manage resources and conserve the biodiversity upon which life on Earth depends.

## ACKNOWLEDGEMENTS

We express our sincere gratitude to Olaf Schubert for administrating the collection of the specimens from the Malaise traps in the field. Without his dedicated work this study would not have been possible. We would also like to thank Dr Marina Querejeta Coma for her diligent work on the laboratory portion of this study, as well as Niklas Franzen and Jonas Isemann for their assistance in support of the laboratory work. We are grateful to the team at the Centre for Biodiversity Genomics in Guelph (Ontario, Canada) for their support and help, and particularly to Sujeevan Ratnasingham for developing

the BOLD database (BOLD; www.boldsystems.org) infrastructure and the BIN management tools.

## AUTHOR CONTRIBUTIONS

Obtained funding: G.H., A.H., J.Mü., P.D.N.H. Designed the research: G.H., J.Mo., L.A.H., J.Mü. Analysed the data: L.A.H., J.Mü. Wrote the paper: L.A.H., J.Mo. Contributed additions/corrections to the manuscript: P.D.N.H., L.H., G.H., A.H, S.S., D.D.

## ETHICAL APPROVAL

Specimens were collected by Malaise traps, which were deployed in 2016 and 2018 in the National Park Bavarian Forest. Fieldwork permits were issued by the responsible state environmental offices of Bavaria (*Bayerisches Staatsministerium für Umwelt und Gesundheit, Munich, Germany, project: "Barcoding Fauna Bavarica"; confirmed by the regional governments "Bezirksregierungen"*).

## DATA ACCESSIBILITY

Table generated from OTU table and BLAST results: Dryad https://doi.org/10.5061/dryad.xd2547dcb (Hardulak et al., 2019). BOLD data set "DS-BWPST Database of Pest Species of Insects in Germany" https://doi.org/10.5883/DS-BWPST. BOLD data set "DS-BFNWARN Bundesamt für Naturschutz Warnliste, Arthropoden" https://doi.org/10.5883/DS-BFNWARN.

## ORCID

*Laura A. Hardulak* https://orcid.org/0000-0002-4821-7343
*Jérôme Morinière* https://orcid.org/0000-0001-9167-6409
*Stefan Schmidt* https://orcid.org/0000-0001-5751-8706
*Jörg Müller* https://orcid.org/0000-0002-1409-1586
*Paul D. N. Hebert* https://orcid.org/0000-0002-3081-6700

## REFERENCES

Andersen, K. S. S., Kirkegaard, R. H., Karst, S. M., & Albertsen, M. (2018). ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *BioRxiv*, 299537.

Armstrong, K. F., & Ball, S. L. (2005). DNA barcodes for biosecurity: Invasive species identification. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1462), 1813–1823. https://doi.org/10.1098/rstb.2005.1713

Astrin, J. J., Höfer, H., Spelda, J., Holstein, J., Bayer, S., Hendrich, L., … Muster, C. (2016). Towards a DNA barcode reference database for spiders and harvestmen of Germany. *PLoS One*, *11*(9), e0162624 (24 pp plus supplements).

Aylagas, E., Borja, Á., & Rodríguez-Ezpeleta, N. (2014). Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). *PLoS One*, *9*(3), e90529. https://doi.org/10.1371/journal.pone.0090529

Baselga, A., & Orme, C. D. L. (2012). betapart: An R package for the study of beta diversity. *Methods in Ecology and Evolution*, *3*(5), 808–812. https://doi.org/10.1111/j.2041-210x.2012.00224.x

Bellard, C., Cassey, P., & Blackburn, T. M. (2016). Alien species as a driver of recent extinctions. *Biology Letters*, *12*(2), 20150623. https://doi.org/10.1098/rsbl.2015.0623

Beng, K. C., Tomlinson, K. W., Shen, X. H., Surget-Groba, Y., Hughes, A. C., Corlett, R. T., & Slik, J. 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), 1–13. https://doi.org/10.1038/srep24965

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *45*(D1), D37–D42. https://doi.org/10.1093/nar/gkw1070

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., … De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in ecology & evolution*, *29*(6), 358–367.

Borrell, Y. J., Miralles, L., Do Huu, H., Mohammed-Geba, K., & Garcia-Vazquez, E. (2017). DNA in a bottle—Rapid metabarcoding survey for early alerts of invasive species in ports. *PLoS One*, *12*(9), e0183347. https://doi.org/10.1371/journal.pone.0183347

Ceballos, G., & Ehrlich, P. R. (2018). The misunderstood sixth mass extinction. *Science*, *360*(6393), 1080–1081.

Chamberlain, S. (2018). *Bold: Interface to Bold systems API*. R package version 0.8.6. Retrieved from https://CRAN.R-project.org/package=bold

Chivian, E., & Bernstein, A. (Eds.) (2008). *Sustaining life: How human health depends on biodiversity*. Oxford, UK: Oxford University Press.

Cock, M. J. W. (2003). *Biosecurity and forests: An introduction – With particular emphasis on forest pests*. FAO Forest Health and Biosecurity Working Paper FBS/2E, 2003.

Comtet, T., Sandionigi, A., Viard, F., & Casiraghi, M. (2015). DNA (meta) barcoding of biological invasions: A powerful tool to elucidate invasion processes and help managing aliens. *Biological Invasions*, *17*(3), 905–922. https://doi.org/10.1007/s10530-015-0854-y

Corlett, R. T. (2017). A bigger toolbox: Biotechnology in biodiversity conservation. *Trends in Biotechnology*, *35*(1), 55–65. https://doi.org/10.1016/j.tibtech.2016.06.009

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930. https://doi.org/10.1658/1100-9233(2003)014[0927:vaporf]2.0.co;2

Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20.

Ehrenfeld, J. G. (2010). Ecosystem consequences of biological invasions. *Annual review of ecology, evolution, and systematics*, *41*, 59–80.

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). 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/10.1002/ece3.3192

Elbrecht, V., & Steinke, D. (2019). Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. *Freshwater Biology*, *64*(2), 380–387.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., … Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: Potential for studying past and present ecosystems. *Molecular Ecology*, *21*(8), 1821–1833. https://doi.org/10.1111/j.1365-294X.2012.05537.x

Fiala, T., & Holuša, J. (2019). Occurrence of the invasive bark beetle *Phloeosinus aubei* on common Juniper trees in the Czech Republic. *Forests*, *10*(1), 12. https://doi.org/10.3390/f10010012

Geiger, M. F., Astrin, J. J., Borsch, T., Burkhardt, U., Grobe, P., Hand, R., … Wägele, W. (2016). 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*, *59*(9), 661–670. https://doi.org/10.1139/gen-2015-0185

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*, e10671 (22 pp plus supplements).

Geijzendorffer, I. R., Regan, E. C., Pereira, H. M., Brotons, L., Brummitt, N., Gavish, Y., … Walters, M. (2016). Bridging the gap between biodiversity data and policy reporting needs: An essential biodiversity variables perspective. *Journal of Applied Ecology*, *53*(5), 1341–1350. https://doi.org/10.1111/1365-2664.12417

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, *6*(4), e17497. https://doi.org/10.1371/journal.pone.0017497

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*, 28. https://doi.org/10.1186/1472-6785-12-28

Hardulak, L. A., Morinière, J., Hausmann, A., Hendrich, L., Schmidt, S., Doczkal, D., … Haszprunar, G. (2019). DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.xd2547dcb

Haszprunar, G. (2009). Barcoding Fauna Bavarica–eine Chance für die Entomologie. *Nachrichtenblatt Der Bayerischen Entomologen*, *58*(1/2), 45–47.

Hausmann, A., Godfray, H. C. J., Huemer, P., Mutanen, M., Rougerie, R., van Nieukerken, E. J., … Hebert, P. D. N. (2013). Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. *PLoS One*, *8*(12), e84518 (11 pp).

Hausmann, A., Haszprunar, G., & Hebert, P. D. N. (2011). DNA barcoding the geometrid fauna of Bavaria (Lepidoptera): Successes, surprises, and questions. *PLoS One*, *6*(2), e17134 (9 pp).

Hausmann, A., Haszprunar, G., Segerer, A. H., Speidel, W., Behounek, G., & Hebert, P. D. N. (2011). Now DNA-barcoded: The butterflies and larger moths of Germany. *Spixiana*, *34*(1), 47–58.

Havemann, N., Gossner, M. M., Hendrich, L., Morinière, J., Niedringhaus, R., Schäfer, P., & Raupach, M. J. (2018). From water striders to water bugs: the molecular diversity of aquatic Heteroptera (Gerromorpha, Nepomorpha) of Germany based on DNA barcodes. *PeerJ*, *6*, e4577 (30 pp).

Hawlitschek, O., Morinière, J., Lehmann, G. U. C., Lehmann, A. W., Kropf, M., Dunz, 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

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(Suppl_1), S96–S99.

Hendrich, L., Morinière, J., Haszprunar, G., Hebert, P. D., 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

Hijmans, R. J., Williams, E., Vennes, C., & Hijmans, M. R. J. (2017). *Package 'geosphere'*. Spherical Trigonometry. R package version 1.5-7, 2017.

Honrado, J. P., Pereira, H. M., & Guisan, A. (2016). Fostering integration between biodiversity monitoring and modelling. *Journal of Applied Ecology*, *53*(5), 1299–1304. https://doi.org/10.1111/1365-2664.12777

IUCN. (2019). *The IUCN red list of threatened species*. Version 2019–1. Retrieved from http://www.iucnredlist.org

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … 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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … 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.

Keller, R. P., Geist, J., Jeschke, J. M., & Kühn, I. (2011). Invasive species in Europe: Ecology, status, and policy. *Environmental Sciences Europe*, *23*(1), 23 (17 pp).

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*(1), 17668. https://doi.org/10.1038/s41598-017-17333-x

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), 2076–2081. https://doi.org/10.1073/pnas.1424997112

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*, 34. https://doi.org/10.1186/1742-9994-10-34

Lindenmayer, D. B., Gibbons, P., Bourke, M., Burgman, M., Dickman, C. R., Ferrier, S., … Zerger, A. (2012). Improving biodiversity monitoring. *Austral Ecology*, *37*(3), 285–294. https://doi.org/10.1111/j.1442-9993.2011.02314.x

Macher, J.-N., Macher, T.-H., & Leese, F. (2017). Combining NCBI and BOLD databases for OTU assignment in metabarcoding and metagenomic datasets: The BOLD_NCBI _Merger. *Metabarcoding and Metagenomics*, *1*, e22262. https://doi.org/10.3897/mbmg.1.22262

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Meyerson, L. A., & Reaser, J. K. (2002). A unified definition of biosecurity. *Science*, *295*(5552), 44. https://doi.org/10.1126/science.295.5552.44a

Morinière, J., Balke, M., Doczkal, D., Geiger, M. F., Hardulak, L. A., Haszprunar, G., … 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*, *9*(4), 900–928. https://doi.org/10.1111/1755-0998.13022

Morinière, J., Cancian de Araujo, B., Lam, A. W., Hausmann, A., Balke, M., Schmidt, 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., Hendrich, L., Balke, M., Beermann, A. J., König, T., Hess, M., … Haszprunar, G. (2017). A DNA barcode library for Germany′s mayflies, stoneflies and caddisflies (Ephemeroptera, Plecoptera and Trichoptera). *Molecular Ecology Resources*, *17*(6), 1293–1307. https://doi.org/10.1111/1755-0998.12683

Morinière, J., Hendrich, L., Hausmann, A., Hebert, P., Haszprunar, G., & Gruppe, A. (2014). Barcoding Fauna Bavarica: 78% of the Neuropterida fauna barcoded! *PLoS One*, *9*(10), e109719 (8 pp plus supplements).

Müller, J., Bußler, H., Goßner, M., Rettelbach, T., & Duelli, P. (2008). The European spruce bark beetle *Ips typographus* in a national park: From pest to keystone species. *Biodiversity and Conservation*, *17*(12), 2979–3001. https://doi.org/10.1007/s10531-008-9409-1

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'hara, R. B., Simpson, G. L., … Wagner, H. (2010). *Vegan: community ecology package*. R package version 1.17-4. Retrieved from http://CRAN.R-project.org/package=vegan

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture

determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2), 407–419.

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338. https://doi.org/10.1111/mec.14478

Pschorn-Walcher, H., & Schwenke, W. (1982). *Die Forstschädlinge Europas. 3 Band – Schmetterlinge*. Hamburg, Germany: Parey.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rabitsch, W., Gollasch, S., Isermann, M., Starfinger, U., & Nehring, S. (2013). "Erstellung einer Warnliste in Deutschland noch nicht vorkommender invasiver Tiere und Pflanzen" Bundesamt für Naturschutz. *BfN-Skripten*, *331*, 126–131.

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.

Raupach, M. J., Hannig, K., Morinière, J., & Hendrich, L. (2016). A DNA barcode library for ground beetles (Insecta, Coleoptera, Carabidae) of Germany: The genus *Bembidion* Latreille, 1802 and allied taxa. *ZooKeys*, *592*, 121–141. https://doi.org/10.3897/zookeys.592.8316

Raupach, M. J., Hannig, K., Morinière, J., & Hendrich, L. (2018). A DNA barcode library for ground beetles of Germany: The genus Amara Bonelli, 1810 (Insecta, Coleoptera, Carabidae). *ZooKeys*, *759*, 57–80. https://doi.org/10.3897/zookeys.759.24129

Raupach, M. J., Hendrich, L., Küchler, S. M., Deister, F., Morinière, J., & Gossner, M. M. (2014). Building-up of a DNA barcode library for true bugs (Insecta: Hemiptera: Heteroptera) of Germany reveals taxonomic uncertainties and surprises. *PLoS One*, *9*(9), e106940 (13 pp).

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

Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., … Fritzlar, F. (2017). Using taxonomic consistency with semiautomated data preprocessing for high quality DNA barcodes. *Methods in Ecology and Evolution*, *8*(12), 1878–1887.

Sala, O. E., Chapin, F. S., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., … Leemans, R. (2000). Global biodiversity scenarios for the year 2100. *Science*, *287*(5459), 1770–1774.

Schmeller, D. S., Julliard, R., Bellingham, P. J., Böhm, M., Brummitt, N., Chiarucci, A., … Belnap, J. (2015). Towards a global terrestrial species monitoring program. *Journal for Nature Conservation*, *25*, 51–57. https://doi.org/10.1016/j.jnc.2015.03.003

Schmid-Egger, C., Straka, J., Ljubomirov, T., Blagoev, G. A., Morinière, J., & Schmidt, S. (2019). DNA barcodes identify 99 per cent of apoid wasp species (Hymenoptera: Ampulicidae, Crabronidae, Sphecidae) from the Western Palearctic. *Molecular Ecology Resources*, *19*(2), 476–484. https://doi.org/10.1111/1755-0998.12963

Schmidt, S., Schmid-Egger, C., Morinière, J., Haszprunar, G., & Hebert, P. D. N. (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.

Schmidt, S., Taeger, A., Morinière, J., Liston, A., Blank, S. M., Kramp, K., … Stahlhut, J. (2017). Identification of sawflies and horntails (Hymenoptera, 'Symphyta') through DNA barcodes: Successes and caveats. *Molecular Ecology Resources*, *17*(4), 670–685. https://doi.org/10.1111/1755-0998.12614

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

Silva-Santos, R., Ramirez, J. L., Galetti, P. M. Jr, & Freitas, P. D. (2018). Molecular evidences of a hidden complex scenario in *Leporinus* cf. *friderici*. *Frontiers in Genetics*, *9*, 47. https://doi.org/10.3389/fgene.2018.00047

Spelda, J., Reip, H. S., Oliveira Biener, U., & Melzer, R. R. (2011). Barcoding Fauna Bavarica: Myriapoda – a contribution to DNA sequence-based identifications of centipedes and millipedes (Chilopoda, Diplopoda). *ZooKeys*, *115*, 123–139. https://doi.org/10.3897/zookeys.156.2176

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

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Wickham, H. (2016). *ggplot2*. Cham, Switzerland: Springer.

Williamson, M. (1996). *Biological invasions*. London, UK: Chapman & Hall.

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/British Ecological Society*, *3*(4), 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zhang, B. C. H. (1994). *Index of economically important Lepidoptera*. Wallingford, UK: CAB International.

Zizka, V. M., Leese, F., Peinert, B., & Geiger, M. F. (2019). DNA metabarcoding from sample fixative as a quick and voucher-preserving biodiversity assessment method. *Genome*, *62*(3), 122–136. https://doi.org/10.1139/gen-2018-0048

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Hardulak LA, Morinière J, Hausmann A, et al. DNA metabarcoding for biodiversity monitoring in a national park: Screening for invasive and pest species. *Mol Ecol Resour*. 2020;00:1–16. https://doi.org/10.1111/1755-0998.13212

## Publication II - A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based Biomonitoring

Citation: 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., and 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.

Received: 19 December 2018 | Revised: 22 March 2019 | Accepted: 25 March 2019

DOI: 10.1111/1755-0998.13022

Check for updates

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

# A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring

Jérôme Morinière[1] | Michael Balke[1] | Dieter Doczkal[1] | Matthias F. Geiger[2] | Laura A. Hardulak[1] | Gerhard Haszprunar[1] | Axel Hausmann[1] | Lars Hendrich[1] | Ledis Regalado[1] | Björn Rulik[2] | Stefan Schmidt[1] | Johann-Wolfgang Wägele[2] | Paul D. N. Hebert[3]

[1]SNSB-Zoologische Staatssammlung, München, Germany

[2]Zoological Research Museum Alexander Koenig - Leibniz Institute for Animal Biodiversity, Bonn, Germany

[3]Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada

**Correspondence**
Jérôme Morinière and Dieter Doczkal, SNSB-Zoologische Staatssammlung, München, Germany.
Emails: moriniere@snsb.de; doczkal@snsb.de

**Funding information**
Bavarian State Ministry of Science and the Arts; German Federal Ministry of Education and Research, Grant/Award Number: BMBF FKZ 01LI1101 and BMBF FKZ 01LI1501

## Abstract

This study summarizes results of a DNA barcoding campaign on German Diptera, involving analysis of 45,040 specimens. The resultant DNA barcode library includes records for 2,453 named species comprising a total of 5,200 barcode index numbers (BINs), including 2,700 *COI* haplotype clusters without species-level assignment, so called "dark taxa." Overall, 88 out of 117 families (75%) recorded from Germany were covered, representing more than 50% of the 9,544 known species of German Diptera. Until now, most of these families, especially the most diverse, have been taxonomically inaccessible. By contrast, within a few years this study provided an intermediate taxonomic system for half of the German Dipteran fauna, which will provide a useful foundation for subsequent detailed, integrative taxonomic studies. Using DNA extracts derived from bulk collections made by Malaise traps, we further demonstrate that species delineation using BINs and operational taxonomic units (OTUs) constitutes an effective method for biodiversity studies using DNA metabarcoding. As the reference libraries continue to grow, and gaps in the species catalogue are filled, BIN lists assembled by metabarcoding will provide greater taxonomic resolution. The present study has three main goals: (a) to provide a DNA barcode library for 5,200 BINs of Diptera; (b) to demonstrate, based on the example of bulk extractions from a Malaise trap experiment, that DNA barcode clusters, labelled with globally unique identifiers (such as OTUs and/or BINs), provide a pragmatic, accurate solution to the "taxonomic impediment"; and (c) to demonstrate that interim names based on BINs and OTUs obtained through metabarcoding provide an effective method for studies on species-rich groups that are usually neglected in biodiversity research projects because of their unresolved taxonomy.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. *Molecular Ecology Resources* Published by John Wiley & Sons Ltd

**KEYWORDS**
barcode library, biodiversity monitoring, *CO1*, cryptic diversity, Diptera, DNA barcoding, Germany, metabarcoding, mitochondrial DNA

## 1 | INTRODUCTION

Recent evidence for major declines in insect populations has provoked intense public concern. Detailed research on economically important groups, such as pollinators, have linked declines in wild bees to pesticide contamination, climate change, habitat fragmentation and degradation (Potts et al., 2010; Vanbergen & the Insect Pollinators Initiative, 2013). Other studies using mass collecting methods suggest the declines may be general, as evidenced by reductions in the biomass of flying insects by 75% over a few decades (Hallmann et al., 2017; Sorg, Schwan, Stenmans, & Müller, 2013) or even within a few years (Lister & Garcia, 2018). However, the evidence for general declines has failed to ascertain if impacts span all insect groups and all size ranges. The failure to track the status of individual lineages reflects the fact that despite advances in taxonomic practices (e.g., integrative taxonomy), our knowledge of most insect species is limited (Brix, Leese, Riehl, & Kihara, 2015; Cruaud, Rasplus, Rodriguez, & Cruaud, 2017; Pante, Schoelinck, & Puillandre, 2014; Riedel, Sagata, Suhardjono, Tänzler, & Balke, 2013; Wheeler, Raven, & Wilson, 2004). Even in Germany, a country with more than 250 years of taxonomic and faunistic research activity, many groups remain poorly known. This gap, which hampers ecological baseline research, is particularly serious for the two hyperdiverse insect orders, the Diptera and Hymenoptera (Geiger, Moriniere, et al., 2016; Klausnitzer, 2006). With at least 9,500 (Schumann, Bährmann, & Stark, 1999; Schumann, Doczkal, & Ziegler, 2011) and 9,600 (Dathe & Blank, 2004) recorded species in Germany, respectively, these two groups comprise over half of its insect alpha diversity (Völkl, Blick, Kornacker, & Martens, 2004). Moreover, it is likely that the true diversity of these two groups is seriously underestimated, a conclusion reinforced by the extraordinarily high numbers of DNA barcode clusters retrieved by simultaneous analysis of arthropods using high-throughput sequencing (HTS; metabarcoding) from insect collections at single monitoring sites (Morinière et al., 2016). As only about 1,000 (Santos, Sampronha, & Santos, 2017) new species of Diptera are described each year from the million or more species awaiting description, the taxonomic impediment in this group will not be resolved without the adoption of new approaches, such as modern molecular genetic methods and integrative taxonomy (Fujita, Leache, Burbrink, McGuire, & Moritz, 2012; Padial, Miralles, Riva, & Vences, 2010; Schlick-Steiner, Arthofer, & Steiner, 2014; Schlick-Steiner et al., 2010).

The known dipteran fauna of Germany includes roughly half of the almost 20,000 species recorded for Europe (as defined in Fauna Europaea, https://fauna-eu.org/; Pape, 2009). Although this is the highest number of Diptera species recorded from any European country, the inventory is certainly very incomplete. A recent checklist for the Empidoidea of Germany (Meyer & Stark, 2015) added 123 species new to Germany, an increase of 12.5%, Jaschhof (2009) added 34 species of Lestremiinae, an increase of 24.3%, and the collecting efforts for different barcoding campaigns resulted in more than 100 species from various families new to Germany among the identified material (Reimann & Rulik, 2015; Heller & Rulik, 2016; B. Rulik unpublished, D. Doczkal unpublished), with many more expected among the unidentified material. Rapid progress in inventorying is hampered by a lack of experts, also known as the taxonomic impediment (de Carvalho et al., 2007). For example, the German Dipterologist's working group (http://www.ak-diptera.de/index.htm, Accessed 18 December 2018) shows that experts were lacking for one-third of the dipteran families, and that most other families had just one or two experts, often voluntary (i.e., unpaid) taxonomists (in the sense defined by Fontaine et al., 2012). A few families such as the Culicidae (https://mueckenatlas.com/), the Asilidae (Wolff, Gebel, & Geller-Grimm, 2018) and the Syrphidae (Ssymank, Doczkal, Rennwald, & Dziock, 2011) are fairly well explored, but several of the species-richest families (e.g., Cecidomyiidae, Ceratopogonidae, Phoridae, Chloropidae, Sphaeroceridae, Anthomyiidae) have received little attention. Malaise traps are widely used as method of choice to collect arthropods and especially flying insects for biodiversity assessments in terrestrial ecosystems, with Diptera being among the most commonly caught taxa (Doczkal, 2017; Hallmann et al., 2017; Hebert et al., 2016; Karlsson, Pape, Johansson, Liljeblad, & Ronquist, 2005; Matthews & Matthews, 1971; Ssymank et al., 2018). The analysis of specimens from two Malaise traps deployed for a single summer in Germany within the Global Malaise Trap Program (GMTP; http://biodiversitygenomics.net/projects/gmp/) revealed similar trends; here Diptera was the most diverse order being represented by 2,500 species, slightly more than half of all the species that were collected and 70.3% of all individuals (26,189) that were analysed (Geiger, Moriniere, et al., 2016).

Taxonomists working on Diptera have long been well aware of the immense number of undescribed species (Bickel et al., 2009) with estimates of global Diptera species diversity ranging from 400,000 to 800,000 species compared with ~160,000 named species (Borkent et al., 2018; Pape, Blagoderov, & Mostovski, 2011). Hebert et al. (2016), applying DNA barcoding to Canadian insects, proposed that the actual number of species could be much higher, suggesting the possible presence of 1.8 million species in just one family, the Cecidomyiidae (gall midges) alone. Although this estimate may be too high, it is very likely that this single family includes more species than are currently described for the order.

At a time when hundreds or possibly thousands of species become extinct each year (Chivian & Bernstein, 2008), a comprehensive species inventory based on accurately identified specimens

represents the foundation for all conservation and biodiversity initiatives. However, the inventory of biodiversity cannot be completed through morphological approaches alone. Both the speed and costs associated with sequence characterization of a standardized DNA fragment can be improved using DNA barcoding. Usually DNA barcoding studies provide a basis for establishing the reference sequence libraries required to identify specimens of known species (Gwiazdowski, Foottit, Maw, & Hebert, 2015; Hebert, Cywinska, Ball, & Dewaard, 2003). Herein we additionally show that it is also an efficient method for registering unknown and taxonomically challenging species—so called "dark taxa" (Page, 2016). Sequenced taxa can subsequently be associated with established binomens by taxonomic specialists using a reverse taxonomy approach, based on accurately identified museum specimens (ideally type specimens) and expert knowledge. During this process, specimens that belong to unnamed molecular character-based units (operational taxonomic units [OTUs] or barcode index numbers [BINs]) will either be referenced to known species or they may represent overlooked species that are new to science (Geiger, Moriniere, et al., 2016). A curated and comprehensive DNA barcode reference library enables fast and reliable species identifications in those many cases where time, personnel and taxonomic expertise are limited. Furthermore, such a library also supports large-scale biodiversity monitoring that relies upon metabarcoding bulk samples (Hajibabaei, Shokralla, Zhou, Singer, & Baird, 2011; Hajibabaei, Spall, Shokralla, & Konynenburg, 2012; Shokralla, Spall, Gibson, & Hajibabaei, 2012), like those obtained from Malaise traps (Gibson et al., 2014; Leray & Knowlton, 2015; Morinière et al., 2016; Yu et al., 2012).

The results reported in this study derive from two major DNA barcoding campaigns: "Barcoding Fauna Bavarica" (BFB, www.faunabavarica.de; Haszprunar, 2009) and the "German Barcode of Life" project (GBOL, www.bolgermany.de; Geiger, Astrin, et al., 2016). Since 2009, DNA barcodes from over 23,000 German species of Metazoa have been assembled, reflecting the analysis of nearly 250,000 specimens that are curated at the SNSB-Zoologische Staatssammlung München (ZSM, see www.barcoding-zsm.de) and ~180,000 specimens curated at the Zoologisches Forschungsmuseum Alexander Koenig Bonn (ZFMK). These records represent a major contribution to the global DNA barcode library that is maintained in the Barcode of Life Data System (BOLD, www.boldsystems.org; Ratnasingham & Hebert, 2007). Currently, the DNA barcode library created by the ZSM researchers represents the second-most comprehensive library of any nation. Previous studies have reported on barcoding results for Coleoptera (Hendrich et al., 2015; Raupach, Hannig, Moriniere, & Hendrich, 2016; Raupach, Hannig, Morinière, & Hendrich, 2018; Rulik et al., 2017), Ephemeroptera, Plecoptera and Trichoptera (Morinière et al., 2017), Heteroptera (Havemann et al., 2018; Raupach et al., 2014), Hymenoptera (Schmid-Egger et al., 2019; Schmidt, Schmid-Egger, Morinière, Haszprunar, & Hebert, 2015; Schmidt et al., 2017), Lepidoptera (Hausmann, Haszprunar, & Hebert, 2011; Hausmann, Haszprunar, Segerer, et al., 2011), Neuroptera (Morinière et al., 2014), Orthoptera (Hawlitschek et al., 2017), Araneae and Opiliones (Astrin et al., 2016), and Myriapoda (Spelda, Reip, Oliveira Biener, & Melzer, 2011; Wesener et al., 2015). Concerning DNA barcoding studies performed for Diptera, no comprehensive study encompassing this entire highly diverse order has been published, but data have been used to revise smaller units thereof: for example, for Calliphoridae (Jordaens et al., 2013; Nelson, Wallman, & Dowton, 2007; Reibe, Schmitz, & Madea, 2009), Ceratopogonidae (Stur & Borkent, 2014), Chironomidae (Carew, Pettigrove, Cox, & Hoffmann, 2007; Carew, Pettigrove, & Hoffmann, 2005; Cranston et al., 2013; Ekrem, Stur, & Hebert, 2010; Ekrem, Willassen, & Stur, 2007; Montagna, Mereghetti, Lencioni, & Rossaro, 2016; Pfenninger, Nowak, Kley, Steinke, & Streit, 2007; Sinclair & Gresens, 2008; Stur & Ekrem, 2011), Culicidae (Ashfaq et al., 2014; Cywinska, Hunter, & Hebert, 2006; Kumar, Rajavel, Natarajan, & Jambulingam, 2007; Versteirt et al., 2015; Wang et al., 2012), Hybotidae (Nagy, Sonet, Mortelmans, Vandewynkel, & Grootaert, 2013), Muscidae (Renaud, Savage, & Adamowicz, 2012), Psychodidae (Gutiérrez, Vivero, Vélez, Porter, & Uribe, 2014; Krüger, Strüven, Post, & Faulde, 2011; Kumar, Srinivasan, & Jambulingam, 2012; Nzelu et al., 2015), Sciaridae (Eiseman, Heller, & Rulik, 2016; Heller, Köhler, Menzel, Olsen, & Gammelo, 2016; Heller & Rulik, 2016; Latibari, Moravvej, Heller, Rulik, & Namaghi, 2015; Ševčík, Kaspřák, & Rulik, 2016), Simuliidae (Rivera & Currie, 2009), Syrphidae (Jordaens et al., 2015) and Tachinidae (Pohjoismäki, Kahanpää, & Mutanen, 2016).

This publication presents the first results of the Diptera campaign and it provides coverage for 5,200 BINs (Ratnasingham & Hebert, 2013). It covers ~55% of the known Diptera fauna from Germany. According to the checklist of German Diptera (Schumann et al., 1999) and the three additions published so far (Schumann, 2002, 2004, 2010) 9,544 species of Diptera have been recorded from Germany. The Diptera library now includes a total of 2,453 reliable species identifications, and 2,700 BINs, which possess either interim species names or just higher-level taxonomy (genus or family; "dark taxa"). Although it has been shown that BINs correspond closely to biological species of most insect orders (Hausmann et al., 2013), there are other studies reporting difficulties in determining species through DNA barcodes within Diptera. In particular, well-studied groups such as the syrphids represent a problem, because here additional genes for a clear type assignment must be consulted in many genera (Mengual, Ståhls, Vujić, & Marcos-Garcia, 2006; Rojo, Ståhls, Pérez-Bañón, & Marcos-García, 2006). Further examples of problems in species delineation due to barcode gaps, at least for some genera, are the Tachinidae and the Calliphoridae (Nelson et al., 2012; Pohjoismäki et al., 2016; Whitworth, Dawson, Magalon, & Baudry, 2007). In one of the few studies dealing with DNA barcoding in Diptera it was shown, that less than 70% of a composition of about 450 species covering 12 families of Diptera could be reliably identified by DNA barcoding, as there was wide overlap between intra- and interspecific genetic variability on the *COI* gene (Meier, Shiyang, Vaidya, & Ng, 2006). However we find that more than 88% of the studied species, identified based on morphology or BIN matches to the BOLD database, can be unambiguously identified using their DNA barcode sequences. BINs enable the creation of an interim

taxonomic system in a structured, transparent and sustainable way and thus become a valuable foundation for subsequent detailed, integrative taxonomic studies. Furthermore, the BIN system enables analyses that are equivalent to studies based on named species, that is where the underlying specimens are identified by specialists using traditional methods (i.e., morphology). The latter will play a special role in the processing, classification and genetic inventorying of less-explored "dark taxa," which have been treated and processed with less priority by previous DNA barcoding activities. Moreover, this automated approach of delineating species is less affected by operational taxonomic biases, so it can provide more objective identifications than conventional approaches (Mutanen et al., 2016; Packer, Gibbs, Sheffield, & Hanner, 2009; Schmidt et al., 2015). Using DNA extracts derived from bulk collections made by Malaise traps, we further demonstrate that species delineation using interim names based on BINs and OTUs constitutes an effective method for biodiversity studies using DNA metabarcoding. As the reference libraries continue to grow and gaps in the species catalogue are subsequently filled, BIN lists assembled by metabarcoding will provide improved taxonomic resolution.

The present study has three main goals: (a) to provide a DNA barcode library for 5,200 BINs of Diptera; (b) to demonstrate, based on the example of bulk extractions from a Malaise trap experiment, that DNA barcode clusters, labelled with globally unique identifiers (such as OTUs and/or BINs), provide a pragmatic, accurate solution to the "taxonomic impediment"; and (c) to demonstrate that interim names based on BINs and OTUs obtained through metabarcoding is an effective method for studies on species-rich groups that are usually neglected in biodiversity research projects because of their unresolved taxonomy.

## 2 | MATERIALS AND METHODS

### 2.1 | Fieldwork, specimens and taxonomy

A network of 130 (professional and voluntary) taxonomists and citizen scientists collected and contributed specimens to the DNA barcoding projects, primarily from various German states, but also from surrounding European countries (Austria, Belgium, Czech Republic, France, Italy). Most specimens (94.5%, 42,587 of 45,040 with *COI* sequences >500 bp) were collected by Malaise traps, which were deployed from 2009 to 2016. The study sites included more than 683 localities in state forests, public lands and protected areas such as the Nationalparks "Bayerischer Wald" and "Berchtesgadener Land," the EU habitats directive site "Landskrone," as well as alpine regions at altitudes up to 2,926 m (Zugspitze). Detailed information on collection sites and dates is available in Appendix S1. Since 2009, more than five million specimens of Diptera were collected by hand collecting, sweep netting, and by Malaise-, window- and pitfall-trapping. However, most voucher specimens have been extracted from Malaise trap samples. Twenty to 100 Malaise traps were deployed in each of seven years (2011–2017) mostly across habitats in Bavaria and Baden-Wurttemberg; one trap was placed in Rhineland-Palatinate. Samples were screened morphologically to maximize the diversity of species submitted for sequence characterization. Most vouchers were derived from Germany (44,511), but others were collected in France (222), Czech Republic (147), Belgium (106), Austria (70) and other Central European countries (18). All samples and specimens are now stored in the SNSB-ZSM or ZFMK except for a few held in private collections. From the entire collection, ~3,000,000 specimens of potential interest, most of which derived from the huge Malaise trap experiments in the framework of the GMTP, were identified to family level mostly by D.D. and to a minor extent by B.R. and experienced specialists using appropriate literature (Oosterbroek, 2006 and references therein; Papp & Darvas, 1997, 1998, 2000a, 2000b, Schumann et al., 2011). From this material, 59,000 specimens were submitted for sequence analysis through the DNA barcoding pipeline (including sample preparation, high-quality imaging and metadata acquisition for each specimen) established at the ZSM to support its involvement in national and international DNA barcoding projects. Most samples (>99%) were stored in 96% EtOH before DNA extraction. Specimen ages generally ranged from 1 to 5 years (43,112 specimens, 96%); only 4% were more than 5 years old. The number of specimens analysed per species ranged from one to 1,356 (i.e., *Megaselia rufa*) (Wood, 1908; see Appendix S1). When taxonomic expertise was available, specimens were sent to specialists to obtain as many species-level identifications as possible.

### 2.2 | Laboratory protocols

A tissue sample was removed from each specimen and transferred into 96-well plates at the SNSB-ZSM for subsequent DNA extraction. For specimens with a body length >2 mm a single leg or a leg segment was removed for DNA extraction. The whole voucher was used for some very small specimens (e.g., ≤1 mm, such as small Cecidomyiidae, Chironomidae and Sciaridae), but replacement vouchers from the same locality were retained. In other cases (vouchers from Malaise traps), DNA was extracted from the whole voucher at the CCDB (Guelph, Canada) using "voucher-recovery" protocols (DeWaard et al., 2019) and the specimens were repatriated to the SNSB-ZSM and ZFMK for identification and curation. DNA extraction plates with the tissue samples were sent to the Canadian Center for DNA Barcoding (CCDB) where they were processed using standard protocols. All protocols for DNA extraction, PCR amplifications and Sanger sequencing procedures are available online (ccdb.ca/resources/). All samples were PCR-amplified with a cocktail of standard and modified Folmer primers CLepFolF (5′-ATTCAACCAATCATAAAGATATTGG) and CLepFolR (5′TAAACTTCTGGATGTCCAAAAAATCA) for the barcode fragment (5′ *COI*; see Hernández-Triana et al., 2014), and the same primers were employed for subsequent bidirectional Sanger sequencing reactions (see also Ivanova, Dewaard, & Hebert, 2006; deWaard, Ivanova, Hajibabaei, & Hebert, 2008, DeWaard et al., 2019). Voucher information such as locality data, habitat, altitude, collector, identifier, taxonomic classifications, habitus images, DNA barcode sequences, primer pairs and trace files for 40,753

specimens are publicly accessible in the "DS-DIPBFGBL—A DNA Barcode reference library of German Diptera (BFB—Barcoding Fauna Bavarica & GBOL—German Barcode of Life" data set on BOLD (http://www.boldsystems.org – data set DOI: dx.doi.org/10.5883/DS-DIPBFGBL), whereas 4,420 specimen records will be stored in the private data set "DS-DIPBFGBP—A DNA Barcode reference library of German Diptera (BFB—Barcoding Fauna Bavarica & GBOL—German Barcode of Life)—private records for future publication" for subsequent publication by the authors and associated taxonomists.

## 2.3 | Data analysis

Sequence divergences for the *COI*-5P barcode region (mean and maximum intraspecific variation and minimum genetic distance to the nearest-neighbouring species) were calculated using the "*Barcode Gap Analysis*" tool on BOLD, employing the Kimura 2-parameter (K2P) distance metric (Puillandre, Lambert, Brouillet, & Achaz, 2012). The program MUSCLE was applied for sequence alignment restricting analysis to sequences with a minimum length of 500 bp. Neighbour-joining (NJ) trees were calculated following alignment based on K2P distances. The "BIN Discordance" analysis on BOLD was used to reveal cases where species assigned to different species shared a BIN, and those cases where a particular species was assigned to two or more BINs. Sequences are grouped into clusters of closely similar *COI* barcode sequences, which are assigned a globally unique identifier, termed a "barcode index number" or BIN (Ratnasingham & Hebert, 2013). This system enables tentative species identifications when taxonomic information is lacking. The BIN system involves a three-step online pipeline, which clusters similar barcode sequences algorithmically into OTUs being "named" by a number. For the majority of studied insect orders, specimens sharing a BIN very often represent a close species-proxy as delineated by traditional taxonomy (e.g., for Lepidoptera, Hausmann et al., 2013). However, some genera or families throughout the insects exhibit problems with species delineation based on DNA barcodes, due to high intra- or low interspecific genetic distances (e.g., cryptic diversity, BIN sharing or the barcode gap; see Hubert & Hanner, 2015). Within the Diptera, this phenomenon has been well documented (Meier et al., 2006), at least in some families, such as calliphorid, syrphid and tachinid species (Mengual et al., 2006; Nelson et al., 2012; Pohjoismäki et al., 2016; Rojo et al., 2006; Whitworth et al., 2007), but may also occur in families of "dark taxa" as well.

Every other "disagreement/conflict" case is the starting point for re-evaluation of both molecular and morphological data. We follow the concept of Integrative Taxonomy (Fujita et al., 2012; Padial et al., 2010; Schlick-Steiner et al., 2014, 2010) to infer whether there are previously overlooked species ("cryptic taxa") in the sample, or whether barcode divergence between species is too low or absent to allow valid species to be delineated using only *COI* characteristics.

## 2.4 | Reverse-taxonomy approach

When sequenced specimens could only be assigned to a category above the species level (family, subfamily or genus), we used interim species names (such as TachIntGen1 sp.BOLD:AAG2112) based on the corresponding BIN, so these specimens could be included in the "Barcode Gap Analysis" in order to provide more comprehensive estimates of the distribution of genetic divergences among both species assigned to Linnaean species and those with BIN assignments. This analysis was conducted on all specimens at the same time after updating the interim taxonomy where necessary. For specimen records, which lack lower taxonomy (e.g., those uploaded only as "Diptera"), we applied the highest "conflict-free" taxonomy—for example the genus name, when other specimens within that BIN had the same identification—using a BIN match with the public data on BOLD (e.g., *Melanagromyza* sp. BOLD:ACP6151). All specimens, which could not be identified to species or genus level, and where the vouchers were in acceptable condition (e.g., unbroken antennae and/or legs after retrieval from Malaise trap), were selected using the corresponding BINs for identification by taxonomic specialists. Interim names were subsequently moved into the "Voucher status" field in the BOLD metadata tables after all analyses were performed.

## 2.5 | Metabarcoding and bioinformatic data analysis

The potential utility of the DNA barcode library for biomonitoring Diptera was tested with field samples, focusing on an early warning system for pest and invasive species based on metabarcoding (L. A. Hardulak et al. in prep). In this study, nine Malaise traps were deployed in the Bayerischer Wald National Park and its surroundings during the vegetated period (May–September) in 2016. Trap bottles were changed twice monthly, producing a total of 90 bulk samples of macroinvertebrates. All specimens were dried and ground with a stainless steel pestle (no size-sorting step), and tissue lysis of insect powder per trap sample was performed overnight, using a solution of 90% insect lysis buffer and 10% proteinase K. DNA extraction was performed with the DNEasy Blood & Tissue kit (Qiagen). A minibarcode region was amplified by PCR, using forward and reverse NGS primers (Leray et al., 2013) targeting a 313-bp-long coding region of mitochondrial *COI*. High-throughput sequencing was performed on an Illumina MiSeq using version 2 (2 × 250 bp, 500 cycles, maximum of 20 million reads) chemistry at the Sequencing Service Unit of the Ludwig-Maximilians University (LMU, Munich, Germany; see Appendix S5 for a more detailed metabarcoding protocol).

Sequence processing was performed with the VSEARCH version 2.4.3 suite (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) and CUTADAPT version 1.14 (Martin, 2011). Forward and reverse reads in each sample were merged with the VSEARCH program "fastq_mergepairs" with a minimum overlap of 40 bp, yielding ~313-bp sequences. Forward and reverse primers were removed with CUTADAPT, using the "discard_untrimmed" option to discard sequences for which primers were not detected at ≥90% identity. Quality filtering was done with the "fastq_filter" in VSEARCH, keeping sequences with zero expected errors ("fastq_maxee" 1). Sequences were dereplicated with "derep_fulllength," first at the sample level, and then concatenated into a fasta file, which was then dereplicated. Chimeric sequences were removed from the fasta file using "uchime_denovo." The remaining

sequences were then clustered into OTUs at 97% identity employing "cluster_size," a greedy, centroid-based clustering program. OTUs were BLASTed against the Diptera database downloaded from BOLD including taxonomy and BIN information in GENEIOUS (version 9.1.7; Biomatters) following the methods described in Morinière et al. (2016). The resulting csv file, which included BIN, Hit-%-ID value, family, genus and species information for each out, was exported from Geneious and combined with the OTU table generated by the bioinformatic pipeline. The combined results table was then filtered by Hit-%-ID value and total read numbers per OTU. All entries with identifications below 97% and total read numbers below 0.01% of the summed reads per sample were removed from the analysis. OTUs were then assigned to the respective BIN (Appendix S2). Presence–absence overviews of selected Diptera taxa (BINs) within the metabarcoding study were created; one-sided Pearson correlation coefficients were calculated to estimate the percentage of "dark taxa" with mid-range body size versus the number of species reported in Germany, both with the inclusion and with the exclusion of families with 0% "dark taxa." (R version 3.4.4 [2018–03-15], R Core Team, 2018).

## 3 | RESULTS

### 3.1 | DNA barcoding/developing a reference library

From the 59,102 specimens submitted for Sanger sequencing, 50,963 *COI*-5P sequences (86.23%) were recovered. Length of the recovered sequence varied with the sequencing protocol; 12.54% (7,410 specimens) were bidirectionally sequenced and yielded a full-length (658 bp) barcode while the rest (43,533) were unidirectionally sequenced yielding 69.95% (41,339) with sequences <658 to >500 bp and 3.75% (2,214 specimens) with sequences <500 bp. No sequence information was recovered from 13.77% (8,139) of the specimens. Barcode recovery was most successful for EtOH-preserved specimens less than 10 years old. For the subsequent analyses we selected 45,040 specimens with high-quality DNA barcode sequences (≥500 bp), which fulfilled the requirements for being assigned to a BIN. This data set included ~5,200 BINs (2,500 were assigned a total of 2,453 Linnean species while 2,700 lacked a species designation, 52.4% of the data set). These BINs included one or more representatives from 88 of the 117 (75%) dipteran families known from Germany (Figure 1, Table 1; Appendix S3, Krona graph in Figure S2). More than one-third (1,829) of the BINs were new to BOLD.

Inspection of the *COI* sequence clusters using NJ trees (created with analytical tools on BOLD) and using the TaxCl-approach for detecting taxonomic incongruences (Rulik et al., 2017) revealed high congruence with morphology-based identifications. Among the 2,453 taxa assigned a Linnean binomen based on morphological identifications and "conflict-free" BIN matches, 88.67% (2,138) were unambiguously discriminated by their *COI* sequences. Another 122 species (4.97%), representing 8.7% of all studied specimens (3,951 individuals), were assigned to more than one BIN, resulting in a total of 255 BINs (Table 1; Appendix S3). For purposes of re-identification, the species in this subset can also be unambiguously assigned

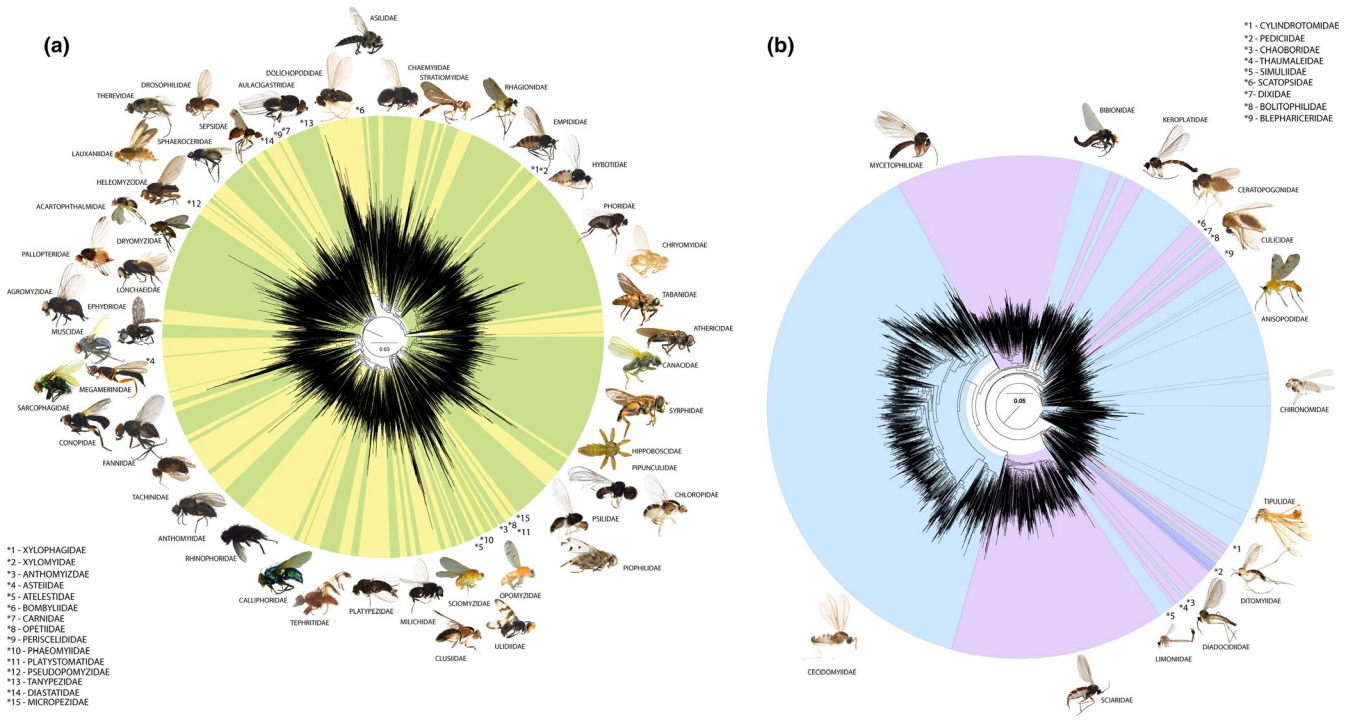


**FIGURE 1** Illustrative circular neighbour-joining (NJ) trees for (a) all Brachycera and (b) all Nemtatocera within the Diptera barcode library; each line in the trees corresponds to one barcode index number (BIN). NJ tree calculations were performed on the BOLD database. A more detailed observation of the BIN diversity for each family can be studied within the Krona graph within the supporting information (Figure S2)

**TABLE 1** Families of Diptera reported in Germany. Information on BIN count, and on the numbers of named and unnamed species within the reference database

| Infraorder | Family | Species reported in Germany | BINs | Ratio barcoded/species (%) | Size (mm) | Total number of taxa/with barcode | Unnamed/with barcode | % of dark taxa |
|---|---|---|---|---|---|---|---|---|
| Brachycera | Acartophthalmidae | 2 | 1 | 50 | 1.0–2.5 | 2 | 0 | 0 |
| Brachycera | Acroceridae | 11 | 0 | 0 | 2.5–20.0 | N/A | N/A | N/A |
| Brachycera | Agromyzidae | 552 | 218 | 39 | 1.0–6.0 | 214 | 149 | 70 |
| Nematocera | Anisopodidae (& Mycetobiidae) | 8 | 7 | 88 | 4.0–12.0 | 7 | 2 | 29 |
| Brachycera | Anthomyiidae | 227 | 188 | 83 | 4.0–12.0 | 178 | 64 | 36 |
| Brachycera | Anthomyzidae | 14 | 5 | 36 | 1.3–4.5 | 5 | 0 | 0 |
| Brachycera | Asilidae | 81 | 18 | 22 | 8.0–20.0 | 18 | 6 | 33 |
| Brachycera | Asteiidae | 7 | 3 | 43 | 1.0–3.0 | 3 | 0 | 0 |
| Brachycera | Atelestidae | 3 | 3 | 100 | 1.5–3.5 | 3 | 0 | 0 |
| Brachycera | Athericidae | 5 | 3 | 60 | 7.5–10.0 | 3 | 1 | 33 |
| Brachycera | Aulacigastridae | 1 | 0 | 0 | 2.0–5.0 | 0 | 0 | N/A |
| Nematocera | Bibionidae (& Pleciidae) | 21 | 12 | 57 | 2.0–15.0 | 10 | 2 | 20 |
| Nematocera | Blephariceridae | 7 | 2 | 29 | 3.0–15.0 | 2 | 1 | 50 |
| Nematocera | Bolitophilidae | 22 | 14 | 64 | 4.0–7.0 | 13 | 7 | 54 |
| Brachycera | Bombyliidae | 40 | 6 | 15 | 1.0–20.0 | 6 | 1 | 17 |
| Brachycera | Braulidae | 1 | 0 | 0 | 1.2–2.5 | 2 | 0 | 0 |
| Brachycera | Calliphoridae | 62 | 35 | 56 | 4.0–16.0 | 39 | 6 | 15 |
| Brachycera | Camillidae | 4 | 0 | 0 | 2.0–3.5 | N/A | N/A | N/A |
| Brachycera | Campichoetidae | 3 | 0 | 0 | 2.5–4.0 | N/A | N/A | N/A |
| Brachycera | Canacidae | 2 | 9 | 450 | 1.6–5.0 | 9 | 1 | 11 |
| Nematocera | Canthyloscelidae | 1 | 0 | 0 | 2.5–9.0 | N/A | N/A | N/A |
| Brachycera | Carnidae | 11 | 7 | 64 | 1.0–2.5 | 7 | 7 | 100 |
| Nematocera | Cecidomyiidae | 836 | 927 | 111 | 0.5–3.0 | 926 | 882 | 95 |
| Nematocera | Ceratopogonidae | 332 | 131 | 39 | 1.0–5.0 | 128 | 97 | 76 |
| Brachycera | Chamaemyiidae | 29 | 17 | 59 | 1.0–5.0 | 17 | 13 | 76 |
| Nematocera | Chaoboridae | 7 | 2 | 29 | 2.0–10.0 | 2 | 0 | 0 |
| Nematocera | Chironomidae | 696 | 455 | 65 | 1.0–10.0 | 438 | 286 | 65 |
| Brachycera | Chloropidae | 198 | 101 | 51 | 1.0–5.0 | 101 | 59 | 58 |
| Brachycera | Chyromyidae | 5 | 2 | 40 | 0.5–8.0 | 2 | 0 | 0 |
| Brachycera | Clusiidae | 9 | 6 | 67 | 1.5–8.0 | 7 | 3 | 43 |
| Brachycera | Coelopidae | 2 | 0 | 0 | 2.5–9.0 | N/A | N/A | N/A |
| Brachycera | Coenomyiidae | 1 | 0 | 0 | 14.0–20.0 | N/A | N/A | N/A |
| Brachycera | Conopidae | 52 | 9 | 17 | 5.0–15.0 | 9 | 0 | 0 |
| Brachycera | Cremifaniidae | 1 | 0 | 0 | 1.5–2.6 | N/A | N/A | N/A |
| Brachycera | Cryptochetidae | 1 | 0 | 0 | 2.0–4.0 | N/A | N/A | N/A |
| Nematocera | Culicidae | 46 | 8 | 17 | 3.0–9.0 | 7 | 0 | 0 |
| Nematocera | Cylindrotomidae | 4 | 1 | 25 | 11.0–16.0 | 1 | 0 | 0 |
| Nematocera | Diadocidiidae | 4 | 3 | 75 | 3–4.5.0 | 3 | 0 | 0 |
| Brachycera | Diastatidae | 6 | 8 | 133 | 2.5–4.0 | 8 | 2 | 25 |
| Nematocera | Ditomyiidae | 4 | 1 | 25 | 6.0–8.0 | 1 | 0 | 0 |
| Nematocera | Dixidae | 16 | 4 | 25 | 3.0–5.5 | 4 | 1 | 25 |
| Brachycera | Dolichopodidae | 356 | 112 | 31 | 1.0–9.0 | 112 | 58 | 52 |

(Continues)

**TABLE 1** (Continued)

| Infraorder | Family | Species reported in Germany | BINs | Ratio barcoded/species (%) | Size (mm) | Total number of taxa/with barcode | Unnamed/with barcode | % of dark taxa |
|---|---|---|---|---|---|---|---|---|
| Brachycera | Drosophilidae | 59 | 28 | 47 | 1.5–7.0 | 28 | 5 | 18 |
| Brachycera | Dryomyzidae | 3 | 2 | 67 | 5.0–18.0 | 2 | 1 | 50 |
| Brachycera | Eginiidae | 1 | 0 | 0 | 2.0–18.0 | N/A | N/A | N/A |
| Brachycera | Empididae (& Brachystomatidae) | 383 | 161 | 42 | 1.0–12.0 | 161 | 107 | 66 |
| Brachycera | Ephydridae | 177 | 130 | 73 | 1.0–11.0 | 132 | 16 | 12 |
| Brachycera | Fanniidae | 56 | 46 | 82 | 2.0–5.0 | 44 | 13 | 30 |
| Brachycera | Gasterophilidae | 4 | 0 | 0 | 9.0–16.0 | N/A | N/A | N/A |
| Brachycera | Helcomyzidae | 3 | 0 | 0 | 6.0–11.0 | N/A | N/A | N/A |
| Brachycera | Heleomyzidae (& Heteromyzidae) | 74 | 58 | 78 | 1.2–12.0 | 55 | 26 | 47 |
| Nematocera | Hesperinidae | 1 | 0 | 0 | 4.0–6.0 | N/A | N/A | N/A |
| Brachycera | Hilarimorphidae | 2 | 0 | 0 | 2.0–7.0 | N/A | N/A | N/A |
| Brachycera | Hippoboscidae | 12 | 7 | 58 | 2.5–10.0 | 7 | 1 | 14 |
| Brachycera | Hybotidae | 229 | 140 | 61 | 1.0–6.0 | 139 | 83 | 60 |
| Brachycera | Hypodermatidae | 5 | 0 | 0 | 10.0–22 | N/A | N/A | N/A |
| Nematocera | Keroplatidae | 60 | 30 | 50 | 4.0–15.0 | 30 | 12 | 40 |
| Brachycera | Lauxaniidae | 67 | 25 | 37 | 2.0–7.0 | 25 | 11 | 44 |
| Nematocera | Limoniidae | 280 | 96 | 34 | 2.0–11.0 | 91 | 50 | 55 |
| Brachycera | Lonchaeidae | 47 | 16 | 34 | 3.0–6.0 | 16 | 9 | 56 |
| Brachycera | Lonchopteridae | 9 | 5 | 56 | 2.0–5.0 | 6 | 0 | 0 |
| Brachycera | Megamerinidae | 1 | 1 | 100 | 6.0–9.0 | 1 | 0 | 0 |
| Brachycera | Micropezidae | 13 | 5 | 38 | 3.0–16.0 | 4 | 1 | 25 |
| Brachycera | Microphoridae | 6 | 0 | 0 | 1.5–3.0 | N/A | N/A | N/A |
| Brachycera | Milichiidae | 13 | 17 | 131 | 1.0–6.0 | 16 | 9 | 56 |
| Brachycera | Muscidae | 317 | 174 | 55 | 2.0–18.0 | 167 | 66 | 40 |
| Nematocera | Mycetophilidae | 573 | 306 | 53 | 2.0–15.0 | 301 | 89 | 30 |
| Brachycera | Neottiophilidae | 1 | 0 | 0 | 1.5–7.0 | N/A | N/A | N/A |
| Brachycera | Nycteribiidae | 8 | 0 | 0 | 1.5–5.0 | N/A | N/A | N/A |
| Brachycera | Odiniidae | 9 | 0 | 0 | 2.0–5.0 | N/A | N/A | N/A |
| Brachycera | Oestridae | 6 | 0 | 0 | 9.0–18.0 | N/A | N/A | N/A |
| Brachycera | Opetiidae | 1 | 1 | 100 | 2.0–5.0 | 1 | 0 | 0 |
| Brachycera | Opomyzidae | 15 | 4 | 27 | 2.0–5.0 | 4 | 1 | 25 |
| Brachycera | Otitidae | 26 | 0 | 0 | 2.5–11.0 | N/A | N/A | N/A |
| Brachycera | Pallopteridae | 16 | 8 | 50 | 2.5–7.0 | 7 | 0 | 0 |
| Nematocera | Pediciidae | 36 | 13 | 36 | 5.0–35.0 | 13 | 3 | 23 |
| Brachycera | Periscelididae | 6 | 1 | 17 | 1.0–5.0 | 1 | 0 | 0 |
| Brachycera | Phaeomyiidae | 3 | 2 | 67 | 3.0–11.0 | 2 | 0 | 0 |
| Brachycera | Phoridae | 364 | 289 | 79 | 0.5–6.0 | 276 | 166 | 60 |
| Brachycera | Piophilidae | 12 | 12 | 100 | 1.5–7.0 | 12 | 4 | 33 |
| Brachycera | Pipunculidae | 111 | 42 | 38 | 2.0–12.0 | 40 | 7 | 18 |
| Brachycera | Platypezidae | 23 | 4 | 17 | 1.5–6.0 | 4 | 0 | 0 |
| Brachycera | Platystomatidae | 3 | 2 | 67 | 3.0–11.0 | 2 | 0 | 0 |
| Brachycera | Pseudopomyzidae | 1 | 1 | 100 | 1.7–2.5 | 1 | 0 | 0 |

(Continues)

**TABLE 1** (Continued)

| Infraorder | Family | Species reported in Germany | BINs | Ratio barcoded/species (%) | Size (mm) | Total number of taxa/with barcode | Unnamed/with barcode | % of dark taxa |
|---|---|---|---|---|---|---|---|---|
| Brachycera | Psilidae | 30 | 12 | 40 | 2.5–10.0 | 12 | 8 | 67 |
| Nematocera | Psychodidae | 143 | 51 | 36 | 2.0–6.0 | 50 | 25 | 50 |
| Nematocera | Ptychopteridae | 8 | 0 | 0 | 7.0–15.0 | N/A | N/A | N/A |
| Brachycera | Pyrgotidae | 1 | 0 | 0 | 8.0–9.0 | N/A | N/A | N/A |
| Brachycera | Rhagionidae | 35 | 20 | 57 | 2.0–20.0 | 20 | 10 | 50 |
| Brachycera | Rhinophoridae | 10 | 9 | 90 | 2.0–11.0 | 7 | 1 | 14 |
| Brachycera | Sarcophagidae | 130 | 49 | 38 | 3.0–22.0 | 49 | 17 | 35 |
| Brachycera | Scatophagidae | 57 | 0 | 0 | 3.0–12.0 | 0 | 0 | N/A |
| Nematocera | Scatopsidae | 47 | 30 | 64 | 0.5–4.0 | 30 | 24 | 80 |
| Brachycera | Scenopinidae | 3 | 0 | 0 | 2.0–7.0 | N/A | N/A | N/A |
| Nematocera | Sciaridae | 342 | 310 | 91 | 1.0–6.0 | 284 | 81 | 29 |
| Brachycera | Sciomyzidae | 78 | 19 | 24 | 2.0–14.0 | 18 | 4 | 22 |
| Brachycera | Sepsidae | 31 | 15 | 48 | 2.0–6.0 | 13 | 1 | 8 |
| Nematocera | Simuliidae | 50 | 19 | 38 | 1.2–6.0 | 18 | 9 | 50 |
| Brachycera | Sphaeroceridae | 137 | 79 | 58 | 0.7–5.5 | 77 | 31 | 40 |
| Brachycera | Stratiomyidae | 66 | 21 | 32 | 2.0–25.0 | 22 | 6 | 27 |
| Brachycera | Strongylophthalmyiidae | 1 | 0 | 0 | 3.0–5.5 | N/A | N/A | N/A |
| Brachycera | Syrphidae | 440 | 242 | 55 | 3.5–35.0 | 297 | 24 | 8 |
| Brachycera | Tabanidae | 58 | 46 | 79 | 6.0–30.0 | 45 | 3 | 7 |
| Brachycera | Tachinidae | 494 | 214 | 43 | 2.0–20.0 | 211 | 76 | 36 |
| Brachycera | Tanypezidae | 1 | 1 | 100 | 5.0–8.0 | 1 | 0 | 0 |
| Brachycera | Tephritidae | 110 | 28 | 25 | 2.5–10.0 | 27 | 5 | 19 |
| Brachycera | Tethinidae | 10 | 0 | 0 | 1.5–3.5 | N/A | N/A | N/A |
| Nematocera | Thaumaleidae | 15 | 13 | 87 | 3.0–5.0 | 13 | 1 | 8 |
| Brachycera | Therevidae | 32 | 4 | 13 | 2.5–15.0 | 4 | 1 | 25 |
| Brachycera | Thyreophoridae | 2 | 0 | 0 | 1.5–7.0 | N/A | N/A | N/A |
| Nematocera | Tipulidae | 123 | 46 | 37 | 7.0–35.0 | 46 | 15 | 33 |
| Nematocera | Trichoceridae | 18 | 24 | 133 | 3.0–9.0 | 24 | 17 | 71 |
| Brachycera | Trixoscelididae | 4 | 0 | 0 | 2.0–4.0 | N/A | N/A | N/A |
| Brachycera | Ulidiidae | 4 | 9 | 225 | 2.5–11.0 | 9 | 4 | 44 |
| Brachycera | Xylomyidae | 3 | 1 | 33 | 6.0–20.0 | 1 | 0 | 0 |
| Brachycera | Xylophagidae | 4 | 1 | 25 | 5.0–11.0 | 1 | 0 | 0 |

*Note*: Additional information on the average body size of the specimens in each family is included.

to a current species. For 34 of these taxa, the maximum intraspecific variation (maxISP) was <3% (range: 1.1%–3.0%), cases which may reflect either young sibling species or high intraspecific variation arising from secondary contact between phylogeographical lineages. Another 88 species showed considerably higher divergences with maxISP ranging from 3% to 6% in 48 species and from 6% to 12% in another 40 species, cases that are strong candidates for overlooked cryptic diversity. Most of these cases involved species whose members were assigned to two BINs (112 species), but specimens of nine species were assigned to three BINs and those of one other to four BINs. Another 156 species (6.56%), representing 2.9% of all specimens (1,316 specimens), involved two or more named species that shared a BIN (Table 2). Ten of these species pairs possessed shallow but consistent divergences within the BIN, meaning that *COI* sequences enabled species identification (e.g., *Chrysotoxum bicinctum* Linnaeus, 1758 and *Chrysotoxum festivum* Linnaeus, 1758; *Sericomyia lappona* Linnaeus, 1758 and *Sericomyia silentis* Harris 1776; *Paragus majoranae* Rondani, 1857 and *Paragus pecchiolii*, Rondani, 1857). Interestingly, almost two-thirds (105/156) of the species exhibiting BIN sharing (168) were hoverflies (Syrphidae), a family that has seen intensive taxonomic study.

Appendix S1 provides species names, sample IDs, BIN assignment and collection information. All project data are available under the publicly accessible DOI: dx.doi.org/10.5883/DS-DIPBFGBL.

### 3.2 | Performance of the reference library for metabarcoding of Malaise trap samples

Among the 90 Malaise trap samples from the Bavarian Forest National Park (L. A. Hardulak et al. in prep.), metabarcoding revealed 1,735 dipteran OTUs, comprising 536,376 reads: 5,960 average reads per sample, matching at 97% or higher to a taxon in the DNA barcode library downloaded from BOLD (average read count was 6,928 per sample with a total of 2,809 OTUs matched to the Diptera database with ≥90%). Multiple OTU matches to a single BIN were merged. Using the Diptera data, we identified a total of 1,403 BINs including representatives of 71 families (1,385 species) within the metabarcoding data set (Appendix S2). Almost one-third (498/1403) of these BINs belonged to "dark taxa." Figure 2 illustrates examples of presence/absence overviews for the families Muscidae, Cecidomyiidae, Chironomidae and Syrphidae for selected Malaise trap sites.

Among families containing "dark taxa," the percentage of unnamed taxa was inversely correlated with body size ($r$ = −0.41, $p$ = 0.0004) and positively with numbers of species reported from Germany (r = 0.33, $p$ = 0.0037) (Figure 3; Appendix S4).

## 4 | DISCUSSION

This study summarizes the results of a DNA barcoding campaign on German Diptera, work based on the characterization of 45,040 specimens. The resultant DNA barcode reference library included records for 5,200 BINs (2,453 named species comprising 2,500 BINs plus 2,700 unnamed BINs) belonging to 88 families, covering ~ 50% of the Diptera fauna reported for Germany (Schumann, 2002, 2004, 2010; Schumann et al., 1999). Until now, most of these families, especially some of the most diverse, have been taxonomically

**TABLE 2** All cases of high intraspecific sequence variation at *COI*; cases of multiple BINs and/or cryptic diversity candidates (CDC)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Agromyzidae | *Napomyza cichorii* | *CDC (2)* | 2.47 | 3.71 | BOLD:AAP2990 |
| | | | | | BOLD:AAX3741 |
| | *Phytomyza continua* | *CDC (2)* | 2.84 | 5.44 | BOLD:AAM6330 |
| | | | | | BOLD:AAY2701 |
| | *Phytomyza ranunculi* | *CDC (2)* | 3.26 | 6.43 | BOLD:AAY3895 |
| | | | | | BOLD:ACL2003 |
| Anthomyiidae | *Anthomyia liturata* | *CDC (2)* | 0.87 | 1.98 | BOLD:ACE4539 |
| | | | | | BOLD:ACE4540 |
| | *Delia nuda* | *CDC (2)* | 1.06 | 1.87 | BOLD:ACJ0544 |
| | | | | | BOLD:ACJ0545 |
| | *Hydrophoria lancifer* | *CDC (2)* | 0.61 | 3.04 | BOLD:AAG2460 |
| | | | | | BOLD:ADC1814 |
| | *Pegomya flavifrons* | *CDC (2)* | 2.5 | 8.83 | BOLD:AAG2479 |
| | | | | | BOLD:AAG6754 |
| | *Pegomya solennis* | *CDC (2)* | 0.85 | 2.67 | BOLD:ACD8686 |
| | | | | | BOLD:ACM6225 |
| | *Pegomya winthemi* | *CDC (2)* | 0.54 | 5.53 | BOLD:AAG1783 |
| | | | | | BOLD:ABA6845 |
| Bibionidae | *Bibio clavipes* | *CDC (2)* | 1.2 | 2.46 | BOLD:ACC6151 |
| | | | | | BOLD:ACR0881 |
| | *Bibio nigriventris* | *CDC (2)* | 1 | 3.13 | BOLD:ABX1732 |
| | | | | | BOLD:ACU5368 |
| Bolitophilidae | *Bolitophila austriaca* | *CDC (2)* | 1.27 | 2.18 | BOLD:AAG4863 |
| | | | | | BOLD:ACI5612 |
| Ceratopogonidae | *Brachypogon sociabilis* | *CDC (2)* | 1.24 | 2.31 | BOLD:ABW3958 |
| | | | | | BOLD:ACE8195 |
| | *Ceratopogon grandiforceps* | *CDC (2)* | 2.63 | 3.94 | BOLD:ABW3984 |
| | | | | | BOLD:ACP4327 |
| | *Forcipomyia* sp. 4ES | *CDC (2)* | 2.18 | 5.98 | BOLD:AAM6200 |
| | | | | | BOLD:ACQ8860 |

(Continues)

**TABLE 2** (Continued)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Chironomidae | *Brillia bifida* | *CDC (2)* | 2.31 | 6.93 | BOLD:AAD7726 |
| | | | | | BOLD:ADI4999 |
| | *Cricotopus bicinctus* | *CDC (2)* | 1.86 | 3.2 | BOLD:AAI6018 |
| | | | | | BOLD:AAT9677 |
| | *Gymnometriocnemus brumalis* | *CDC (2)* | 0.5 | 2.41 | BOLD:ACD4501 |
| | | | | | BOLD:ACU9207 |
| | *Limnophyes natalensis* | *CDC (2)* | 1.51 | 2.89 | BOLD:AAB7361 |
| | | | | | BOLD:ACT1270 |
| | *Limnophyes* sp. 4SW | *CDC (2)* | 1.49 | 4.03 | BOLD:ACR9428 |
| | | | | | BOLD:ACU4225 |
| | *Mesosmittia flexuella* | *CDC (2)* | 0.79 | 2.02 | BOLD:ADE7569 |
| | | | | | BOLD:ACU4856 |
| | *Orthocladius fuscimanus* | *CDC (2)* | 2 | 2.66 | BOLD:AAV5075 |
| | | | | | BOLD:ACX3046 |
| | *Parametriocnemus stylatus* | *CDC (2)* | 0.76 | 2.03 | BOLD:AAI2687 |
| | | | | | BOLD:ACT9205 |
| | *Paraphaenocladius exagitans* | *CDC (3)* | 2.54 | 5.88 | BOLD:AAE3719 |
| | | | | | BOLD:ACQ4724 |
| | | | | | BOLD:ACT8523 |
| | *Paraphaenocladius impensus* | *CDC (4)* | 6.85 | 11.99 | BOLD:AAC4200 |
| | | | | | BOLD:ACT2714 |
| | | | | | BOLD:ACT5784 |
| | | | | | BOLD:ACU4175 |
| | *Paratanytarsus laccophilus* | *CDC (2)* | 2.09 | 3.14 | BOLD:AAC8842 |
| | | | | | BOLD:ACF2457 |
| | *Polypedilum convictum* | *CDC (2)* | 2.45 | 4.61 | BOLD:AAW4661 |
| | | | | | BOLD:ACT9278 |
| | *Smittia reissi* | *CDC (2)* | 1.72 | 3.47 | BOLD:ACS9748 |
| | | | | | BOLD:ACU4112 |
| Conopidae | *Myopa testacea* | *CDC (2)* | 3.72 | 3.72 | BOLD:AAK8836 |
| | | | | | BOLD:AAK8838 |
| Dolichopodidae | *Microphor anomalus* | *CDC (2)* | 5.47 | 11 | BOLD:ACH9042 |
| | | | | | BOLD:ACH9043 |
| | *Microphor holosericeus* | *CDC (2)* | 4.06 | 12.7 | BOLD:ACB6469 |
| | | | | | BOLD:ACH6989 |
| Empididae | *Hemerodromia adulatoria* | *CDC (2)* | 8.52 | 8.52 | BOLD:ACJ6728 |
| | | | | | BOLD:ACJ6729 |
| | *Kowarzia barbatula* | *CDC (2)* | 7.21 | 10.71 | BOLD:ACJ6935 |
| | | | | | BOLD:ACJ7236 |
| | *Kowarzia tenella* | *CDC (2)* | 5.39 | 10.8 | BOLD:ACJ6935 |
| | | | | | BOLD:ACJ7236 |

(Continues)

**TABLE 2** (Continued)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Ephydridae | *Allotrichoma laterale* | *CDC (2)* | 6.44 | 6.44 | BOLD:ABA8753 |
| | | | | | BOLD:ACF1575 |
| | *Ditrichophora fuscella* | *CDC (2)* | 3.81 | 7.62 | BOLD:ABA8605 |
| | | | | | BOLD:ABA8606 |
| | *Ditrichophora palliditarsis* | *CDC (2)* | 3.87 | 6.57 | BOLD:AAX8675 |
| | | | | | BOLD:ABA8748 |
| | *Halmopota salinarius* | *CDC (2)* | 2.43 | 3.81 | BOLD:ABA7826 |
| | | | | | BOLD:ABA7827 |
| | *Hydrellia flaviceps* | *CDC (2)* | 4.22 | 6.33 | BOLD:ABA8652 |
| | | | | | BOLD:ABV8173 |
| | *Philygria flavipes* | *CDC (2)* | 1.19 | 2.03 | BOLD:ABA8663 |
| | | | | | BOLD:ACK3229 |
| | *Polytrichophora duplosetosa* | *CDC (2)* | 2.05 | 4.11 | BOLD:ABA8627 |
| | | | | | BOLD:ABA8628 |
| | *Scatella obsoleta* | *CDC (2)* | 1.25 | 2.5 | BOLD:ABA7493 |
| | | | | | BOLD:ABA7494 |
| | *Scatophila signata* | *CDC (2)* | 3.3 | 3.3 | BOLD:ABA7651 |
| | | | | | BOLD:ABA7652 |
| Fanniidae | *Fannia postica* | *CDC (2)* | 2.35 | 7.03 | BOLD:ABW2012 |
| | | | | | BOLD:ACG3518 |
| Heleomyzidae | *Heleomyza serrata* | *CDC (2)* | 0.37 | 3.54 | BOLD:ABX8716 |
| | | | | | BOLD:ACV1127 |
| Lauxaniidae | *Minettia longipennis* | *CDC (2)* | 0.96 | 1.45 | BOLD:ACR0546 |
| | | | | | BOLD:ACR0548 |
| Limoniidae | *Chionea lutescens* | *CDC (2)* | 1.1 | 1.1 | BOLD:ABV5195 |
| | | | | | BOLD:ADD1050 |
| | *Euphylidorea meigenii* | *CDC (2)* | 1.91 | 4.88 | BOLD:ABV4905 |
| | | | | | BOLD:ACU9122 |
| Milichiidae | *Phyllomyza equitans* | *CDC (2)* | 1.39 | 4.05 | BOLD:ACB3455 |
| | | | | | BOLD:ACD3072 |
| Muscidae | *Helina evecta* | *CDC (3)* | 1.83 | 4.27 | BOLD:AAE3133 |
| | | | | | BOLD:ACB3279 |
| | | | | | BOLD:ADB5997 |
| | *Mydaea humeralis* | *CDC (2)* | 1.95 | 5.84 | BOLD:AAE0058 |
| | | | | | BOLD:ACD1934 |

(Continues)

**TABLE 2** (Continued)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Mycetophilidae | *Boletina dispecta* | *CDC (3)* | 9.01 | 11.2 | BOLD:AAY5579 |
| | | | | | BOLD:AAY5580 |
| | | | | | BOLD:AAY5581 |
| | *Brevicornu griseicolle* | CDC(2) | 9.06 | 13.6 | BOLD:ACU9474 |
| | | | | | BOLD:ABA1563 |
| | *Brevicornu sericoma* | *CDC (2)* | 1.99 | 4.58 | BOLD:AAY6368 |
| | | | | | BOLD:ABA1564 |
| | *Phronia obtusa* | *CDC (2)* | 0.83 | 1.18 | BOLD:AAY8505 |
| | | | | | BOLD:ACJ2989 |
| | *Stigmatomeria crassicornis* | *CDC (2)* | 0.56 | 1.86 | BOLD:AAY6370 |
| | | | | | BOLD:ACU7541 |
| | *Zygomyia angusta* | *CDC (3)* | 3.29 | 14.88 | BOLD:AAY5526 |
| | | | | | BOLD:AAY5527 |
| | | | | | BOLD:ABW0168 |
| | *Zygomyia valida* | *CDC (2)* | 9.51 | 14.5 | BOLD:AAY5526 |
| | | | | | BOLD:ABW0168 |
| Pallopteridae | *Toxoneura aff. modesta* | *CDC (2)* | 3.41 | 5.13 | BOLD:ACB4053 |
| | | | | | BOLD:ACV1580 |
| Phoridae | *Megaselia consetigera* | *CDC (2)* | 0.65 | 2.63 | BOLD:ACG2938 |
| | | | | | BOLD:ACX1476 |
| | *Megaselia glabrifrons* | *CDC (2)* | 0.66 | 1.78 | BOLD:ACG3433 |
| | | | | | BOLD:ACI6910 |
| | *Megaselia longicostalis* | *CDC (3)* | 1.32 | 5.72 | BOLD:AAG3263 |
| | | | | | BOLD:ADA4916 |
| | | | | | BOLD:AAG7025 |
| | *Megaselia lutea* | *CDC (2)* | 2.14 | 6.46 | BOLD:AAG3351 |
| | | | | | BOLD:ACG3608 |
| | *Megaselia nigriceps* | *CDC (3)* | 0.76 | 7.16 | BOLD:AAG7022 |
| | | | | | BOLD:AAY6384 |
| | | | | | BOLD:ACF7950 |
| | *Megaselia pulicaria complex* | *CDC (3)* | 5.85 | 11.96 | BOLD:AAL9073 |
| | | | | | BOLD:AAP4698 |
| | | | | | BOLD:AAU8534 |
| | *Megaselia rufa* | *CDC (2)* | 1.83 | 8.31 | BOLD:ACD9573 |
| | | | | | BOLD:ACD9606 |
| | *Megaselia ruficornis* | *CDC (2)* | 5.46 | 17.53 | BOLD:ACF7708 |
| | | | | | BOLD:ACG4585 |
| | *Megaselia sepulchralis* | *CDC (2)* | 2.27 | 4.27 | BOLD:ACF7622 |
| | | | | | BOLD:ACZ9853 |
| | *Megaselia subpalpalis* | *CDC (2)* | 1.05 | 2.17 | BOLD:AAL9083 |
| | | | | | BOLD:ACZ7449 |
| | *Megaselia tarsella* | *CDC (3)* | 0.45 | 5.61 | BOLD:ACE0332 |
| | | | | | BOLD:ACF7226 |
| Psychodidae | *Psychoda nr. albipennis* | *CDC (2)* | 1.55 | 3.45 | BOLD:ABA0876 |
| | | | | | BOLD:ACN5049 |

(Continues)

**TABLE 2** (Continued)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Rhinophoridae | *Rhinomorinia sarcophagina* | *CDC (2)* | 0.75 | 1.78 | BOLD:ACD9526 |
| | | | | | BOLD:ACG3259 |
| Sciaridae | *Bradysia brevispina* | *CDC (2)* | 2.86 | 8.4 | BOLD:ACE4845 |
| | | | | | BOLD:ACI5443 |
| | *Bradysia inusitata* | *CDC (2)* | 6.61 | 6.61 | BOLD:ACE7273 |
| | | | | | BOLD:ACH4332 |
| | *Bradysia praecox* | *CDC (2)* | 1.09 | 2.35 | BOLD:ACF3561 |
| | | | | | BOLD:ACU9870 |
| | *Bradysia regularis* | *CDC (2)* | 0.1 | 1.67 | BOLD:ACC1391 |
| | | | | | BOLD:ACQ7807 |
| | *Bradysia tilicola* | *CDC (2)* | 2.87 | 6.03 | BOLD:AAN6444 |
| | | | | | BOLD:ACP0919 |
| | *Bradysia trivittata* | *CDC (2)* | 0.57 | 3.57 | BOLD:AAH3947 |
| | | | | | BOLD:ACB1143 |
| | *Bradysiopsis vittata* | *CDC (2)* | 2.24 | 4.62 | BOLD:ACC1999 |
| | | | | | BOLD:ACR0949 |
| | *Corynoptera grothae* | *CDC (2)* | 4.75 | 9.36 | BOLD:ACK0158 |
| | | | | | BOLD:ACO7236 |
| | *Corynoptera luteofusca* | *CDC (2)* | 8.16 | 11.8 | BOLD:ACJ1951 |
| | | | | | BOLD:ACQ8494 |
| | *Corynoptera polana* | *CDC (2)* | 1.95 | 3.81 | BOLD:ACF6941 |
| | | | | | BOLD:ACF7764 |
| | *Corynoptera subtilis* | *CDC (2)* | 2.91 | 6.26 | BOLD:ACD5314 |
| | | | | | BOLD:ACT9420 |
| | *Corynoptera tetrachaeta* | *CDC (2)* | 4.16 | 4.16 | BOLD:ACG5327 |
| | | | | | BOLD:ACL4032 |
| | *Corynoptera tridentata* | *CDC (2)* | 9.95 | 9.95 | BOLD:ACJ1561 |
| | | | | | BOLD:ACJ9791 |
| | *Epidapus atomarius* | *CDC (2)* | 0.07 | 3.98 | BOLD:ACD4767 |
| | | | | | BOLD:ACX3063 |
| | *Leptosciarella fuscipalpa* | *CDC (2)* | 5.24 | 9.24 | BOLD:ACE2641 |
| | | | | | BOLD:ACQ8733 |
| | *Leptosciarella scutellata* | CDC (3) | 4.84 | 7.98 | BOLD:ACD6061 |
| | | | | | BOLD:ACG4078 |
| | | | | | BOLD:ACI9623 |
| | *Pnyxiopsis degener* | *CDC (2)* | 1.83 | 5.17 | BOLD:ACE2293 |
| | | | | | BOLD:ACF9729 |
| | *Scatopsciara neglecta* | *CDC (2)* | 0.53 | 1.78 | BOLD:ACC7986 |
| | | | | | BOLD:ACQ2637 |
| | *Scatopsciara subciliata* | *CDC (2)* | 1.93 | 4.32 | BOLD:AAH4004 |
| | | | | | BOLD:ACA8369 |
| | *Sciara hemerobioides* | *CDC (2)* | 4.1 | 4.1 | BOLD:ACQ8933 |
| | | | | | BOLD:ACR4627 |
| | *Trichosia morio* | *CDC (2)* | 0.78 | 3.99 | BOLD:ACD5342 |
| | | | | | BOLD:ACO9950 |

(Continues)

**TABLE 2** (Continued)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Simuliidae | *Simulium cryophilum* | *CDC (2)* | 1.49 | 3.14 | BOLD:ACU9243 |
| | | | | | BOLD:AAU1818 |
| Sphaeroceridae | *Opacifrons coxata* | *CDC (2)* | 6.41 | 14 | BOLD:ACP2618 |
| | | | | | BOLD:ACP5793 |
| | *Spelobia clunipes* | *CDC (2)* | 2.89 | 6.93 | BOLD:AAG7312 |
| | | | | | BOLD:ACF9400 |
| Syrphidae | *Cheilosia albipila* | *CDC (2)* | 2.51 | 6.88 | BOLD:AAW3610 |
| | | | | | BOLD:AAZ1026 |
| | *Cheilosia chrysocoma* | *CDC (2)* | 3.69 | 3.69 | BOLD:ABY6892 |
| | | | | | BOLD:ACJ5068 |
| | *Cheilosia derasa* | *CDC (2)* | 0.58 | 3.47 | BOLD:AAY9044 |
| | | | | | BOLD:AAW3649 |
| | *Cheilosia flavipes* | *CDC (2)* | 8.79 | 8.79 | BOLD:AAW3610 |
| | | | | | BOLD:AAY9045 |
| | *Cheilosia impressa* | *CDC (2)* | 1.95 | 5.74 | BOLD:AAW3651 |
| | | | | | BOLD:AAW3615 |
| | *Cheilosia lenis* | *CDC (2)* | 3.85 | 7.86 | BOLD:AAY8876 |
| | | | | | BOLD:AAY8875 |
| | *Cheilosia mutabilis* | *CDC (2)* | 1.94 | 2.74 | BOLD:AAY9746 |
| | | | | | BOLD:AAY9747 |
| | *Cheilosia personata* | *CDC (2)* | 1.35 | 1.88 | BOLD:ACH1700 |
| | | | | | BOLD:ACX0819 |
| | *Cheilosia proxima* | *CDC (3)* | 3.28 | 6.91 | BOLD:AAW3607 |
| | | | | | BOLD:AAW3651 |
| | | | | | BOLD:ABY8734 |
| | *Cheilosia vernalis-agg.* | *CDC (2)* | 2.07 | 3.84 | BOLD:ACF0974 |
| | | | | | BOLD:ACJ5218 |
| | *Eupeodes nitens* | *CDC (2)* | 3.97 | 3.97 | BOLD:AAB2384 |
| | | | | | BOLD:ACH1529 |
| | *Melanogaster nuda* | *CDC (2)* | 0.81 | 2.44 | BOLD:AAY8880 |
| | | | | | BOLD:ACH5745 |
| | *Merodon rufus* | *CDC (2)* | 0.68 | 1.09 | BOLD:ADI8358 |
| | | | | | BOLD:AAQ1380 |
| | *Paragus pecchiolii* | *CDC (2)* | 0.96 | 4.86 | BOLD:ABA3664 |
| | | | | | BOLD:ACG8255 |
| | *Parasyrphus punctulatus* | *CDC (2)* | 1.11 | 2.65 | BOLD:AAZ4514 |
| | | | | | BOLD:ACG4772 |
| | *Pipiza noctiluca* | *CDC (2)* | 1.54 | 3.92 | BOLD:AAL4100 |
| | | | | | BOLD:ACG4983 |
| | *Platycheirus albimanus* | *CDC (2)* | 0.37 | 3.01 | BOLD:AAL7898 |
| | | | | | BOLD:ACJ4919 |
| | *Sericomyia lappona* | *CDC (2)* | 2.06 | 3.9 | BOLD:AAB1553 |
| | | | | | BOLD:ACH1641 |

(Continues)

**TABLE 2** (Continued)

| Family | Species | CDC rank | Mean intraspecific variation | Max. intraspecific variation | BIN |
|---|---|---|---|---|---|
| Tabanidae | *Tabanus bromius* | *CDC (2)* | 2.04 | 2.93 | BOLD:AAF3864 |
| | | | | | BOLD:ACJ5745 |
| | *Tabanus glaucopis* | | 3.27 | 4.43 | BOLD:AAF3858 |
| | | | | | BOLD:AAF3859 |
| Tachinidae | *Actia dubitata* | *CDC (2)* | 2.36 | 2.36 | BOLD:ACP3766 |
| | | | | | BOLD:ACH1972 |
| | *Bessa selecta* | *CDC (2)* | 1.45 | 2.38 | BOLD:ADK1760 |
| | | | | | BOLD:AAW3422 |
| | *Cyzenis albicans* | *CDC (2)* | 1.18 | 2.18 | BOLD:ACB0896 |
| | | | | | BOLD:ACM9631 |
| | *Kirbya moerens* | *CDC (2)* | 1.22 | 1.86 | BOLD:ACJ2730 |
| | | | | | BOLD:ACB0261 |
| | *Peribaea fissicornis* | *CDC (2)* | 2.22 | 8.17 | BOLD:ACH1961 |
| | | | | | BOLD:ACJ2910 |
| | *Phorinia aurifrons* | *CDC (2)* | 3.76 | 11.2 | BOLD:ADK4076 |
| | | | | | BOLD:ACB0795 |

inaccessible because of the lack of specialists. By contrast, within just a few years, this study provided an interim taxonomic identification system for half of the German Diptera fauna. Although half these species still lack a Linnean name, their BIN assignments are useful "taxonomic handles" for work in ecology, conservation biology and other biodiversity research (see Geiger, Moriniere, et al., 2016). The study demonstrates the efficiency of DNA barcoding in the identification of Central European Diptera, reinforcing the results of earlier studies. DNA barcode coverage was nearly complete for many species-poor families (e.g., Megamerinidae, Opetiidae, Phaeomyiidae) known from Germany and the incidence of "dark taxa" in these families was low. Overall, there was a strong inverse relationship between the number of "dark taxa" and average body size: the smaller the average body size of a family, the higher the ratio of "dark taxa" (Figure 3). Among families with the smallest body sizes, our results suggest a higher incidence of cryptic diversity and overlooked species, indicating the number of dipteran species in Germany is likely to be much higher than previously recognized. Among families, such as the "Iteaphila group" (Empidoidea; see Meyer & Stark, 2015), Milichiidae and Trichoceridae, DNA barcoding indicates unexpectedly high levels of diversity as their BIN count is substantially higher than the number of species known from Germany (Schumann et al., 1999). The Cecidomyiidae represent the most impressive example, as we encountered 930 BINs while only 886 species are known from Germany (Table 1; Jaschhof, 2009; Schumann et al., 1999). As such, they represent by far the largest family of Diptera in the studied area. When compared with the other families in Figure 2b, it is clear that the Cecidomyiidae show a lower average interspecific variation, indicating an increased evolutionary rate. As already proposed by Hebert et al. (2016), the extraordinary species—or BIN number—might be linked to their unusual mode of reproduction, namely haplodiploidy. Here, paternally inherited genomes of diploid males are inactivated during embryogenesis (Normark, 2003). The phenomenon of haplodiploidy is known from Hymenoptera (Branstetter et al., 2018; Hansson & Schmidt, 2018) another group known to be rate accelerated, but it is largely unstudied throughout Diptera. Despite the need for more study, we conclude the true diversity of Diptera in Germany, Europe and the world has been seriously underestimated, a conclusion reached in several other studies (Erwin, 1982; Hebert et al., 2016; May, 1988; Ødegaard, 2000).

Within the metabarcoded Malaise trap samples collected over just one season in one region of Germany, we identified 1,735 OTUs with a sequence identity higher than 97% to a dipteran record. This result indicates that metabarcode analysis of bulk samples will be a valuable approach for assessing the diversity of Diptera in Germany (Appendix S2). Variation in overall biodiversity between sampling sites as well as annual phenologies of certain taxa can easily be visualized using presence–absence maps (Figure 2). This will be a useful feature for comparison of large data sets and for monitoring beneficial or pest insects (L. A. Hardulak et al. in preparation). Although a third of the OTUs within the metabarcoding data set could not be assigned to a Linnean species, interim names, such as BIN assignments, make it possible to compare sampling sites. OTUs with lower sequence similarities (<97%) to known taxa can be used to track "dark taxa," those species missing from the reference sequence library. Although such taxa may only be assigned to a family or genus, their records are still valuable for evaluating differences between samples from various environments or sites. At present, dipteran species, although overall present in very high numbers, are extremely underrepresented within environmental assessments in Germany: ~2,000

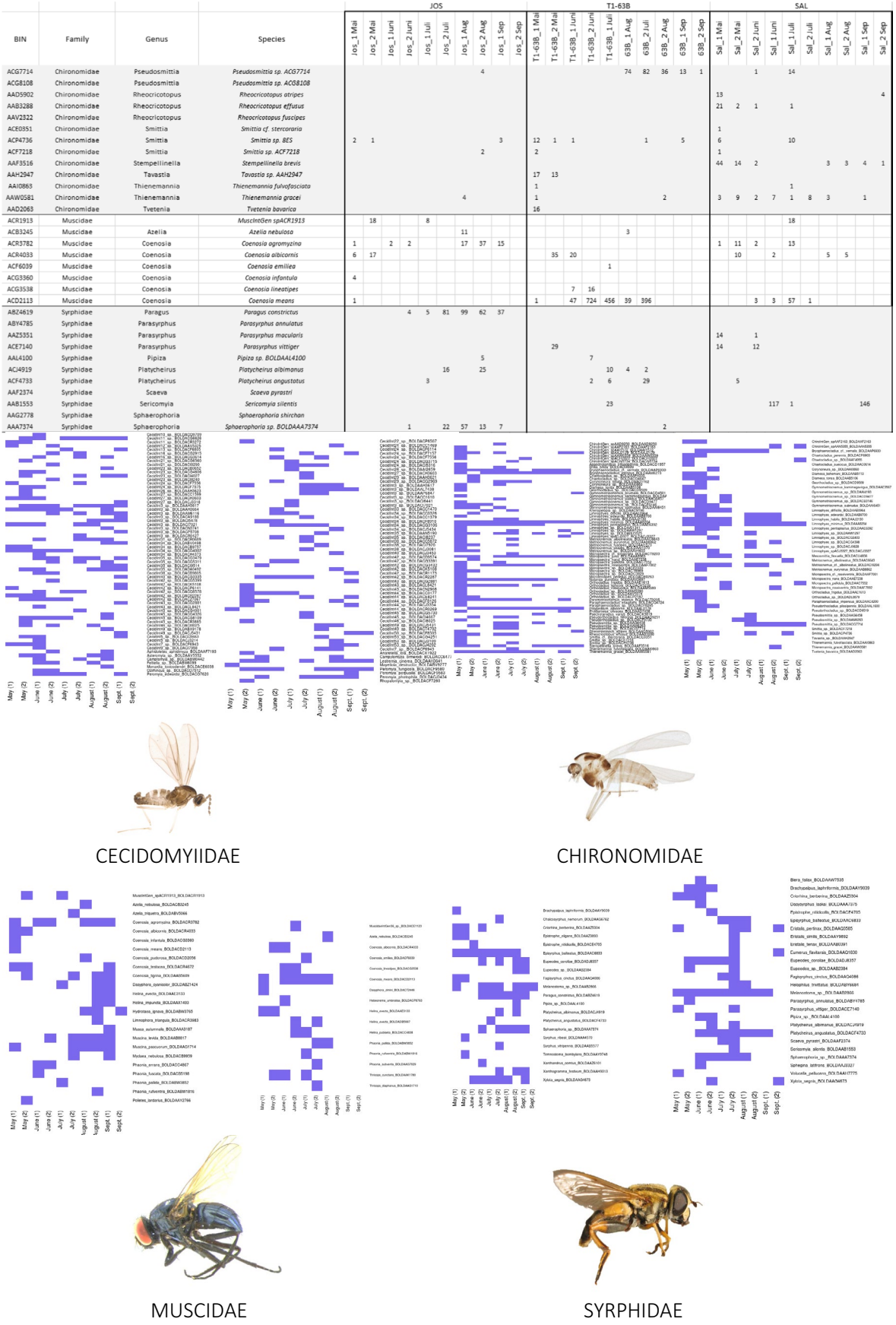


**FIGURE 2** Examples from the metabarcoding results. Presence–absence overviews for three sample sites (Jos, T1-63B and SAL) and illustrative examples for the families Cecidomyiidae, Chironomidae, Muscidae and Syrphidae

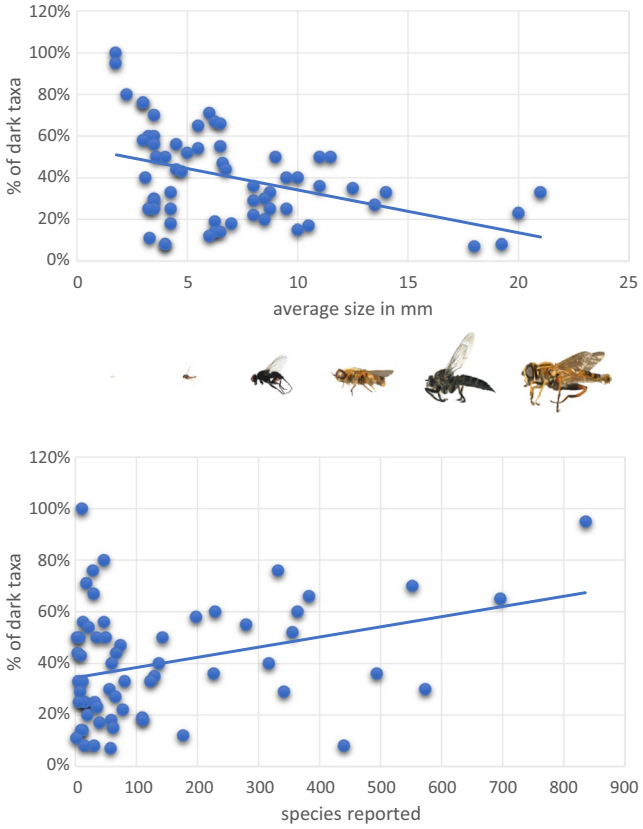


**FIGURE 3** Illustration of the relationship between the percentage of "dark taxa" and average body size (mm), and in number of species reported for a family

species from 11 families (Asilidae, Atelestidae, Ceratopogonidae, Chaoboridae, Dixidae, Dolichopodidae, Empididae, Hybotidae, Psychodidae, Syrphidae, Thaumaleidae) are included in the German red list (Gruttke et al., 2016), but not a single dipteran species is listed among the ~1,000 species being protected according to the European Flora-Fauna-Habitat directive (Council Directive 92/43/EEC on the Conservation of natural habitats and of wild fauna and flora, 1992), which ensures the conservation of a wide range of rare, threatened or endemic animal and plant species in Europe. The present study is a first step to permit the proper evaluation of the status of dipterans and the potential designation of some species as targets for conservation action.

Previous studies have shown the great potential of metabarcoding for biotic assessments in various contexts, including Malaise trap surveys (Morinière et al., 2016), biosurveillance of invasive and pest species (Ashfaq & Hebert, 2016; L. A. Hardulak et al. in prep), macrozoobenthos sampling for assessing water and stream health (Elbrecht & Leese, 2015; Serrana, Miyake, Gamboa, & Watanabe, 2018), faeces analyses for dietary inference (De Barba et al., 2014; Hawlitschek, Fernández-González, Balmori-de la Puente, & Castresana, 2018), species identification for forensic entomology (Chimeno et al., 2018) and for soil biology (Oliverio, Gan, Wickings, & Fierer, 2018). This approach combines the advantages of DNA barcoding, namely the capacity to identify any life stage, body fragment or even trace DNA in the environment, with the ability of high-throughput sequencers to analyse millions of DNA fragments and thousands of specimens at a time. The application of this technology to biodiversity assessments will certainly enable species surveys at larger scales, shorter time and lower costs compared with classical morphological approaches (Douglas et al., 2012; Hajibabaei et al., 2011; Ji et al., 2013; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). The ability to upscale biomonitoring projects is crucial, as is the need to generate biodiversity data fast and with less dependence on often unavailable taxonomic experts. Additionally, data generated by ongoing metabarcoding studies, such as from annual national biomonitoring projects, can be combined and reanalysed, producing recursively more comprehensive species lists, when new reference sequences become available or when taxonomic annotations have been improved. While biomonitoring studies have traditionally employed small subsets of indicator species, metabarcoding will enable comprehensive assessments of biodiversity because even "dark taxa" can be tracked. Furthermore, metabarcoding can enhance the ability to rapidly assess biodiversity patterns to identify regions that are of most significance for conservation.

Although this project aimed to develop a comprehensive DNA barcode library, resource constraints meant that only half the specimens sorted to a family or better taxonomy could be analysed. It is certain that many species and genera currently absent from the reference library remain within this sorted material, making the remaining samples a valuable resource for future extension of the reference library. Our work has also highlighted the potential of DNA barcoding and metabarcoding to aid efforts to conserve the world's fauna. Because these technologies greatly enhance our ability to identify, and thus conserve, biodiversity, they should be pursued—vigorously. As our study has provided several thousands of voucher-based DNA barcode records, we invite the global community of dipteran taxonomists to improve identifications for the many "dark taxa" encountered in our study by identifying these vouchers using reverse taxonomic approaches.

The present study represents an important component of a decade of work directed toward creating a comprehensive DNA barcode library for German animal species. Because Diptera represents the largest and taxonomically most challenging insect order, they have received less attention than other orders (e.g., Lepidoptera, Coleoptera, freshwater orders) with lower species richness and more taxonomic expertise. Our work on Diptera has not only confirmed that this order is extremely species-rich, but also that several of its most diverse families include a large proportion of "dark taxa." The present study represents a cornerstone for subsequent research on these unexplored groups of Diptera. This paper presents the results of one of the most comprehensive studies on DNA barcoding of Diptera, with a coverage of over 80% of German families. Due to the general lack of taxonomy in many groups of Diptera, only a fraction of the specimens could be identified to species level. Most specimens for the study were obtained from just three Malaise traps deployed as a component of the Global Malaise programme (see http://biodiversitygenomics.net/projects/gmp/). Voucher specimens are still being identified by external specialists, a process that

**TABLE 3** All cases of low intraspecific sequence variation at *COI*; cases of BIN sharing (BS)

| Family | Species | BS rank | Mean intraspecific variation | Max intraspecific | BIN |
|---|---|---|---|---|---|
| Anthomyiidae | *Hylemya nigrimana* | *BS (2)* | 0.34 | 0.52 | BOLD:ABA6492 |
| | *Hylemya vagans* | | 0.37 | 1.58 | |
| Calliphoridae | *Calliphora loewi* | *BS (2)* | 1.07 | 1.07 | BOLD:AAB6579 |
| | *Calliphora vicina* | | 0.84 | 2.59 | |
| | *Lucilia caesar* | *BS (3)* | 0.95 | 3.07 | BOLD:AAA7470 |
| | *Lucilia caesarillustris* | | 0.7 | 2.43 | |
| | *Lucilia illustris* | | N/A | 0 | |
| Dolichopodidae | *Medetera petrophiloides* | *BS (2)* | 0.35 | 1.22 | BOLD:ACA1124 |
| | *Medetera truncorum* | | N/A | 0 | |
| | *Sphyrotarsus argyrostomus* | *BS (2)* | 0.91 | 1.37 | BOLD:ADB6106 |
| | *Sphyrotarsus hygrophilus* | | N/A | 0 | |
| Empididae | *Kowarzia madicola* | *BS (2)* | 0 | 0 | BOLD:ACJ7236 |
| | *Kowarzia tenella* | | 5.39 | 10.8 | |
| | *Kowarzia barbatula* | *BS (2)* | 4.8 | 11.3 | BOLD:ACJ6935 |
| | *Kowarzia tenella* | | 5.39 | 10.8 | |
| Ephydridae | *Allotrichoma bezzii* | *BS (4)* | 0.13 | 0.31 | BOLD:ACF1575 |
| | *Allotrichoma filiforme* | | 0.08 | 0.15 | |
| | *Allotrichoma laterale* | | 6.44 | 6.44 | |
| | *Allotrichoma schumanni* | | 0 | 0 | |
| | *Ephydra macellaria* | *BS (3)* | N/A | 0 | BOLD:AAG2729 |
| | *Ephydra murina* | | N/A | 0 | |
| | *Ephydra riparia* | | 2.83 | 2.83 | |
| | *Hydrellia nigricans* | *BS (2)* | 0.23 | 0.31 | BOLD:ABA8624 |
| | *Hydrellia subalbiceps* | | 0.31 | 0.46 | |
| | *Notiphila cinerea* | *BS (2)* | 0.26 | 0.46 | BOLD:ABA7513 |
| | *Notiphila graecula* | | 0 | 0 | |
| | *Notiphila riparia* | *BS (2)* | 0.16 | 0.35 | BOLD:AAX5585 |
| | *Notiphila subnigra* | | 0.41 | 0.62 | |
| | *Philygria flavipes* | *BS (2)* | 1.19 | 2.03 | BOLD:ACK3229 |
| | *Philygria punctatonervosa* | | 0.15 | 0.15 | |
| | *Psilopa compta* | *BS (2)* | 0.08 | 0.16 | BOLD:AAG6948 |
| | *Psilopa nitidula* | | 0.38 | 0.77 | |
| Iteaphila-group | *Anthepiscopus indet.* | *BS (2)* | 0.14 | 0.48 | BOLD:ACD9492 |
| | *Anthepiscopus* sp. 1 | | 11.3 | 11.3 | |
| | *Anthepiscopus* sp. 1 | *BS (2)* | 11.3 | 11.3 | BOLD:ACJ7111 |
| | *Anthepiscopus* sp. 4 | | 0.91 | 1.58 | |
| | *Iteaphila* sp. 1 | *BS (2)* | 0.07 | 0.15 | BOLD:ACD3033 |
| | *Iteaphila* sp. 2 | | 4.49 | 9.77 | |
| Lonchopteridae | *Lonchoptera lutea* | *BS (2)* | 0.39 | 1.09 | BOLD:ABX0277 |
| | *Lonchoptera nitidifrons* | | N/A | 0 | |

(Continues)

**TABLE 3** (Continued)

| Family | Species | BS rank | Mean intraspecific variation | Max intraspecific | BIN |
|---|---|---|---|---|---|
| Muscidae | *Hydrotaea dentipes* | *BS (2)* | 2.15 | 9.78 | BOLD:AAZ9882 |
| | *Hydrotaea similis* | | 0 | 0 | |
| Mycetophilidae | *Boletina gripha* | *BS (2)* | 0.52 | 0.9 | BOLD:AAF6783 |
| | *Boletina groenlandica* | | N/A | 0 | |
| | *Mycetophila distigma* | *BS (2)* | N/A | 0 | BOLD:AAY8340 |
| | *Mycetophila flava* | | 0.19 | 0.19 | |
| | *Zygomyia angusta* | *BS (2)* | 4.6 | 15.4 | BOLD:AAY5526 |
| | *Zygomyia valida* | | 14.5 | 14.5 | |
| | *Zygomyia angusta* | *BS (2)* | 4.6 | 15.4 | BOLD:ABW0168 |
| | *Zygomyia valida* | | 14.5 | 14.5 | |
| Phoridae | *Triphleba bicornuta* | *BS (2)* | N/A | 0 | BOLD:ACF0365 |
| | *Triphleba* sp. BOLD:ACF0365 | | 0.66 | 1.22 | |
| Sarcophagidae | *Sarcophaga depressifrons* | *BS (2)* | 0 | 0 | BOLD:ABV4597 |
| | *Sarcophaga haemorrhoa* | | 0.47 | 0.7 | |
| Simuliidae | *Simulium balcanicum* | *BS (2)* | N/A | 0 | BOLD:AAM4036 |
| | *Simulium equinum* | | 1.59 | 2.66 | |
| Syrphidae | *Baccha elongata* | *BS (6)* | N/A | 0 | BOLD:ABA3006 |
| | *Baccha elongata s.s.* | | 0 | 0 | |
| | *Baccha obscuripennis* | | 1.23 | 2.02 | |
| | *Baccha* sp. BOLDABA3006 | | N/A | 0 | |
| | *Brachypalpus laphriformis* | | 0.56 | 1.54 | BOLD:AAY9039 |
| | *Brachypalpus valgus* | | N/A | 0 | |
| | *Cheilosia albipila* | *BS (2)* | 2.51 | 6.88 | BOLD:AAW3610 |
| | *Cheilosia flavipes* | | 8.79 | 8.79 | |
| | *Cheilosia barbata* | *BS (3)* | 0.1 | 0.3 | BOLD:AAW3615 |
| | *Cheilosia impressa* | | 1.95 | 5.74 | |
| | *Cheilosia* sp. BOLDAAW3615 | | 0 | 0 | |
| | *Cheilosia chloris* | *BS (8)* | 0.57 | 1.42 | BOLD:ACF0974 |
| | *Cheilosia chlorus* | | 0.12 | 0.18 | |
| | *Cheilosia chlorus-group* | | N/A | 0 | |
| | *Cheilosia fraterna* | | 0.55 | 0.87 | |
| | *Cheilosia melanura* | | 0.06 | 0.2 | |
| | *Cheilosia ruficollis* | | N/A | 0 | |
| | *Cheilosia* sp. BOLDACF0974 | | 0.47 | 0.71 | |
| | *Cheilosia vernalis-agg.* | | 2.07 | 3.84 | |

(Continues)

**TABLE 3** (Continued)

| Family | Species | BS rank | Mean intraspecific variation | Max intraspecific | BIN |
|---|---|---|---|---|---|
| | *Cheilosia crassiseta* | *BS (6)* | N/A | 0 | BOLD:AAW3647 |
| | *Cheilosia impudens* | | N/A | 0 | |
| | *Cheilosia nigripes* | | N/A | 0 | |
| | *Cheilosia* sp. BIOUG17085-G07 | | 0.75 | 1.94 | |
| | *Cheilosia aff. grisella* | | N/A | 0 | |
| | *Cheilosia antiqua* | | N/A | 0 | |
| | *Cheilosia faucis* | *BS (2)* | 0.7 | 0.88 | BOLD:AAY8874 |
| | *Cheilosia nivalis* | | 0 | 0 | |
| | *Cheilosia grisella* | *BS (2)* | 0.18 | 0.18 | BOLD:AAW3619 |
| | *Cheilosia pubera* | | 0.49 | 0.87 | |
| | *Cheilosia canicularis* | *BS (2)* | 0.08 | 0.38 | BOLD:ACI2500 |
| | *Cheilosia montana* | | N/A | 0 | |
| | *Cheilosia carbonaria* | *BS (2)* | 0.37 | 0.37 | BOLD:AAY8876 |
| | *Cheilosia lenis* | | 3.85 | 7.86 | |
| | *Chrysotoxum bicinctum* | *BS (2)* | 0.86 | 2 | BOLD:AAJ0967 |
| | *Chrysotoxum festivum* | | 0 | 0 | |
| | *Dasysyrphus hilaris* | *BS (3)* | 0.35 | 0.52 | BOLD:AAA7375 |
| | *Dasysyrphus laskai* | | 0.3 | 0.3 | |
| | *Dasysyrphus venustus* | | N/A | 0 | |
| | *Dasysyrphus lenensis* | *BS (3)* | 0.58 | 0.58 | BOLD:AAB2865 |
| | *Dasysyrphus pinastri* | | 1.25 | 2.1 | |
| | *Dasysyrphus* sp. BOLDAAB2865 | | 0.12 | 0.17 | |
| | *Eupeodes bucculatus* | *BS (5)* | 1.14 | 3.13 | BOLD:AAB2384 |
| | *Eupeodes nielseni* | | 0.15 | 0.37 | |
| | *Eupeodes nitens* | | 3.97 | 3.97 | |
| | *Eupeodes* sp. BOLDAAB2384 | | 0.39 | 1.03 | |
| | *Eupeodes luniger* | | 0.53 | 1.05 | |
| | *Melanogaster aerosa* | *BS (2)* | N/A | 0 | BOLD:AAQ4015 |
| | *Melanogaster hirtella* | | 0.26 | 0.7 | |
| | *Melanostoma dubium* | *BS (7)* | 0 | 0 | BOLD:AAB2866 |
| | *Melanostoma mellinum* | | 0.58 | 1.21 | |
| | *Melanostoma mellinum-agg.* | | N/A | 0 | |
| | *Melanostoma scalare* | | 0.49 | 1.3 | |
| | *Melanostoma* sp. A | | 0 | 0 | |
| | *Melanostoma* sp. B | | 0.11 | 0.16 | |
| | *Melanostoma* sp. BOLDAAB2866 | | 0.63 | 2.69 | |
| | *Merodon avidus* | *BS (2)* | N/A | 0 | BOLD:AAQ1379 |
| | *Merodon avidus* B | | 0.55 | 1.03 | |

(Continues)

**TABLE 3** (Continued)

| Family | Species | BS rank | Mean intraspecific variation | Max intraspecific | BIN |
|---|---|---|---|---|---|
| | *Paragus aff. haemorrhous* | *BS (5)* | N/A | 0 | BOLD:ABZ4619 |
| | *Paragus constrictus* | | N/A | 0 | |
| | *Paragus haemorrhous* | | 0.26 | 0.87 | |
| | *Paragus* sp. BOLDABZ4619 | | 0.07 | 0.37 | |
| | *Paragus tibialis* | | N/A | 0 | |
| | *Paragus majoranae* | *BS (2)* | 0.87 | 0.87 | BOLD:ABA3664 |
| | *Paragus pecchiolii* | | 0.96 | 4.86 | |
| | *Parasyrphus lineola* | *BS (2)* | 0.19 | 0.39 | BOLD:ACE7140 |
| | *Parasyrphus vittiger* | | 0.63 | 1.44 | |
| | *Pipiza bimaculata* | *BS (4)* | N/A | 0 | BOLD:AAL4100 |
| | *Pipiza nocticula* | | N/A | 0 | |
| | *Pipiza noctiluca-agg.* | | N/A | 0 | |
| | *Pipiza* sp. BOLDAAL4100 | | 0.55 | 1.65 | |
| | *Platycheirus angustatus* | *BS (3)* | 0.84 | 2.02 | BOLD:ACF4733 |
| | *Platycheirus europaeus* | | 1.95 | 1.95 | |
| | *Platycheirus* sp. BOLDACF4733 | | 0.21 | 1.15 | |
| | *Platycheirus clypeatus* | *BS (5)* | 0.38 | 0.88 | BOLD:AAA9506 |
| | *Platycheirus fulviventris* | | 1.04 | 1.04 | |
| | *Platycheirus occultus* | | 0.51 | 1.04 | |
| | *Platycheirus perpallidus* | | N/A | 0 | |
| | *Platycheirus* sp. BOLDAAA9506 | | 0.9 | 2.03 | |
| | *Platycheirus melanopsis* | *BS (2)* | 0.25 | 0.62 | BOLD:AAP0412 |
| | *Platycheirus tatricus* | | N/A | 0 | |
| | *Platycheirus nielseni* | *BS (3)* | 0 | 0 | BOLD:AAC6630 |
| | *Platycheirus peltatus* | | 0.24 | 0.72 | |
| | *Platycheirus peltatus-group* | | N/A | 0 | |
| | *Platycheirus scutatus* | *BS (3)* | 0.05 | 0.19 | BOLD:AAG4665 |
| | *Platycheirus scutatus-group* | | 0.44 | 0.71 | |
| | *Platycheirus splendidus* | | N/A | 0 | |
| | *Scaeva dignota* | *BS (2)* | N/A | 0 | BOLD:AAF2374 |
| | *Scaeva pyrastri* | | 0.25 | 0.91 | |
| | *Scaeva pyrastri* | *BS (2)* | 0.25 | 0.91 | BOLD:AAF2374 |
| | *Scaeva dignota* | | N/A | 0 | |
| | *Sericomyia lappona* | *BS (2)* | 2.06 | 3.9 | BOLD:AAB1553 |
| | *Sericomyia silentis* | | 0.05 | 0.24 | |

(Continues)

**TABLE 3** (Continued)

| Family | Species | BS rank | Mean intraspecific variation | Max intraspecific | BIN |
|---|---|---|---|---|---|
| | *Sphaerophoria bankowskae* | *BS (9)* | N/A | 0 | BOLD:AAA7374 |
| | *Sphaerophoria infuscata* | | 0.24 | 0.38 | |
| | *Sphaerophoria interrupta* | | 0 | 0 | |
| | *Sphaerophoria interrupta-group* | | 0.49 | 0.75 | |
| | *Sphaerophoria philanthus* | | N/A | 0 | |
| | *Sphaerophoria rueppellii* | | N/A | 0 | |
| | *Sphaerophoria* sp. BOLDAAA7374 | | 0.31 | 6.54 | |
| | *Sphaerophoria taeniata* | | N/A | 0 | |
| | *Sphaerophoria virgata* | | N/A | 0 | |
| | *Sphegina montana* | *BS (2)* | N/A | 0 | BOLD:ABX4867 |
| | *Sphegina sibirica* | | 0.4 | 0.41 | |
| | *Temnostoma apiforme* | *BS (2)* | 0.52 | 0.52 | BOLD:AAV6543 |
| | *Temnostoma meridionale* | | 0.35 | 0.52 | |
| Stratiomyidae | *Beris geniculata* | *BS (2)* | N/A | 0 | BOLD:AAW3384 |
| | *Beris morrisii* | | 0.48 | 1.47 | |
| Tachinidae | *Lydella stabulans* | *BS (2)* | 0.12 | 0.44 | BOLD:AAP8653 |
| | *Lydella thompsoni* | | 0.68 | 1.31 | |
| | *Medina luctuosa* | *BS (3)* | 1.35 | 1.35 | BOLD:AAG6902 |
| | *Medina melania* | | | | |

is labour intensive and time consuming, especially for taxonomically challenging taxa.

Our study presents results from one of the most comprehensive DNA barcoding projects on Diptera, a megadiverse, and, almost certainly, most diverse insect order. Our results strongly support the conclusion that DNA barcoding will enable the discovery and identification of most dipteran taxa. Some cases of low interspecific variation were observed in the Syrphidae, Tachinidae and Calliphoridae where additional markers may be needed for species identification (Haarto & Ståhls, 2014; Nelson et al., 2012; Pohjoismäki et al., 2016; Whitworth et al., 2007). However, in most cases, there was congruence between BINs and species defined by traditional morphological methods, supporting the use of DNA barcoding as a species identification tool for Diptera. This conclusion and the finding that many of the species we encountered represent "dark taxa" indicates that DNA barcoding will speed the discovery of genetic entities that will eventually gain recognition as biological species. Our data release aims at making these results accessible to the scientific community through a public data portal so they will be available for taxonomic research, biodiversity studies and barcoding initiatives at national and international levels.

In summary, the application of DNA barcoding enabled a comprehensive assessment of German Diptera, including several highly diverse families, which would otherwise have been excluded due to a lack of taxonomic expertise. By selecting morphospecies from the pool of specimens collected by the year-long deployment of Malaise traps in ecosystems ranging from alpine to lowland settings, we constructed a reference library for most dipteran families known from Germany. Due to the diversity of sampling sites, we encountered a wide range of taxa from microendemics to wide-ranging generalists with varied seasonal phenologies. We emphasize that DNA barcoding and the resultant barcode reference libraries provide an easy, intuitive introduction to molecular genetics, an approach accessible to undergraduate students in a way that genome sequencing is not. Because DNA barcoding workflows have been implemented in many laboratories around the world and because current primer sets reliably generate amplicons, this method is ideal for educational purposes. Democratization of the method, the analytical tools and data through the BOLD database (Ratnasingham & Hebert, 2007) further facilitates its use in real world situations. The approach has the additional advantage of allowing students to not only work with "real organisms," but also to solve long-standing taxonomic puzzles.

The latter work leads students to probe the historical literature, to regale in past expeditions in search of type locations or type material, and potentially to end the chase by describing a new species. However, it is critical that senior taxonomists and professors need to recognize these possibilities and encourage their students to embrace this approach as it offers such a clear solution to the taxonomic impediment.

Germany has a tradition of more than 250 years of entomological research, and the number of Diptera species recorded is the highest for any European country comprising almost half of the European fauna. Despite this long effort, knowledge of its Diptera fauna must be regarded as fragmentary. In accordance with the species accumulation curve presented by Pape (2009) for the British Isles, additional species were revealed from current collecting efforts for practically every species-rich family. Recording "new" species is slowed by the lack of experts for many of these families as well as by the lack of up-to-date identification keys. A particularly important result of our study is that the estimated number of dipteran species in Germany is certainly much higher than formerly thought. High proportions of unrecorded species were evident for the Agromyzidae, Anthomyiidae, Cecidomyiidae, Ceratopogonidae, Chironomidae, Chloropidae, Phoridae, Sciaridae and Sphaeroceridae, and to a lesser extent for the Empidoidea, Limoniidae, Mycetophilidae and others. Further studies point to an enormous under-estimation of the species diversity in the Cecidomyiidae (Borkent et al., 2018; Hebert et al., 2016). Although our data do not allow for an accurate projection for the size of the total species numbers, it seems quite likely that this single family contains thousands of unrecorded species in Germany.

## ACKNOWLEDGEMENTS

We express our extreme gratitude to the taxonomists, citizen scientists and nature enthusiasts who supported this campaign by collecting thousands of dipteran species. The realization of this mammoth task would not have been possible without the help of Adaschkiewitz, W., Assum, Babiy, P. P., Bährmann, R., Baranov, V., Beermann, A., Behounek, G., Belleuvre, N., Blick, T., Bolz, R., Brandt-Floren, C., Brenzinger, S., Brown, A., Burmeister E. G., Charabidze, D., Chimeno, C., Claussen, C., Dettinger-Klemm, A., Diller, E., Drozd, P., Dunz, A., Duschl Miesbach, M., Dworschak, W., Eckert, I., Esser, J., Fahldieck, M., Fiedler, Fittkau, E. J. (+), Flügel, H. J.,Forster, W., Forstner, P., Fünfstück, J., Fütterer, S., Fuhrmann, S., Gabriel, I., Gammelmo, O., Gerecke, R., Glaw, F., Guggemoos, T., , Haberberger, S., Hable, J., Haeselbarth, E., Hansen, L. O., Hartop, E., Hawlitschek, O., Heller, K., Hessing, R., Hierlmeier, V., Hilbig, D., Höglund, J. R., Höhne, F., Honold, D., Jaschof, M., Jon, T., Kamin, J., Kappert, J., Kehlmaier, C., Kilian, D., Kirsch, H., Kjaerandsen, J., Kleiner, M., Koehler, F., Koehler, J., Kölbl, N., Koenig, T., Kolbeck, H., Kraus, G., Kraus, W., Kuehbandner, M., Kuehlhorn, F., Kuhlmann, M., Kusdas, K., Kvifte, G. M., Lindner, S., Loennve, O. J., Lucas, W., Mair, K., Mandery, K., Mengual, X., Merkel-Wallner, G., Mortelmans, J., Müller, Mueller, R., Mueller-Kroehling, S., Neumann, C., Olberg, S., Olsen, K. M., Pavlova, A., Pötter, L., Steven, M., Plassmann, E., Podhorna, J., Prescher, S., Prozeller, M., Pushkar, V., Reckel, F., Rehm, T., Reiff, N., Reimann, A., Reiso, S., Rennwald, K., Richter, B., Riedel, G., Rohrmoser, S., Rozo, P., Rudzinksi, H. G., Ruf, T., Salomon, C., Schacht, W. (+), Schäfer, A., Scheingraber, M., Scheler, Schmieder, F., Schödl, M., Schoenitzer, K., Schrott, S., Schubart, C., Schubert, C., Schubert, W., Schwarz, K., Schwemmer, R., Sedlak, G., Segerer, A., Sellmayer, G., Spelda, J., Spies, M., Ssymank, A., Steigemann, U., Stenhouse, G., Stoecklein, F., Stuke, J. H., Stur, E., Tänzler, K., Tänzler, R., Telfer, A., Toussaint, C., Toussaint, E., Treiber, R., Troester, M., v. Tschirnhaus, M., v. d. Dunk, K., , Vallenduuk, H., van Ess, L., Velterop, J., Voith, J., Volf, M., Wachtel, F., Wagner, R., Warncke, K., Weber, D., Weiffenbach, H., Weigand, A. M., Weixler, K., Wiedenbrug, S., Windmaisser, T., Winqvist, K., Woodley, N., E., Zahn, A. The project was funded by grants from the Bavarian State Ministry of Science and the Arts (2009-2018: Barcoding Fauna Bavarica, BFB) and the German Federal Ministry of Education and Research (German Barcode of Life: 2012-2019, BMBF FKZ 01LI1101 and 01LI1501). We are grateful to the team at the Centre for Biodiversity Genomics in Guelph (Ontario, Canada) for their great support and help and particularly to Sujeevan Ratnasingham for developing the BOLD database (BOLD; www.boldsystems.org) infrastructure and the BIN management tools. The sequencing work was supported, in part, by funding from the Government of Canada to Genome Canada through the Ontario Genomics Institute, while the Ontario Ministry of Research and Innovation and NSERC supported development of the BOLD informatics platform. We also thank all the students who assisted in the ZSM-barcoding projects (barcoding-zsm.de) for picking countless legs and photographing countless specimens. We would like to express our thanks to Dr Vedran Bozicevic (AIM GmbH, Munich, Germany) for assisting with the KRONA file to enable inspection of BIN diversity. Fieldwork permits were issued by the responsible state environmental offices of Bavaria (Bayerisches Staatsministerium für Umwelt und Gesundheit, Munich, Germany, project: "Barcoding Fauna Bavarica"; confirmed by the regional governments "Bezirksregierungen") and Rhineland-Palatinate ("Struktur- und Genehmigungsdirektion Nord", Axel Schmidt [Koblenz, Germany]).

## AUTHOR CONTRIBUTIONS

Obtained funding: G.H., W.W., A.H., P.D.N.H. Collected the samples: D.D., B.R. Conceived and designed the experiments: J.M., L.A.H., M.G., B.R. Analysed the data: J.M., L.A.H., M.F.G., L.R., B.R. Wrote the paper: J.M., L.A.H., S.S., M.B., D.D. Contributed (additions/corrections) to the manuscript: P.D.N.H., A.H., M.F.G., L.H., G.H.

## DATA AVAILABILITY

All specimen data have been made publicly available within the BOLD workbench - a DOI for the dataset has been added.

## ORCID

*Jérôme Morinière* https://orcid.org/0000-0001-9167-6409
*Stefan Schmidt* https://orcid.org/0000-0001-5751-8706

## REFERENCES

Ashfaq, M., & Hebert, P. D. N. (2016). DNA barcodes for bio-surveillance: Regulated and economically important arthropod plant pests. *Genome*, *59*(11), 933–945. https://doi.org/10.1139/gen-2016-0024

Ashfaq, M., Hebert, P. D. N., Mirza, J. H., Khan, A. M., Zafar, Y., & Mirza, M. S. (2014). Analyzing mosquito (Diptera: Culicidae) diversity in Pakistan by DNA barcoding. *PLoS ONE*, *9*(5), e97268. https://doi.org/10.1371/journal.pone.0097268

Astrin, J. J., Höfer, H., Spelda, J., Holstein, J., Bayer, S., Hendrich, L., … Muster, C. (2016). Towards a DNA barcode reference database for spiders and harvestmen of Germany. *PLoS ONE*, *11*(9), e0162624 (24 pp, supplements). https://doi.org/10.1371/journal.pone.0162624

Bickel, D., T. Pape, & R. Meier (Eds.) (2009). *Diptera diversity: status, challenges and tools* (pp. 459). Leiden, Netherlands: Brill.

Borkent, A., Brown, B., Adler, P. H., Amorim, D. D. S., Barber, K., Bickel, D., … Capellari, R. S. (2018). Remarkable fly (Diptera) diversity in a patch of Costa Rican cloud forest. *Zootaxa*, *4402*(1), 53–90.

Branstetter, M. G., Childers, A. K., Cox-Foster, D., Hopper, K. R., Kapheim, K. M., Toth, A. L., & Worley, K. C. (2018). Genomes of the Hymenoptera. *Current Opinion in Insect Science*, *25*, 65–75. https://doi.org/10.1016/j.cois.2017.11.008

Brix, S., Leese, F., Riehl, T., & Kihara, T. C. (2015). A new genus and new species of Desmosomatidae Sars, 1897 (Isopoda) from the eastern South Atlantic abyss described by means of integrative taxonomy. *Marine Biodiversity*, *45*(1), 7–61. https://doi.org/10.1007/s12526-014-0218-3

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.

Chimeno, C., Morinière, J., Podhorna, J., Hardulak, L., Hausmann, A., Reckel, F., … Haszprunar, G. (2018). DNA barcoding in forensic entomology – Establishing a DNA reference library of potentially forensic relevant arthropod species. *Journal of Forensic Sciences*, *64*(2), 593–601. https://doi.org/10.1111/1556-4029.13869

ChivianE., & BernsteinA. (Eds.) (2008). *Sustaining life: How human health depends on biodiversity*. Oxford, UK: Oxford University Press.

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. *Raffles Bulletin of Zoology*, *61*(2), 779–793.

Cruaud, P., Rasplus, J. Y., Rodriguez, L. J., & Cruaud, A. (2017). High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. *Scientific Reports*, *7*, 41948. https://doi.org/10.1038/srep41948

Cywinska, A., Hunter, F. F., & Hebert, P. D. (2006). Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*, *20*(4), 413–424. https://doi.org/10.1111/j.1365-2915.2006.00653.x

Dathe, H. H., & Blank, S. M.. (2004). Nachträge zum Verzeichnis der Hautflügler Deutschlands, Entomofauna Germanica Band 4 (Hymenoptera). (1). *Entomologische Nachrichten Und Berichte*, *48*(3–4), 179–182.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, *14*(2), 306–323. https://doi.org/10.1111/1755-0998.12188

de Carvalho, M. R., Bockmann, F. A., Amorim, D. S., Brandão, C. R. F., de Vivo, M., de Figueiredo, J. L., … 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

DeWaard, J. R., Ivanova, N. V., Hajibabaei, M., & Hebert, P. D. N. (2008). Assembling DNA barcodes. *Environmental Genomics*, *410*, 275–294.

DeWaard, J. R., Levesque-Beaudin, V., deWaard, S. L., Ivanova, N. V., McKeown, J. T. A., Miskie, R., … Hebert, P. D. N. (2019). Expedited assessment of terrestrial arthropod diversity by coupling Malaise traps with DNA barcoding. *Genome*, *62*(3), 85–95. https://doi.org/10.1139/gen-2018-0093

Doczkal, D. (2017). Vorsortierung der Proben und Vollständigkeit der Erfassung. In A. Ssymank, & D. Doczkal (Eds.), *Biodiversität des südwestlichen Dinkelbergrandes und des Rheintals bei Grenzach-Whylen, eine Bestandsaufnahme im südwestlichen Einfallstor Deutschlands für neue Arten in der Folge des Klimawandels*. Mauritiana (Altenburg) 34, 1–910.

Douglas, W. Y., 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

Eiseman, C. S., Heller, K., & Rulik, B. (2016). A new leaf-mining dark-winged fungus gnat (Diptera: Sciaridae), with notes on other insect associates of marsh marigold (Ranunculaceae: *Caltha palustris* L.). *Proceedings of the Entomological Society of Washington*, *118*(4), 519–533.

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., 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. https://doi.org/10.1016/j.ympev.2006.11.021

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

Erwin, T. L. (1982). Tropical forests: Their richness in Coleoptera and other arthropod species. *The Coleopterists Bulletin*, *36*(1), 74–75.

Fontaine, B., van Achterberg, K., Alonso-Zarazaga, M. A., Araujo, R., Asche, M., Aspöck, H., … Bouchet, P. (2012). New species in the Old World: Europe as a frontier in biodiversity exploration, a test bed for 21st Century taxonomy. *PLoS ONE*, *7*(5), e36881. https://doi.org/10.1371/journal.pone.0036881

Fujita, M. K., Leache, A. D., Burbrink, F. T., McGuire, J. A., & Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*, *27*(9), 480–488. https://doi.org/10.1016/j.tree.2012.04.012

Geiger, M. F., Astrin, J. J., Borsch, T., Burkhardt, U., Grobe, P., Hand, R., … Monje, C. (2016). 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*, *59*(9), 661–670.

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*, e10671. https://doi.org/10.3897/BDJ.4.e10671

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., … Hajibabaei, M. (2014). Simultaneous assesssment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8007–8012.

Gruttke, H., Binot-Hafke, M., Balzer, S., Haupt, H., Hofbauer, N., Ludwig, G., & Ries, M. (2016). Rote Liste gefährdeter Tiere, Pflanzen und Pilze Deutschlands. Band 4: Wirbellose Tiere (Teil 2). *Naturschutz Und Biologische Vielfalt*, *70*(4), 598.

Gutiérrez, M. A. C., Vivero, R. J., Vélez, I. D., Porter, C. H., & Uribe, S. (2014). DNA barcoding for the identification of sand fly species (Diptera, Psychodidae, Phlebotominae) in Colombia. *PLoS ONE*, *9*(1), e85496.

Gwiazdowski, R. A., Foottit, R. G., Maw, H. E. L., & Hebert, P. D. N. (2015). The Hemiptera (Insecta) of Canada: Constructing a reference library of DNA barcodes. *PLoS ONE*, *10*(4), e0125635.

Haarto, A., & Ståhls, G. (2014). When mtDNA COI is misleading: Congruent signal of ITS2 molecular marker and morphology for North European *Melanostoma* Schiener, 1860 (Diptera, Syrphidae). *ZooKeys*, *431*, 93–134.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., & Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, *6*(4), e17497 (7 pp).

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.

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., … Goulson, D. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE*, *12*(10), e0185809.

Hansson, C., & Schmidt, S. (2018). Revision of the European species of Euplectrus Westwood (Hymenoptera, Eulophidae), with a key to European species of Euplectrini. *Journal of Hymenoptera Research*, *67*, 1.

Hardaluk, L. (in prep.). Metabarcoding in the Nationalpark Bayerischer Wald - screening for invasive and pest invertebrates in bulk samples.

Haszprunar, G. (2009). Barcoding Fauna Bavarica–eine Chance für die Entomologie. *Nachrichtenblatt Der Bayerischen Entomologen Bayer Ent*, *58*(1/2), 45.

Hausmann, A., Godfray, H. C. J., Huemer, P., Mutanen, M., Rougerie, R., van Nieukerken, E. J., … 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., Haszprunar, G., & Hebert, P. D. N. (2011). DNA barcoding the geometrid fauna of Bavaria (Lepidoptera): Successes, surprises, and questions. *PLoS ONE*, *6*(2), e17134. https://doi.org/10.1371/journal.pone.0017134

Hausmann, A., Haszprunar, G., Segerer, A. H., Speidel, W., Behounek, G., & Hebert, P. D. N. (2011). Now DNA-barcoded: The butterflies and larger moths of Germany. *Spixiana*, *34*(1), 47–58.

Havemann, N., Gossner, M. M., Hendrich, L., Morinière, J., Niedringhaus, R., Schäfer, P., & Raupach, M. J. (2018). From water striders to water bugs: The molecular diversity of aquatic Heteroptera (Gerromorpha, Nepomorpha) of Germany based on DNA barcodes. *PeerJ*, *6*, e4577. https://doi.org/10.7717/peerj.4577

Hawlitschek, O., Fernández-González, A., Balmori-de la Puente, A., & Castresana, J. (2018). A pipeline for metabarcoding and diet analysis from fecal samples developed for a small semi-aquatic mammal. *PLoS ONE*, *13*(8), e0201763. https://doi.org/10.1371/journal.pone.0201763

Hawlitschek, O., Morinière, J., Lehmann, G. U. C., Lehmann, A. W., Kropf, M., Dunz, 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

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., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Levesque-Beaudin, V., Milton, M. A., … Jannetta, P. (2016). Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150333.

Heller, K., Köhler, A., Menzel, F., Olsen, K. M., & Gammelo, Ø. (2016). Two formerly unrecognized species of Sciaridae (Diptera) revealed by DNA barcoding. *Norwegian Journal of Entomology*, *63*(1), 96–115.

Heller, K., & Rulik, B. (2016). *Ctenosciara alexanderkoenigi* sp. n. (Diptera: Sciaridae), an exotic invader in Germany? *Biodiversity Data Journal*, *4*, e6460.

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

Hernández-Triana, L. M., Prosser, S. W., Rodríguez-Perez, M. A., Chaverri, L. G., Hebert, P. D. N., & Ryan Gregory, T. (2014). Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Molecular Ecology Resources*, *14*(3), 508–518.

Hubert, N., & Hanner, R. (2015). DNA barcoding, species delineation and taxonomy: A historical perspective. *DNA Barcodes*, *3*(1), 44–58. https://doi.org/10.1515/dna-2015-0006

Ivanova, N. V., Dewaard, J. R., & Hebert, P. D. N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, *6*(4), 998–1002. https://doi.org/10.1111/j.1471-8286.2006.01428.x

Jaschhof, M. (2009). Eine aktualisierte Artenliste der Holzmücken Deutschlands, unter besonderer Berücksichtigung der Fauna Bayerns (Diptera, Cecidomyiidae, Lestremiinae). *Spixiana*, *32*(1), 139–151.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … 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

Jordaens, K., Goergen, G., Virgilio, M., Backeljau, T., Vokaer, A., & De Meyer, M. (2015). DNA barcoding to improve the taxonomy of the Afrotropical hoverflies (Insecta: Diptera: Syrphidae). *PLoS ONE*, *10*(10), e0140264. https://doi.org/10.1371/journal.pone.0140264

Jordaens, K., Sonet, G., Braet, Y., De Meyer, M., Backeljau, T., Goovaerts, F., … Desmyter, S. (2013). DNA barcoding and the differentiation between North American and West European *Phormia regina* (Diptera, Calliphoridae, Chrysomyinae). *ZooKeys*, *365*, 149–174. https://doi.org/10.3897/zookeys.365.6202

Karlsson, D., Pape, T., Johansson, K. A., Liljeblad, J., & Ronquist, F. (2005). The Swedish Malaise Trap Project, or how many species of Hymenoptera and Diptera are there in Sweden? *Entomologisk Tidsskrift*, *126*, 43–53.

Klausnitzer, B. (2006). Stiefkinder der Entomologie in Mitteleuropa. *Beiträge Zur Entomologie*, *56*, 360–368.

Krüger, A., Strüven, L., Post, R. J., & Faulde, M. (2011). The sandflies (Diptera: Psychodidae, Phlebotominae) in military camps in northern Afghanistan (2007–2009), as identified by morphology and DNA 'barcoding'. *Annals of Tropical Medicine & Parasitology*, *105*(2), 163–176. https://doi.org/10.1179/136485911X12899838683241

Kumar, N. P., Rajavel, A. R., Natarajan, R., & Jambulingam, P. (2007). DNA barcodes can distinguish species of Indian mosquitoes (Diptera:

Culicidae). *Journal of Medical Entomology*, *44*(1), 01–07. https://doi.org/10.1093/jmedent/41.5.01

Kumar, N. P., Srinivasan, R., & Jambulingam, P. (2012). DNA barcoding for identification of sand flies (Diptera: Psychodidae) in India. *Molecular Ecology Resources*, *12*(3), 414–420. https://doi.org/10.1111/j.1755-0998.2012.03117.x

Latibari, M. H., Moravvej, G., Heller, K., Rulik, B., & Namaghi, H. S. (2015). New records of Black Fungus Gnats (Diptera: Sciaridae) from Iran, including the reinstatement of *Bradysia cellarum* Frey. *Studia Dipterologica*, *22*(1), 39–44.

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences USA*, *112*(7), 2076–2081. https://doi.org/10.1073/pnas.1424997112

Leray, M., Yang, Y. J., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Lister, B. C., & Garcia, A. (2018). Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proceedings of the National Academy of Sciences*, *115*(44), E10397–E10406. https://doi.org/10.1073/pnas.1722477115

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Matthews, R. W., & Matthews, J. R. (1971). The Malaise trap: Its utility and potential for sampling insect populations. *The Great Lakes Entomologist*, *4*(4), 4.

May, R. M. (1988). How many species are there on earth? *Science*, *241*(4872), 1441–1449.

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (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

Mengual, X., Ståhls, G., Vujić, A., & Marcos-Garcia, M. A. (2006). Integrative taxonomy of Iberian *Merodon* species (Diptera, Syrphidae). *Zootaxa*, *1377*, 1–26.

Meyer, H., & Stark, A. (2015). Verzeichnis und Bibliografie der Tanzfliegenverwandten Deutschlands (Diptera: Empidoidea: Atelestidae, Brachystomatidae, Dolichopodidae s. l., Empididae, Hybotidae, "Iteaphila-Gruppe", Oreogetonidae). Studia Dipterologica Supplement 19.

Montagna, M., Mereghetti, V., Lencioni, V., & Rossaro, B. (2016). Integrated taxonomy and DNA barcoding of alpine midges (Diptera: Chironomidae). *PLoS ONE*, *11*(3), e0149673. https://doi.org/10.1371/journal.pone.0149673

Morinière, J., Cancian de Araujo, B., Lam, A. W., Hausmann, A., Balke, M., Schmidt, 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., Hendrich, L., Balke, M., Beermann, A. J., König, T., Hess, M., … Haszprunar, G. (2017). A DNA barcode library for Germany's mayflies, stoneflies and caddisflies (Ephemeroptera, Plecoptera and Trichoptera). *Molecular Ecology Resources*, *17*(6), 1293–1307. https://doi.org/10.1111/1755-0998.12683

Morinière, J., Hendrich, L., Hausmann, A., Hebert, P., Haszprunar, G., & Gruppe, A. (2014). Barcoding Fauna Bavarica: 78% of the Neuropterida fauna barcoded!. *PLoS ONE*, *9*(10), e109719. https://doi.org/10.1371/journal.pone.0109719

Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., … Godfray, H. C. J. (2016). Species-level para-and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology*, *65*(6), 1024–1040. https://doi.org/10.1093/sysbio/syw044

Nagy, Z. T., Sonet, G., Mortelmans, J., Vandewynkel, C., & Grootaert, P. (2013). Using DNA barcodes for assessing diversity in the family Hybotidae (Diptera, Empidoidea). *ZooKeys*, *365*, 263–278. https://doi.org/10.3897/zookeys.365.6070

Nelson, L. A., Lambkin, C. L., Batterham, P., Wallman, J. F., Dowton, M., Whiting, M. F., … Cameron, S. L. (2012). Beyond barcoding: A mitochondrial genomics approach to molecular phylogenetics and diagnostics of blowflies (Diptera: Calliphoridae). *Gene*, *511*(2), 131–142. https://doi.org/10.1016/j.gene.2012.09.103

Nelson, L. A., Wallman, J. F., & Dowton, M. (2007). Using COI barcodes to identify forensically and medically important blowflies. *Medical and Veterinary Entomology*, *21*(1), 44–52. https://doi.org/10.1111/j.1365-2915.2007.00664.x

Normark, B. B. (2003). The evolution of alternative genetic systems in insects. *Annual Review of Entomology*, *48*(1), 397–423.

Nzelu, C. O., Cáceres, A. G., Arrunátegui-Jiménez, M. J., Lañas-Rosas, M. F., Yañez-Trujillano, H. H., Luna-Caipo, D. V., … Kato, H. (2015). DNA barcoding for identification of sand fly species (Diptera: Psychodidae) from leishmaniasis-endemic areas of Peru. *Acta Tropica*, *145*, 45–51. https://doi.org/10.1016/j.actatropica.2015.02.003

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

Oliverio, A. M., Gan, H., Wickings, K., & Fierer, N. (2018). A DNA metabarcoding approach to characterize soil arthropod communities. *Soil Biology and Biochemistry*, *125*, 37–43. https://doi.org/10.1016/j.soilbio.2018.06.026

Oosterbroek, P. (2006). *The European Families of the Diptera*. Uitgeverij: KNNV-Vereniging voor Veldbiologie.

Packer, L., Gibbs, J., Sheffield, C., & Hanner, R. (2009). DNA barcoding and the mediocrity of morphology. *Molecular Ecology Resources*, *9*(Supplement 1), 42–50. https://doi.org/10.1111/j.1755-0998.2009.02631.x

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, *7*(1), 16.

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.

Pante, E., Schoelinck, C., & Puillandre, N. (2014). From integrative taxonomy to species description: One step beyond. *Systematic Biology*, *64*(1), 152–160.

Pape, T. (2009). Palaearctic Diptera - from tundra to desert. In T. Pape, D. Bickel, & R. Meier (Eds.), *Diptera diversity: Status, challenges and tools* (pp. 121–154). Leiden, The Netherlands: Brill.

Pape, T., Blagoderov, V., & Mostovski, M. B. (2011). Order Diptera Linnaeus, 1758. In Z.-Q. Zhang (Ed.), *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness* (pp. 222–229). Woodcroft, South Australia: Magnolia Press.

Papp, L., & B. Darvas (Eds.) (1997). *Contribution to a Manual of Palaearctic Diptera. Vol. 2, Nematocera and Lower Brachycera*. Budapest, Hungary: Science Herald.

Papp, L., & B. Darvas (Eds.) (1998). *Contribution to a Manual of Palaearctic Diptera. Vol. 3, Higher Brachycera*. Budapest, Hungary: Science Herald.

Papp, L., & B. Darvas (Eds.) (2000a). *Contribution to a Manual of Palaearctic Diptera. Vol. 1, General and Applied Dipterology*. Budapest, Hungary: Science Herald.

Papp, L., & B. Darvas (Eds.) (2000b). *Contribution to a Manual of Palaearctic Diptera. Appendix*. Budapest, Hungary: Science Herald.

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.

Pohjoismäki, J. L., Kahanpää, J., & Mutanen, M. (2016). DNA barcodes for the northern European tachinid flies (Diptera: Tachinidae). *PLoS ONE*, *11*(11), e0164933.

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: Trends, impacts and drivers. *Trends in Ecology & Evolution*, *25*(6), 345–353.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), 1864–1877. https://doi.org/10.1111/j.1365-294X.2011.05239.x

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

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.

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

Raupach, M. J., Hannig, K., Moriniere, J., & Hendrich, L. (2016). A DNA barcode library for ground beetles (Insecta, Coleoptera, Carabidae) of Germany: The genus *Bembidion* Latreille, 1802 and allied taxa. *ZooKeys*, *592*, 121–141. https://doi.org/10.3897/zookeys.592.8316

Raupach, M. J., Hannig, K., Morinière, J., & Hendrich, L. (2018). A DNA barcode library for ground beetles of Germany: The genus *Amara* Bonelli, 1810 (Insecta, Coleoptera, Carabidae). *ZooKeys*, *759*, 57–80. https://doi.org/10.3897/zookeys.759.24129

Raupach, M. J., Hendrich, L., Küchler, S. M., Deister, F., Morinière, J., & Gossner, M. M. (2014). Building-up of a DNA barcode library for true bugs (Insecta: Hemiptera: Heteroptera) of Germany reveals taxonomic uncertainties and surprises. *PLoS ONE*, *9*(9), e106940. https://doi.org/10.1371/journal.pone.0106940

Reibe, S., Schmitz, J., & Madea, B. (2009). Molecular identification of forensically important blowfly species (Diptera: Calliphoridae) from Germany. *Parasitology Research*, *106*(1), 257–261. https://doi.org/10.1007/s00436-009-1657-9

Reimann, B., & Rulik, B. (2015). *Dasiops calvus* Morge (Diptera: Lonchaeidae), a lance fly new to the German fauna, revealed by the GBOL-project. *Studia Dipterologica*, *21*(2), 283–287.

Renaud, A. K., Savage, J., & Adamowicz, S. J. (2012). DNA barcoding of Northern Nearctic Muscidae (Diptera) reveals high correspondence between morphological and molecular species limits. *BMC Ecology*, *12*(1), 24. https://doi.org/10.1186/1472-6785-12-24

Riedel, A., Sagata, K., Suhardjono, Y. R., Tänzler, R., & Balke, M. (2013). Integrative taxonomy on the fast track-towards more sustainability in biodiversity research. *Frontiers in Zoology*, *10*(1), 15. https://doi.org/10.1186/1742-9994-10-15

Rivera, J., & Currie, D. C. (2009). Identification of Nearctic black flies using DNA barcodes (Diptera: Simuliidae). *Molecular Ecology Resources*, *9*, 224–236. https://doi.org/10.1111/j.1755-0998.2009.02648.x

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

Rojo, S., Ståhls, G., Pérez-Bañón, C., & Marcos-García, M. Á. (2006). Testing molecular barcodes: Invariant mitochondrial DNA sequences vs the larval and adult morphology of West Palaearctic *Pandasyopthalmus* species (Diptera: Syrphidae: Paragini). *European Journal of Entomology*, *103*(2), 443. https://doi.org/10.14411/eje.2006.058

Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., … Ahrens, D. (2017). Using taxonomic consistency with semi-automated data preprocessing for high quality DNA barcodes. *Methods in Ecology and Evolution*, *8*(12), 1878–1887. https://doi.org/10.1111/2041-210X.12824

Santos, D., Sampronha, S., & Santos, C. M. D. (2017). Advances on dipterology in the 21st century and extinction rates. *Papéis Avulsos De Zoologia*, *57*(33), 433–444. https://doi.org/10.11606/0031-1049.2017.57.33

Schlick-Steiner, B. C., Arthofer, W., & Steiner, F. M. (2014). Take up the challenge! Opportunities for evolution research from resolving conflict in integrative taxonomy. *Molecular Ecology*, *23*(17), 4192–4194. https://doi.org/10.1111/mec.12868

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*, 421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Schmid-Egger, C., Straka, J., Ljubomirov, T., Blagoev, G. A., Morinière, J., & Schmidt, S. (2019). DNA barcodes identify 99 per cent of apoid wasp species (Hymenoptera: Ampulicidae, Crabronidae, Sphecidae) from the Western Palearctic. *Molecular Ecology Resources*, *19*(2), 476–484. https://doi.org/10.1111/1755-0998.12963

Schmidt, S., Schmid-Egger, C., Morinière, J., Haszprunar, G., & Hebert, P. D. N. (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.

Schmidt, S., Taeger, A., Morinière, J., Liston, A., Blank, S. M., Kramp, K., … Stahlhut, J. (2017). Identification of sawflies and horntails (Hymenoptera, 'Symphyta') through DNA barcodes: Successes and caveats. *Molecular Ecology Resources*, *17*(4), 670–685. https://doi.org/10.1111/1755-0998.12614

Schumann, H. (2002). Erster Nachtrag zur „Checkliste der Dipteren Deutschlands ". *Studia Dipterologica*, *9*(2), 437–445.

Schumann, H. (2004). Zweiter Nachtrag zur „Checkliste der Dipteren Deutschlands ". *Studia Dipterologica*, *11*(2), 619–630.

Schumann, H. (2010). Dritter Nachtrag zur „Checkliste der Dipteren Deutschlands". *Studia Dipterologica*, *16*(1/2), 17–27.

Schumann, H., Bährmann, R., & Stark, A. (1999). Checkliste der Dipteren Deutschlands. *Entomofauna Germanica 2*. *Studia Dipterologica Supplement*, *2*, 1–354.

Schumann, H., Doczkal, D., & Ziegler, J. (2011). Diptera - Zweiflügler. In: B. Klausnitzer (Ed.), *Stresemann*, *Exkursionsfauna von Deutschland*. *Vol. 2*, *Wirbellose: Insekten*. 11 (pp. 832–932). Auflage: Spektrum Akademischer Verlag.

Serrana, J. M., Miyake, Y., Gamboa, M., & Watanabe, K. (2018). Comparison of DNA metabarcoding and morphological identification for stream macroinvertebrate biodiversity assessment and monitoring. *bioRxiv*, 436162. https://doi.org/10.1101/436162

Ševčík, J., Kaspřák, D., & Rulik, B. (2016). A new species of *Docosia* Winnertz from Central Europe, with DNA barcoding based on four gene markers (Diptera, Mycetophilidae). *ZooKeys*, *549*, 127–143. https://doi.org/10.3897/zookeys.549.6925

Shokralla, S., Spall, J., Gibson, J., & 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

Sinclair, C. S., & Gresens, S. E. (2008). Discrimination of *Cricotopus* species (Diptera: Chironomidae) by DNA barcoding. *Bulletin of Entomological Research*, *98*(6), 555–563. https://doi.org/10.1017/S0007485308005865

Sorg, M., Schwan, H., Stenmans, W., & Müller, A. (2013). Ermittlung der Biomassen flugaktiver Insekten im Naturschutzgebiet Orbroicher Bruch mit Malaise Fallen in den Jahren 1989 und 2013. *Mitteilungen Entomologischer Verein Krefeld*, *1*, 1–5.

Spelda, J., Reip, H. S., Oliveira Biener, U., & Melzer, R. R. (2011). Barcoding Fauna Bavarica: Myriapoda – a contribution to DNA sequence-based identifications of centipedes and millipedes (Chilopoda, Diplopoda). *ZooKeys*, *115*, 123–139. https://doi.org/10.3897/zookeys.156.2176

Ssymank, A., Doczkal, D., Rennwald, K., & Dziock, F. (2011). Rote Liste und Gesamtartenliste der Schwebfliegen (Diptera: Syrphidae) Deutschlands. *Naturschutz Und Biologische Vielfalt*, *70*(3), 13–83.

Ssymank, A., Sorg, M., Doczkal, D., Rulik, B., Merkel-Wallner, G., & Vischer-Leopold, M. (2018). Praktische Hinweise und Empfehlungen zur Anwendung von Malaisefallen für Insekten in der Biodiversitätserfassung und im Monitoring. *Series Naturalis*, *1*, 1–12.

Stur, E., & Borkent, A. (2014). When DNA barcoding and morphology mesh: Ceratopogonidae diversity in Finnmark, Norway. *ZooKeys*, *463*, 95–131. https://doi.org/10.3897/zookeys.463.7964

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

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

Vanbergen, A. J., & Insect Pollinators Initiative. (2013). Threats to an ecosystem service: Pressures on pollinators. *Frontiers in Ecology and the Environment*, *11*(5), 251–259. https://doi.org/10.1890/120126

Versteirt, V., Nagy, Z. T., Roelants, P., Denis, L., Breman, F. C., Damiens, D., … Van Bortel, W. (2015). Identification of Belgian mosquito species (Diptera: Culicidae) by DNA barcoding. *Molecular Ecology Resources*, *15*(2), 449–457. https://doi.org/10.1111/1755-0998.12318

Völkl, W., Blick, T., Kornacker, P. M., & Martens, H. (2004). Quantitativer Überblick über die rezente Fauna von Deutschland. *Natur Und Landschaft*, *79*(7), 293–295.

Wang, G., Li, C., Guo, X., Xing, D., Dong, Y., Wang, Z., … Zhao, T. (2012). Identifying the main mosquito species in China based on DNA barcoding. *PLoS ONE*, *7*(10), e47051. https://doi.org/10.1371/journal.pone.0047051

Wesener, T., Voigtländer, K., Decker, P., Oeyen, J. P., Spelda, J., & Lindner, N. (2015). First results of the German Barcode of Life (GBOL)–Myriapoda project: Cryptic lineages in German *Stenotaenia linearis* (Koch, 1835) (Chilopoda, Geophilomorpha). *ZooKeys*, *510*, 15–29. https://doi.org/10.3897/zookeys.510.8852

Wheeler, Q. D., Raven, P. H., & Wilson, E. O. (2004). Taxonomy: Impediment or expedient? *Science*, *303*, 285–285. https://doi.org/10.1126/science.303.5656.285

Whitworth, T. L., Dawson, R. D., Magalon, H., & Baudry, E. (2007). DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences*, *274*(1619), 1731–1739.

Wolff, D., Gebel, M., & Geller-Grimm, F. (2018). *Die Raubfliegen Deutschlands*. Quelle & Meyer Bestimmungsbücher.

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.

## SUPPORTING INFORMATION

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

**How to cite this article:** Morinière J, Balke M, Doczkal D, 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;00:1–29. https://doi.org/10.1111/1755-0998.13022

## Publication III - DNA Barcoding in Forensic Entomology – Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Species

Citation: Chimeno, Caroline, Jérôme Morinière, Jana Podhorna, Laura Hardulak, Axel Hausmann, Frank Reckel, Jan E. Grunwald, Randolph Penning, and Gerhard Haszprunar. (2018). “DNA Barcoding in Forensic Entomology - Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Species.” *Journal of Forensic Sciences* 64(2): 593–601.

JOURNAL OF FORENSIC SCIENCES

*J Forensic Sci*, 2018
doi: 10.1111/1556-4029.13869
Available online at: onlinelibrary.wiley.com

**TECHNICAL NOTE**

**PATHOLOGY/BIOLOGY**

*Caroline Chimeno*,[1] *M.Sc.; Jérôme Morinière*,[1] *M.Sc.; Jana Podhorna*,[2] *Ph.D.; Laura Hardulak*,[1] *M.Sc.; Axel Hausmann*,[1] *Ph.D.; Frank Reckel*,[3] *Ph.D.; Jan E. Grunwald*,[3] *Ph.D.; Randolph Penning*,[4] *M.D.; and Gerhard Haszprunar*,[1] *Ph.D.*

# DNA Barcoding in Forensic Entomology – Establishing a DNA Reference Library of Potentially Forensic Relevant Arthropod Species*,†

**ABSTRACT:** Throughout the years, DNA barcoding has gained in importance in forensic entomology as it leads to fast and reliable species determination. High-quality results, however, can only be achieved with a comprehensive DNA barcode reference database at hand. In collaboration with the Bavarian State Criminal Police Office, we have initiated at the Bavarian State Collection of Zoology the establishment of a reference library containing arthropods of potential forensic relevance to be used for DNA barcoding applications. CO1-5P' DNA barcode sequences of hundreds of arthropods were obtained via DNA extraction, PCR and Sanger Sequencing, leading to the establishment of a database containing 502 high-quality sequences which provide coverage for 88 arthropod species. Furthermore, we demonstrate an application example of this library using it as a backbone to a high throughput sequencing analysis of arthropod bulk samples collected from human corpses, which enabled the identification of 31 different arthropod Barcode Index Numbers.

**KEYWORDS:** forensic science, forensic entomology, DNA barcoding, high throughput sequencing, next generation sequencing, Cytochrome C Oxidase 1, DNA reference library, bulk sample analysis

One important task within forensic sciences is estimating the postmortem interval (PMI) of a deceased individual (1). While there are numerous medical techniques in pathology for doing so, they are only applicable within approximately the first 72 h after death as the conditional body changes needed for such analyses disappear with ongoing body decomposition (2). Therefore, in cases where a body is recovered in later stages of decomposition, another field of expertise is needed for such clarification (3). Because arthropod colonization of a carrion source is assumed to coincide fairly with the start of death, a forensic entomologist can—assuming that arthropod colonization is possible and not impeded through various factors such as cold weather—estimate the postmortem interval (PMI), or minimum postmortem interval (PMImin) [see [4] for the ongoing debate on the correct terminology], by studying the arthropods sampled from the decomposing body (3). Generally, when using arthropod specimen in death investigations, the first and most crucial step is accurate species identification (5). This, however, may pose a major problem if entomologists are confronted with large amounts of partial arthropod remains such as exuviae, limbs, and unidentified arthropod biomass. Even intact specimens pose a large burden when wanting to apply morphological methods, as eggs, early larval stages, and sometimes even later stages of many different species share similar features making it close to impossible in certain groups, even for a specialized taxonomist, to distinguish between them based on morphology alone (2). In many cases, collected insects of early instars need to be incubated and raised under constant conditions until the imago stage is achieved and distinguishable features become visible (6).

The application of molecular biology in the field of forensic entomology has gained in importance throughout the last decades, as it offers countless new possibilities in analyzing arthropod specimens (7). DNA barcoding uses a short genetic sequence of the mitochondrial cytochrome c oxidase 1 gene (CO1-5′) as a unique identifier to differentiate between species (8,9), enabling fast and reliable species identification even from small amounts of unknown arthropod biomass (10). However, it is most important to note that the quality of the results derived from such an analysis depend strongly on the reference database used for sequence comparison and associated taxonomic

[1]Zoologische Staatssammlung München (SNSB-ZSM), Münchhausenstrasse 21, 81247 München, Germany.
[2]Mendel University in Brno (MEDELU), Zemedelska 1, Brno 613 00 Czech Republic.
[3]Abteilung II, Bayerisches Landeskriminalamt, Maillingerstraße 15, 80636 München, Germany.
[4]Institute of Legal Medicine, Ludwig-Maximilians Universität, München, Germany.
Corresponding author: Caroline Chimeno, M.Sc. E-mail: ca_chimeno@yahoo.com

*This study was conducted in the framework of the GBOL Project, which is supported by a grant from the German Federal Ministry of Education and Research (FKZ 01LI1101 and 01LI1501).
†Presented at the 13th Meeting of the European Association for Forensic Entomology (EAFE), May 25–28, 2016, in Budapest, Hungary.
Received 17 Feb. 2018; and in revised form 30 April 2018, 11 June 2018; accepted 12 June 2018.

© 2018 American Academy of Forensic Sciences

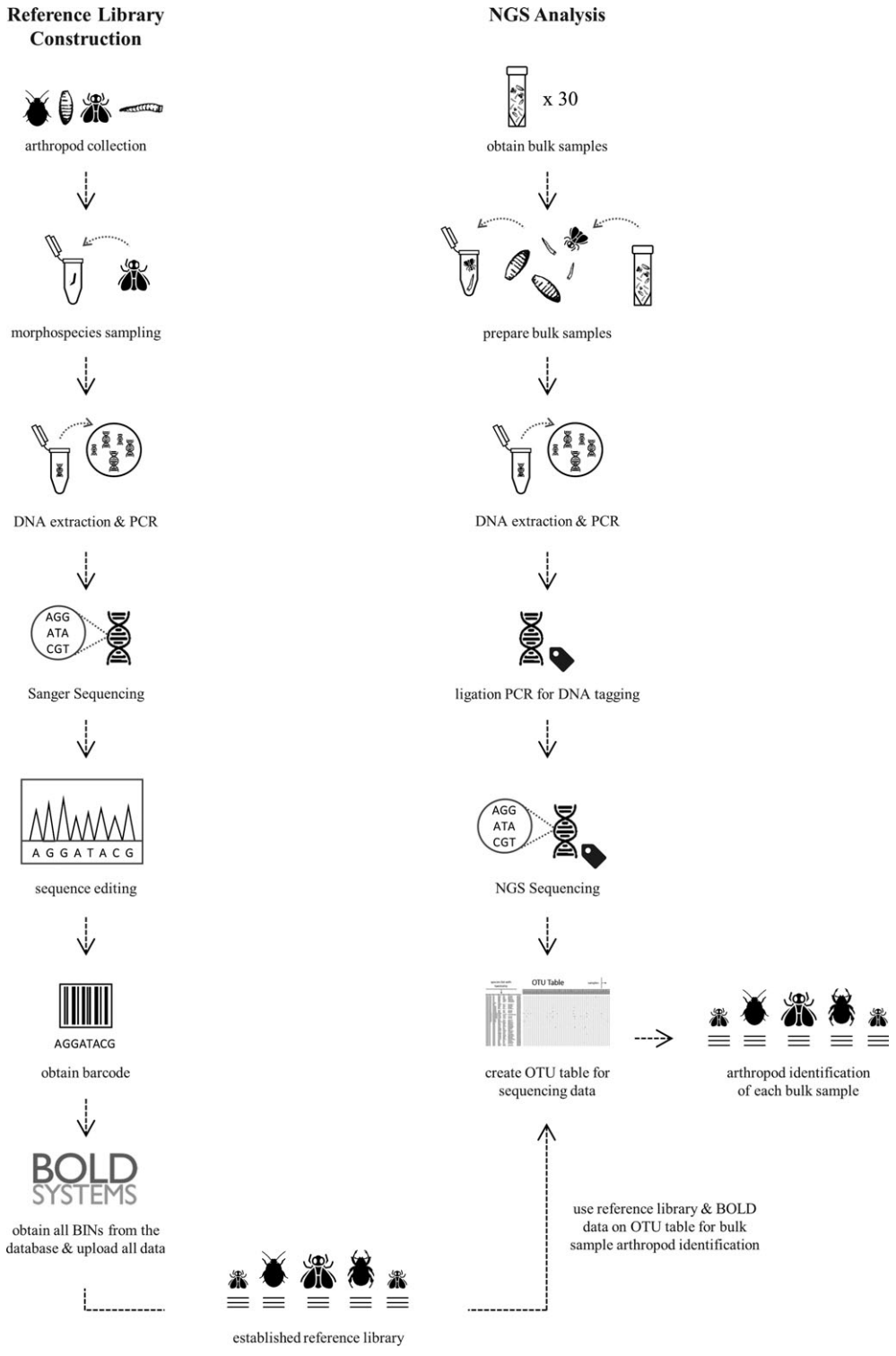


FIG. 1—*Visualization of the study workflow.*

identifications. The establishment of a high-quality reference DNA barcode database is, therefore, an important prerequisite when wanting to guarantee accurate results (11).

Within the framework of the International Barcode of Life (iBOL) campaign and with close cooperation with the Biodiversity Institute of Ontario (BIO, Guelph, Canada), a DNA barcode reference library of approximately 20 000 animal species represented by more than 250 000 specimens has been constructed since 2009 at the Bavarian State Collection of Zoology in Munich (SNSB, ZSM - www.barcoding-zsm.de) and provided for the international Barcode of Life Database (www.boldsystems.org BOLD; [12]). These animal species have been identified by professional taxonomists and skilled experts prior to library construction. This DNA barcode reference library enables a very high standard of accurate species identification, being most comprehensive for arthropods, including Coleoptera (10),

TABLE 1—*Primers and corresponding PCR conditions used in this study.*

| Primer | F/R | Sequence | References |
|---|---|---|---|
| LepF | F | 5′ ATTCAACCAATCATAAAGATATTGG 3′ | (45) |
| Nancy | R | 5′ CCTGGTAAAATTAAAATATAAACTTC 3′ | (46) |
| PCR conditions | | | |
| 2′:94°C– 5x[1′:94°C– 90″:45°C– 90″:72°C]– 35x[1′:93°C– 90″:50°C– 90″:72°C]– 10′:72°C | | | |

TABLE 3—*Success of the DNA barcoding analysis.*

| | |
|---|---|
| Number of processed specimen | 1392 |
| Success rate | 48.7% |
| Number of recovered sequences | 678 |
| Number of sequences ≥500 bp | 502 |
| Number of species with sequences ≥500 bp | 88 |
| Number of BINS to sequences ≥500 bp | 86 |

TABLE 4—*Number of specimens, species and BINs per arthropod order.*

| Order | Nr. of specimens | Nr. of species | Nr. of BINs |
|---|---|---|---|
| Araneae | 1 | 1 | 0 |
| Coleoptera | 41 | 14 | 14 |
| Diptera | 452 | 70 | 73 |
| Hymenoptera | 5 | 1 | 1 |
| Isopoda | 1 | 0 | 0 |
| Mecoptera | 1 | 1 | 0 |
| Pseudoskorpiones | 1 | 1 | 1 |

Hymenoptera (13,14), Lepidoptera (15,16), Neuroptera (17), Orthoptera (18), Heteroptera (8), Ephemeroptera, Plecoptera & Trichoptera (19), Araneae & Opiliones (20), and Myriapoda (11). A data release on the German Diptera is still in progress. With this study, we present the construction of a separate library on BOLD for its unique use within the field of forensic entomology. Although DNA barcoding has been the subject of various publications since its first successful demonstration in forensic sciences in 1994 (21), the majority of these publications (e.g. [22–27]) focus their studies solely on one or few arthropod families and/or species relevant to forensic entomology. Here, we have conducted a broad-scale study on hundreds of arthropod specimens sampled from various carrion sources as well as from human bodies to create a separate internationally accessible database on BOLD containing numerous potentially forensic relevant arthropod species found in Europe. The aim was to ensure high-quality DNA barcoding results for accurate species identification, to be used for future analyses within this field. Furthermore, we wanted to demonstrate the application of such a library using it as a backbone to a high throughput sequencing (HTS) analysis administered to bulk samples or arthropod specimens obtained from human corpses.

This study, therefore, demonstrates the (i) establishment of a reference database containing arthropod species which are relevant in forensic entomology, and (ii) the use of this library as a backbone to merely test its application on bulk extractions of real life arthropod communities found on corpses via HTS, in order to ultimately obtain data for a possible PMImin estimation.

## Materials and Methods

### *Specimens*

For reference library construction, we investigated a total of 1392 arthropod specimens of which the majority (1003 specimens; 72%) were stored in 96% EtOH, while one-third of the samples was dried and pinned. Studied specimens originated from different localities and previous experiments, and were collected within the years of 2008–2015 in Munich (2008, 2015), Brno (2014), Litovel (2014), Zabcice (2014), and Budweis (2014/2015).The source of these arthropods were pigs, birds, and rats. Most of these specimens were imagos, which were previously identified morphologically through various authors and experts using standard keys and literature (3,28–44). The remaining specimens were identified to the order and/or family level only prior to Sanger Sequencing and in these cases, species identification was performed by reverse taxonomy using the SNSB, ZSM DNA barcode reference data on BOLD. These specimens were used for the establishment of a forensic reference library.

Furthermore, in a separate context, we have obtained a total of 30 bulk samples containing arthropods, which were collected by forensic scientists at the Munich Institute of Forensic Pathology (2015/2016) off human bodies originating from forensic cases. All specimens have been collected in prelabeled and prefilled (with 96% EtOH) 50 mL Falcon tubes. Each 50 mL Falcon tube was half-filled with arthropods (approximately 1000–2500 specimens per tube, whereas each sample was dominated by Dipteran larvae and eggs ~80%). Species identification within the bulk samples was not performed morphologically, as most specimens were represented by larvae and/or eggs, and especially because these bulk samples were not intended to be used for reference library construction; their analysis simply served as an application example of metabarcoding technology.

The following workflows for reference library construction and high throughput sequencing analysis are visualized in Fig. 1.

### *DNA Barcode Reference Library Construction*

One to three specimens of each designated morphospecies were selected to be transferred into 96-well plates. A small muscle tissue sample was extracted from large specimen (large larvae, imagos and puparia), whereas the entire specimen was used for small individuals, such as mites, small larvae and eggs.

Tissue lysis was performed using a mixture of Proteinase K and lysis buffer following the manufacturer's instructions (DNeasy blood & tissue kit, Qiagen, Hilden, Germany). The lysis reaction was facilitated by occasional vortexing of the

TABLE 2—*Mini metabarcoding primers used for the PCR amplification of all bulk samples.*

| Primer | F/R | Sequence | References |
|---|---|---|---|
| mlCOIintF | F | 5′ GGWACWGGWTGAACWGTWTAYCCYCC 3′ | (49) |
| dgHco | R | 5′ TAAACTTCAGGGTGACCAAARAAYCA 3′ | (49) |
| PCR conditions | | | |
| 2′:96°C– 3x[15″:96°C– 30″:48°C– 90″:65°C]– 30x[15″:96°C– 30″:55°C—90″:65°C]– 10′:72°C | | | |

TABLE 5—*The reference library of potentially forensic relevant arthropod species.*

| Order | Family | Species | Nr. of Sequences | BINs |
|---|---|---|---|---|
| Araneae | Agelenidae | *Coelotes terrestris* | 1 | – |
| Coleoptera | Dermestidae | *Dermestes frischii* | 1 | BOLD:ACB8353 |
| | | *Dermestes haemorrhoidalis* | 3 | BOLD:AAI9639 |
| | | *Dermestes maculatus* | 2 | BOLD:AAF8298 |
| | | *Dermestes murinus* | 1 | BOLD:ACE1097 |
| | Histeridae | *Saprinus semistriatus* | 12 | BOLD:ABX1714 |
| | Nitidulidae | *Omosita depressa* | 2 | BOLD:ABY0372 |
| | | *Omosita discoidea* | 1 | BOLD:AAK3701 |
| | Silphidae | *Necrodes littoralis* | 3 | BOLD:AAP7891 |
| | | *Nicrophorus humator* | 1 | BOLD:AAF2685 |
| | | *Nicrophorus vespilloides* | 5 | BOLD:AAF3432 |
| | | *Nicrophorus vespillo* | 3 | BOLD:AAG3728 |
| | | *Thanatophilus sinuatus* | 2 | BOLD:AAW6863 |
| | Staphylinidae | *Aleochara curtula* | 1 | BOLD:AAJ2741 |
| | | *Omalium rivulare* | 4 | BOLD:AAN1494 |
| Diptera | Acartophthalmidae | *Acartophthalmus bicolor* | 1 | – |
| | Anthomyiidae | *Anthomyia procellaris* | 3 | BOLD:AAP2970 |
| | | *Emmesomyia grisea* | 1 | – |
| | | *Lasiomma picipes* | 1 | BOLD:ACZ5374 |
| | | *Lasiomma strigilatum* | 1 | BOLD:ACI8977 |
| | Calliphoridae | *Calliphora vicina* | 30 | BOLD:AAB6579 |
| | | *Calliphora vomitoria* | 6 | BOLD:AAA8931 |
| | | *Chryosyma albiceps* | 1 | BOLD:ABX6432 |
| | | *Lucilia ampullacea* | 18 | BOLD:AAC3450 |
| | | *Lucilia caesar* | 26 | BOLD:AAA7470 |
| | | *Lucilia caesarillustris* | 8 | BOLD:AAA7470 |
| | | *Lucilis curpina x sericata* | 1 | BOLD:AAA6618 |
| | | *Lucilia illustris* | 2 | BOLD:AAA7470 |
| | | *Lucilia sericata* | 74 | BOLD:AAA6618 |
| | | *Phormia regina* | 2 | BOLD:AAZ7380 |
| | | *Pollenia amentaria* | 6 | BOLD:ABV5497 |
| | | *Pollenia angustigena* | 1 | BOLD:AAP2825 |
| | | *Pollenis hungarica* | 1 | – |
| | | *Pollenia pediculata* | 3 | BOLD:AAG6745 |
| | | *Pollenia rudis* | 3 | BOLD:AAH3035 |
| | | *Protophormia terraenovae* | 6 | BOLD:AAC9614 |
| | Carnidae | *Meoeura spBOLDAAG6972* | 1 | BOLD:AAG6972 |
| | Drosophilidae | *Drosophila testacea* | 1 | BOLD:ABX1717 |
| | | *Scaptomyza pallida* | 1 | BOLD:ACE9016 |
| | Fanniidae | *Fannia aequilineata* | 1 | BOLD:AAG1746 |
| | | *Fannia armata* | 2 | BOLD:AAU6630 |
| | | *Fannia canicularis* | 14 | BOLD:AAF7101 |
| | | *Fannia lustrator* | 1 | BOLD:ACB3656 |
| | | *Fannia manicata* | 3 | BOLD:ABV8154 |
| | | *Fannia monilis* | 2 | – |
| | | *Fannia prisca* | 34 | BOLD:ACJ6083 |
| | | – | 13 | BOLD:ACR0452 |
| | Heleomyzidae | *Heleomyza serrata* | 4 | BOLD:ABX8716 |
| | | *Neoleria inscripta* | 1 | BOLD:AAU6649 |
| | | *Scoliocentra brachypterna* | 1 | BOLD:ACK9089 |
| | | *Scoliocentra villosa* | 1 | BOLD:ACD3147 |
| | | *Tephrochlamys flavipes* | 3 | BOLD:ACD3340 |
| | Limoniidae | *Limonia nubeculosa* | 1 | – |
| | Muscidae | *Hydrotaea cyrtoneurina* | 2 | BOLD:AAX2553 |
| | | *Hydrotaea dentipes* | 22 | BOLD:AAZ9882<br>BOLD:AAI83769 |
| | | *Hydrotaea ignava* | 20 | BOLD:ABW3765 |
| | | *Hydrotaea irritans* | 3 | BOLD:AAX2545 |
| | | *Hydrotaea meteorica* | 1 | BOLD:ACB3402 |
| | | *Musca domestica* | 8 | BOLD:AAA6020 |
| | | *Muscina levida* | 3 | BOLD:AAB8817 |
| | | *Muscina pascuorum* | 1 | BOLD:AAG1714 |
| | | *Muscina prolapsa* | 1 | BOLD:AAI3240 |
| | | *Muscina stabulans* | 1 | BOLD:AAM4634 |
| | | *Mydaea ancilla* | 3 | BOLD:AAX3222 |
| | | *Phaonia pallida* | 1 | BOLD:ABW3852 |
| | | *Phaonia subventa* | 4 | BOLD:AAG7029 |
| | | *Polietes lardarius* | 2 | BOLD:AAY2766 |
| | | – | 1 | BOLD:ACP3754 |
| | Phoridae | *Megaselia scalaris* | 28 | BOLD:AAG3322 |
| | | *Spelobia sp.* | 1 | BOLD:ACE0514 |

TABLE 5—*Continued.*

| Order | Family | Species | Nr. of Sequences | BINs |
|---|---|---|---|---|
| | | – | 1 | BOLD:ACU4097 |
| | | – | 1 | – |
| | Piophilidae | *Allopiophila vulgaris* | 1 | BOLD:AAG1787 |
| | | *Liopiophila varipes* | 7 | BOLD:AAG1789 |
| | | *Parapiophila vulgaris* | 2 | BOLD:ACU3238 |
| | | *Protopiophila latipes* | 6 | BOLD:AAF6352 |
| | | *Stearibia nigriceps* | 13 | BOLD:AAE9188 |
| | | *Piophila casei* | 2 | BOLD:AAG6813 |
| | Sarcophagidae | *Sarcophaga albiceps* | 1 | BOLD:AAE9461 |
| | | *Sarcophaga argyrostoma* | 5 | BOLD:AAI0975 |
| | | *Sarcophaga caerulescens* | 3 | BOLD:ABZ2577 |
| | | *Sarcophaga carnaria* | 1 | BOLD:AAX9423 |
| | | *Sarcophaga subvicina* | 2 | BOLD:AAG6743 |
| | Scathophagidae | – | 1 | – |
| | Sciaridae | *Scatopsciara vitripennis* | 1 | – |
| | Sepsidae | *Nemopoda nitidula* | 6 | BOLD:AAG5640 |
| | | *Sepsis fulgensxorthocnemis* | 1 | BOLD:AAJ7599 |
| | Simuliidae | *Simulium balcanicum* | 1 | BOLD:AAM4036 |
| | Sphaeroeridae | *Caproica ferruginata* | 5 | BOLD:AAN6407 |
| | | *Coproica hirticula* | 4 | BOLD:ABV3226 |
| | | *Ischiolepta pusilla* | 1 | BOLD:AAV0763 |
| | | *Leptocera caenosa* | 2 | BOLD:AAG7280 |
| | | *Spelobia luteilabris* | 1 | BOLD:AAL7752 |
| | | – | 5 | BOLD:AAG7028 |
| | Trichoceridae | – | 1 | BOLD:ACC0273 |
| Hymenoptera | Braconidae | *Alysia manducator* | 2 | BOLD:AAG1325 |
| | | | 1 | – |
| | Formicidae | – | 1 | – |
| | Ichneumonidae | *Phygadeuon detestator* | 1 | – |
| | | – | 1 | – |
| Isopoda | Trachelipodidae | – | 1 | – |
| Mecoptera | Panorpidae | *Panorpa vulgaris* | 1 | – |
| Pseudoskorpiones | Neobisiidae | *Neobisium carcinoides* | 1 | BOLD:ACR8463 |

sample tubes. Consecutive extraction of genomic DNA was performed with standard barcoding protocols provided by the Canadian Center for DNA Barcoding (CCDB; www.ccdb.ca). All samples were PCR amplified using the Mango-Taq (Bioline, Luckenwalde, Germany) and the primers LepF (45) and Nancy (46) (Table 1). A BigDye Cycle Sequencing PCR (ABI, Darmstadt, Germany) was applied using the same primers in order to prepare the samples for Sanger Sequencing. All samples were sent to a sequencing facility for sequencing.

The received sequences were assembled, respectively, and edited using the Sequencher Sequence Editing software v4.10.1 (Gene Codes, U. S., Ann Arbour), then aligned to one another with MEGA 7 (47) for further corrections. By uploading the edited DNA barcode sequences and trace files to the BOLD database, barcode index numbers (BINs) of these individuals were acquired. Within BOLD, similar CO1 barcode sequences are assigned a globally unique identifier (48). All metadata including the species names, images, voucher numbers as well as regions and countries of origin were given for each specimen. The created library, as well as all uploaded data related to it is accessible on BOLD under the identification BCFOR (available under DOI: XXX-DS-BCFOR).

### *High Throughput Sequencing Analysis on Bulk Samples*

All 30 bulk samples obtained from the morgue were drained from EtOH and placed in an oven (70°C) for at least 6 h until all specimens were completely dry. Tissue lysis of the entire bulk samples was performed for 8 h at 56°C and in this case, the amount of the buffer was adapted to the biomass in each tube. DNA bulk extraction followed the same protocols as mentioned in the paragraph above. PCR products were obtained using the mini metabarcoding NGS primers mlCOIintF and dgHco (49), and the Mango-Taq (Bioline, Luckenwalde, Germany) (Table 2). In a subsequent ligation PCR, amplified PCR products were tagged with unique index sequence tags. DNA concentrations of each amplicon pool were measured using Qubit ds-DNA high sensitivity chemicals (Life Technologies, Darmstadt, Germany) and were then pooled to equal molarity. Amplicon pools were cleaned using MinElute columns (Qiagen, Hilden, Germany) to eliminate unwanted residues. The cleanup products were then sent to a commercial sequencing facility for paired-end sequencing on an Illumina MiSeq (v2 chemicals $2 \times 250$ bp).

The bioinformatics pipeline presented by Morinière et al., (50) was followed for content identification of all bulk samples. The received CO1 mini barcode sequences ~313 bp and ~320 bp and were identified to the BIN-level using the BOLD database, with the representative sequences used for BLAST-ing in Geneious v8.0.3 (Biomatters, Auckland, New Zealand). This was done by merging the paired-end CO1 sequences with BBMerge (BBMap v. 35.80) before clustering with CD-HIT EST v4.6.5 (51), at a 98% similarity cutoff.

An OTU table was created displaying the number of raw sequence reads included in each OTU for each sample. OTU sequences were then BLASTed against a custom database consisting of public and private sequences downloaded from BOLD using Geneious 10.2.3 (http://www.geneious.com, [52]). The search results were imported into a spreadsheet for filtering in order to retain only sequences with 100% similarity. The OTU

TABLE 6—*All cases of unique BINs and BINs with multiple species.*

| Unique BINs | BINs with Multiple Species | Nr. of Species associated to these BINs |
|---|---|---|
| BOLD:AAB8817 | BOLD:AAA6020 | 3 |
| BOLD:AAC9614 | BOLD:AAA6618 | 11 |
| BOLD:AAE9188 | BOLD:AAA7470 | 14 |
| BOLD:AAF2685 | BOLD:AAA8931 | 2 |
| BOLD:AAF6352 | BOLD:AAB6579 | 6 |
| BOLD:AAF7101 | BOLD:AAC3450 | 3 |
| BOLD:AAF8298 | BOLD:AAE9461 | 3 |
| BOLD:AAG1325 | BOLD:AAF3432 | 2 |
| BOLD:AAG1714 | BOLD:AAG1787 | 3 |
| BOLD:AAG1746 | BOLD:AAG3728 | 2 |
| BOLD:AAG1789 | BOLD:AAG6743 | 3 |
| BOLD:AAG3322 | BOLD:AAG6813 | 2 |
| BOLD:AAG5640 | BOLD:AAI9639 | 2 |
| BOLD:AAG6745 | BOLD:AAJ7599 | 2 |
| BOLD:AAG6972 | BOLD:AAM4036 | 4 |
| BOLD:AAG7028 | BOLD:AAP2825 | 2 |
| BOLD:AAG7280 | BOLD:AAX9423 | 4 |
| BOLD:AAG7029 | BOLD:AAZ9882 | 2 |
| BOLD:AAH3035 | BOLD:ABX6432 | 2 |
| BOLD:AAI0975 | BOLD:ABX8716 | 2 |
| BOLD:AAI3240 | BOLD:ABY0372 | 2 |
| BOLD:AAI8769 | BOLD:ACD3340 | 2 |
| BOLD:AAJ0780 | BOLD:ACJ6083 | 4 |
| BOLD:AAK3419 | BOLD:ADH9310 | 2 |
| BOLD:AAK3701 | | |
| BOLD:AAL7752 | | |
| BOLD:AAM4634 | | |
| BOLD:AAN1494 | | |
| BOLD:AAN6407 | | |
| BOLD:AAP7891 | | |
| BOLD:AAU6630 | | |
| BOLD:AAU6649 | | |
| BOLD:AAV0763 | | |
| BOLD:AAV6691 | | |
| BOLD:AAW6863 | | |
| BOLD:AAX2545 | | |
| BOLD:AAX2553 | | |
| BOLD:AAX3222 | | |
| BOLD:AAY2766 | | |
| BOLD:AAZ7380 | | |
| BOLD:ABV3226 | | |
| BOLD:ABV5497 | | |
| BOLD:ABV8154 | | |
| BOLD:ABW3765 | | |
| BOLD:ABW3852 | | |
| BOLD:ABX1714 | | |
| BOLD:ABX1717 | | |
| BOLD:ABZ2577 | | |
| BOLD:ACB3402 | | |
| BOLD:ACB3656 | | |
| BOLD:ACB8353 | | |
| BOLD:ACC0273 | | |
| BOLD:ACD3147 | | |
| BOLD:ACE0514 | | |
| BOLD:ACE1097 | | |
| BOLD:ACE9016 | | |
| BOLD:ACI8977 | | |
| BOLD:ACK9089 | | |
| BOLD:ACP3565 | | |
| BOLD:ACP3754 | | |
| BOLD:ACQ1683 | | |
| BOLD:ACR0452 | | |
| BOLD:ACR4546 | | |
| BOLD:ACR8463 | | |
| BOLD:ACU3075 | | |
| BOLD:ACU3238 | | |
| BOLD:ACU4097 | | |
| BOLD:ACU4303 | | |
| BOLD:ACU4486 | | |
| BOLD:ACU4524 | | |
| BOLD:ACU4758 | | |
| BOLD:ACU4896 | | |
| BOLD:ACU5310 | | |
| BOLD:ACU5491 | | |
| BOLD:ACU5492 | | |
| BOLD:ACZ5374 | | |
| BOLD:ADA9775 | | |
| BOLD:ADG6176 | | |

TABLE 6—*Continued.*

TABLE 7—*BIN association throughout the samples within the reference library.*

| | |
|---|---|
| Number of BINs | 86 |
| Number of species | 88 |
| Percentage dipteran specimens assigned to BINs | 87% |
| Percentage coleopteran specimens assigned to BINs | 100% |

table was then imported into a spreadsheet, and the taxonomic names of the database sequences corresponding to the OTUs were matched to their entries within the table. To guarantee quality results, filtering was done to retain only OTUs with hits ≥10. The annotated OTU table served as the foundation for the assembly of a species list including all bulk samples.

## Results and Discussion

### *Establishment of the Forensic Reference Library*

A total of 1,392 specimens were processed in this study, of which 678 (48.7%) generated CO1 barcode sequences, leading to the identification of 92 species assigned to 102 BINs (Table 3). Sequencing success was highest for the freshly collected samples from 2014 and 2015 (571 of 1003; 57%), whereas the quality of PCR reactions and Sanger sequencing dropped for the older samples from 2008 (107 of 393; 27%). Although, it was not the focus of this study to perform tests using various primers targeting the CO1-5P region, it would have been beneficial to the amplification success of the studied material. Forty-four specimens were identified to the family or genus level only; these included 39 Diptera, 3 Hymenoptera, 1 Isopoda and 1 Mesostigmata, whereas BINs were nonetheless assigned in 18 of these cases (42%). The generated fragment lengths of these 678 samples ranged from short fragments (<500 bp) for 177 specimens (26.1%), to complete barcodes with a maximum length of 658 bp, which was the case for 70 specimens (10.3%). A total of 502 sequences displayed a COI-5P sequence length of ≥500 bp and were thus defined as successful barcodes to be used for the forensic library. All fragments <500 bp were excluded from further analyses and are not included within the library.

Table 4 displays an overview of the established library, which contains 502 barcode compliant sequences providing coverage for 88 different arthropod species belonging to 28 families, of 7 distinct orders. The predominant and most diverse order within the library is Diptera, scoping with a total of 452 specimens over

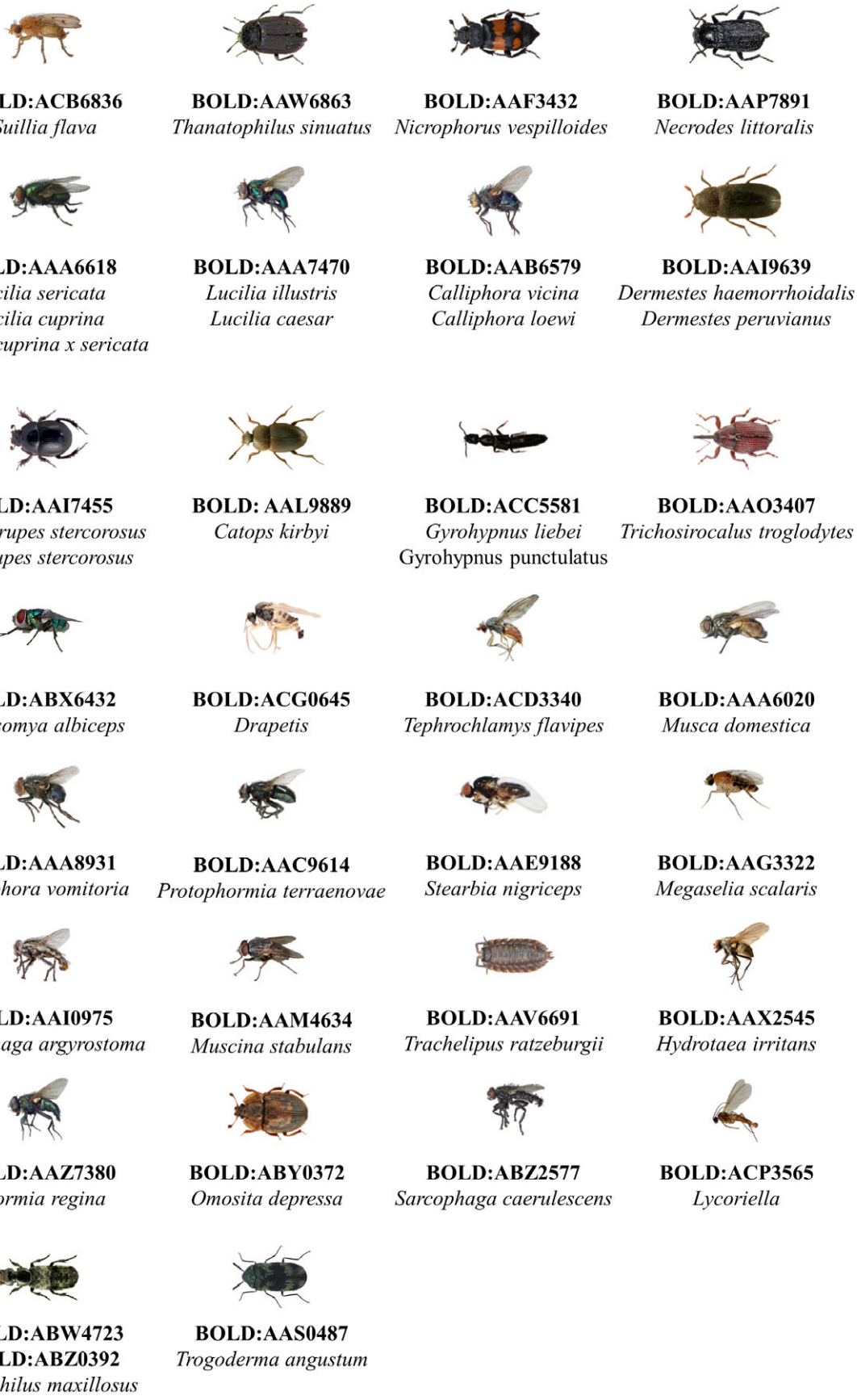


FIG. 2—*Visualization of the 31 detected BINs spanning throughout the bulk samples.*

70 different species belonging to 16 different families. The top three dipteran families in terms of abundancy and biodiversity are, in decreasing order, Calliphoridae (187 specimens, 15 species), Muscidae (73 specimens, 14 species), and Fanniidae (70 specimens, 7 species). The second most abundant order is Coleoptera, with a total of 41 successfully barcoded specimens representing 14 different species from 5 different families. With 14 specimens assigned to 5 different species, Silphidae is the family with the highest biodiversity and greatest abundance. The orders Hymenoptera, Mecoptera, Pseudoscorpiones, Isopoda, and Araneae are also present within the library, although represented by very few to only single specimens (≤5). The entire reference library is displayed in Table 5.

As accurate results of a DNA barcoding analysis can only be guaranteed in cases where a comprehensive and detailed library is used as a backbone, the establishment of this sequence reference library represents an important step toward the common usage of DNA barcoding in forensic entomology. Extending this library through the ongoing addition of newly barcoded species in the framework of future DNA applications would further increase the robustness of species identification. Although our library is yet, with its 88 arthropod species, of moderate size, it is still useful as many up to date studies within forensic entomology focus their research on single to few arthropod species and/or groups, whereas here, we provide an extensive and broader foundation for further applications. With the use of our database as a backbone to future metabarcoding analyses, sequenced specimen data can be easily and quickly compared to our library for rapid identification. This is especially interesting as entomologists sampling from decomposing material are often confronted with immature arthropod stages such as eggs or larvae, with arthropod residue such as exuviae or partial remains, or simply with unidentified arthropod biomass which are extremely difficult to identify without the use of molecular biology.

### *BINs – Barcode Index Numbers*

The international Barcode of Life Database groups sequences into clusters of closely similar COI barcode sequences which are assigned to a globally unique identifier, termed a BIN. Members of a BIN often belong to a single species as delineated by traditional taxonomy (53). Whereas most of the designated BINs (78.76%) were assigned to one unique species only, 24 (24%) BINs comprise more than one species. However, in most cases of "BIN sharing," single CO1 barcodes can be used to identify species within low divergent sequence clusters. All cases of unique BINs and BINs with multiple species are summarized in Table 6. Of the 502 specimens constituting the forensic reference library, 436 (87%) were assigned to a BIN. The highest rate is found among Coleoptera, where all 41 specimens (100%) distributed over 5 different families were successfully associated to a BIN. Dipteran specimens were assigned to a BIN in 87% of cases (Table 7).

### *HTS Analysis and Reference Library Application*

The analysis of the 30 arthropod bulk samples collected from human bodies revealed a total of 31 arthropod BINs. In addition to private and public arthropod data on BOLD, the newly established forensic reference library was used for the identification of the HTS amplicon data. Here, 12 (38.7%) BINs were detected solely through the application of our library in comparison to the data obtained from BOLD, suggesting that identification gaps may occur when applying HTS directly to bulk samples and neglecting prior library construction.The 31 identified BINs provide coverage for arthropod species spanning through 16 different families belonging to the orders of Coleoptera, Diptera, and Isopoda. BINs were predominantly attributed to Diptera (58%) and Coleoptera (39%). Only one BIN was assigned to an isopod species. The most abundant family throughout all orders is Calliphoridae, which was assigned to a total of 7 different BINs. While most BINs are assigned to one species, respectively, some are shared by more than one species, e.g., the BIN BOLD:AAA7470, which is assigned to *Lucilia illustris* and to *Lucilia caesar* or the BIN BOLD:AAA6618, which is assigned to *Lucilia sericata* and to *Lucilia cuprina* (Fig. 2).

Overall, the community structure within the bulk samples shows a clear domination of dipteran over coleopteran species: the Diptera-Coleoptera ratios span from a minimum of 55.60% to a maximum of 100% between the single samples. Diptera displayed a rather high minimum quote of 76%. While some BINS are very abundant throughout the samples, the top three being BOLD:AAA7470 (96.7%), BOLD:AAA6618 (96.7%), and BOLD:AAB6579 (93%), which were detected on nearly all corpses independent of the location of death, other BINs display a very low abundancy, having been recorded only at distinct locations.

To conclude, with a sturdy reference library at hand, the application of HTS is innovative as it enables the analysis of hundreds and thousands of individuals as a whole and the generation of large datasets which can be then evaluated at will. Although we have only tested this method on a small-scale, it is clear that HTS will become an essential tool for future large-scale purposes.

### *Acknowledgments*

We would like to thank Tobias König and Athena Wai Lam for their support on the various laboratory procedures, which were undertaken throughout this study.

## References

1. Alibegović A. Cartilage: a new parameter for the determination of the postmortem interval? J Forensic Leg Med 2014;27:39–45.
2. Amendt J, Krettek R, Nießen G, Zehner R. Forensische entomologie: ein handbuch [Forensic entomology: an introduction]. Frankfurt, Deutschland: Verlag für Polizeiwissehenschaft, 2013.
3. Gennard DE. Forensic entomology: an introduction. Chichester, England: John Wiley & Sons, 2007.
4. Tarone AM, Sanford MR. Is PMI the hypothesis or the null hypothesis? J Med Entomol 2017;54(5):1109–15.
5. Joseph I, Mathew DG, Sathyan P, Vargheese G. The use of insects in forensic investigations: an overview on the scope of forensic entomology. J Forensic Dent Sci 2011;3(2):89–91.
6. Byrd JH, Lord WD, Wallance JR, Tomberlin JK. Collection of entomological evidence during legal investigations. In: Byrd JH, Castner JL, editors. Forensic entomology: the utility of arthropods in legal investigations. 2nd edn. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2009;127–76.
7. Wells JD, Stevens J. Application of DNA-based methods in forensic entomology. Annu Rev Entomol 2008;53:103–20.
8. Raupach MJ, Hendrich L, Küchler SM, Deister F, Morinière J, Gossner MM. Building-up of a DNA barcode library for true bugs (Insecta: Hemiptera: Heteroptera) of Germany reveals taxonomic uncertainties and surprises. PLoS ONE 2014;9(9):e106940.
9. Hebert PDN, Cywinska A, Ball SL, Dewaard JR. Biological identifications through DNA barcodes. Proc Royal Soc B: Biol Sci 2003;270 (1512):313–21.
10. Hendrich L, Morinière J, Haszprunar G, Hebert PDN, Hausmann A, Köhler F, et al. 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 2014;15(4):795–818.
11. Spelda J, Reip HS, Oliveira-Biener U, Melzer RR. Barcoding Fauna Bavarica: Myriapoda – a contribution to DNA sequence-based identifications of centipedes and millipedes (Chilopoda, Diplopoda). ZooKeys 2011;156:123–39.
12. Ratnasingham S, Hebert PDN. BOLD: the Barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 2007;7(3):355–64.
13. Schmidt S, Schmidegger C, Morinière J, Haszprunar G, Hebert PDN. DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoideapartim). Mol Ecol Resour 2015;15(4):985–1000.
14. Schmidt S, Taeger A, Morinière J, Liston A, Blank SM, Kramp K, et al. Identification of sawflies and horntails (Hymenoptera, 'Symphyta') through DNA barcodes: successes and caveats. Mol Ecol Resour 2016;17(4):670–85.
15. Hausmann A, Haszprunar G, Hebert PDN. DNA barcoding the geometrid fauna of Bavaria (Lepidoptera): successes, surprises, and questions. PLoS ONE 2011;6(2):e17134.
16. Hausmann A, Haszprunar G, Segerer AH, Speidel W, Behounek G, Hebert PDN. Now DNA-barcoded: the butterflies and larger moths of Germany. Spixiana 2011;34(1):47–58.
17. Morinière J, Hendrich L, Hausmann A, Hebert PDN, Haszprunar G, Gruppe A. Barcoding Fauna Bavarica: 78% of the Neuropterida fauna barcoded!. PLoS ONE 2014;9(10):e109719.
18. Hawlitschek O, Morinière J, Lehmann GUC, Lehmann AW, Kropf M, Dunz A, et al. DNA barcoding of crickets, katydids, and grasshoppers (Orthoptera) from Central Europe with focus on Austria, Germany, and Switzerland. Mol Ecol Resour 2016;17(5):1037–53.
19. Morinière J, Hendrich L, Balke M, Beermann AJ, König T, Hess M, et al. A DNA barcode library for Germany's mayflies, stoneflies and caddisflies (Ephemeroptera, Plecoptera and Tricoptera). Mol Ecol Resour 2017;17(6):1293–307.
20. Astrin JJ, Höfer H, Spelda J, Holstein J, Bayer S, Hendrich L, et al. Towards a DNA barcode reference database for spiders and harvestmen of Germany. PLoS ONE 2016;11(9):e0162624.
21. Sperling FAH, Anderson GS, Hickey DA. A DNA-based approach to the identification of insect species used for postmortem interval estimation. J Forensic Sci 1994;39(2):418–27.
22. Wallman JF, Donnellan SC. The utility of mitochondrial DNA sequences for the identification of forensically important blowflies (Diptera: Calliphoridae) in southeastern Australia. Forensic Sci Int 2001;120(1–2):60–7.
23. Chen WY, Hung TH, Shiao SF. Molecular identification of forensically important blow fly species (Diptera: Calliphoridae) in Taiwan. J Med Entomol 2004;41(1):47–57.
24. Ames C, Turner B, Daniel B. The use of mitochondrial cytochrome oxidase I gene (COI) to differentiate two UK blowfly species – *Calliphora vicina* and *Calliphora vomitoria*. Forensic Sci Int 2006;164(2–3):179–82.
25. Jordaens K, Sonet G, Braet Y, De Meyer M, Backeljau T, Goovaerts F, et al. DNA barcoding and the differentiation between North American and West European *Phormia regina* (Diptera, Calliphoridae, Chrysomyinae). ZooKeys 2013;365:149–74.
26. Sonet G, Jordaens K, Braet Y, Bourguignon L, Dupont E, Backeljau T, et al. Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. ZooKeys 2013;365:307–28.
27. Schilthuizen M, Scholte C, van Wijk REJ, Dommershuijzen J, van der Horst D, Meijer zu Schlochtern M, et al. Using DNA-barcoding to make the necrobiont beetle family Cholevidae accessible for forensic entomology. Forensic Sci Int 2011;210:91–5.
28. Amendt J, Campobasso CP, Goff ML, Grassberger M, editors. Current concepts in forensic entomology. Dordrecht, Heidelberg, London, New York: Springer, 2010.
29. Szpila K. Key for identification of European and Mediterranean blowflies (Diptera, Calliphoridae) of medical and veterinary importance–adult flies. In: Gennard D, editor. Forensic entomology: an introduction, 2nd edn. Hoboken, NJ: Willey-Blackwell, 2012;77–81.
30. Laštůvka Z, Šťastná P, Suchomel J, Gaisler J. Zoologie [Zoology]. Brno, Czech Republic: Mendelova univerzita v Brně [Mendel University in Brno], 2015
31. Byrd JH, Castner JL. Insects of forensic importance. In: Byrd JH, Castner JL, editors. Forensic entomology: the utility of arthropods in legal investigations. Boca Raton, FL: CRC Press, 2000;43–79.
32. Erzinçlioğlu Z. Blowflies (naturalists' Handbooks No. 23). Slough, Berkshire, U.K.: Richmond Publishing Co. Ltd, 1996.
33. Williams KA, Villet MH. Morphological identification of Lucilia sericata, Lucilia cuprina and their hybrids (Diptera, Calliphoridae). ZooKeys 2014;420:69–85.
34. Ubero-Pascal N, López-Esclapez R, García MD, Arnaldos MI. Morphology of preimaginal stages of Calliphora vicina Robineau-Desvoidy, 1830 (Diptera, Calliphoridae): a comparative study. Forensic Sci Int 2012;219 (1–3):228–43.
35. Akbarzadeh K, Wallman JF, Sulakova H, Szpila K. Species identification of Middle Eastern blowflies (Diptera: Sarcophagidae) of forensic importance. Parasitol Res 2015;114(4):1463–72.
36. Feng DX, Liu DC. Pupal age estimation of forensically important Megaselia spiracularis Schmitz (Diptera: Phoridae). Forensic Sci Int 2013;231 (1–3):199–203.
37. Szpila K, Hall MJR, Pape T, Grzywacz A. Morphology and identification of first instars of the European and Mediterranean blowflies of forensic importance. Part II. Luciliinae. Med Vet Entomol 2013;27 (4):349–66.
38. Szpila K, Hall M, Sukontason K, Tantawi T. Morphology and identification of the first instar larvae of European and Mediterranean blowflies of forensic importance. Part I. Chrysomyinae. Med Vet Entomol 2013;27 (2):181–93.
39. Szpila K, Pape T, Hall MJR, Mądra A. Morphology and identification of the first instar larvae of European and Mediterranean blowflies of forensic importance. Part III: Calliphorinae. Med Vet Entomol 2014;28 (2):133–42.
40. Peacock ER. Adults and larvae of hide, larder and carped beetles and their relatives (Coleoptera: Dermestidae) and of derodontid beetles (Coleoptera: Derodontidae). Handbooks for the identification of British Insects. Vol. 5, Part 3. London, U.K.: Royal Entomological Society of London, 1993.
41. Háva J. World keys to the genera and subgenera of Dermestidae (Coleoptera), with descriptions, nomenclature and distributional records. Acta Musei Nationalis Pragae, Series B, Natural History 2004;60(3–4):149–64.
42. Ortloff A, Zanetti N, Centeno N, Silva R, Bustamante F, Olave A. Ultramorphological characteristics of mature larvae of Nitidula carnaria (Schaller 1783) (Coleoptera: Nitidulidae), a beetle species of forensic importance. Forensic Sci Int 2014;239:e1–9.
43. Frątczak K, Matuszewski S. Instar determination in forensically useful beetles Necrodes littoralis (Silphidae) and Creophilus maxillosus (Staphylinidae). Forensic Sci Int 2014;241:20–6.
44. Frątczak K, Matuszewski S. Classification of forensically-relevant larvae according to instar in a closely related species of carrion beetles (Coleoptera: Silphidae: Silphinae). Forensic Sci Med Pathol 2016;12(2):193–7.
45. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc Natl Acad Sci U S A 2004;101(41):14812–7.
46. Silva-Brandão KL, Wahlberg N, Francini RB, Azeredo-Espin AML, Brown KS, Paluch M, et al. Phylogenetic relationships of butterflies of the tribe Acraeini (Lepidoptera, Nymphalidae, Heliconiinae) and the evolution of host plant use. Mol Phylogenet Evol 2008;46(2):515–31.
47. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33 (7):1870–4.
48. Ratnasingham S, Hebert PDN. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS ONE 2013;8(7): e66213.
49. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, et al. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front Zool 2013;10(1):34.
50. Morinière J, Cancian De Araujo B, 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;11(5):e0155497.
51. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26 (5):680–2.
52. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 2012;28(12):1647–9.
53. Hausmann A, Godfray HCJ, Huemer P, Mutanen M, Rougerie R, et al. Genetic patterns in European geometrid moths revealed by the barcode index number (BIN) system. PLoS ONE 2013;8(12):e84518.

## Publication IV - High Throughput Sequencing as a novel quality control method for industrial yeast starter cultures

Citation: Michel, M., Hardulak, L. A., Meier-Dörnberg, T., Morinière, J., Hausmann, A., Back, W., Haszprunar, G., Jacob, F., and Hutzler, M. (2019). “High throughput sequencing as a novel quality control method for industrial yeast starter cultures.” *BrewingScience* 72 (March/April): 63-68.

M. Michel, L. A. Hardulak, T. Meier-Dörnberg, J. Morinière, A. Hausmann, W. Back, G. Haszprunar, F. Jacob and M. Hutzler

# High Throughput Sequencing as a novel quality control method for industrial yeast starter cultures

**The use of pure starter cultures is largely responsible for the great success of today's fermentation products market. With their implementation, fermentations (e.g., beer fermentation) have become predictable, efficient, controllable and reproducible. The pureness of the applied starter culture is therefore of great interest for obtaining these advantages. Many protocols have been applied for quality control of the pureness of these cultures, and they have improved over the last century. The current state of the art of detecting interfering microorganisms consists of the use of selective media or targeted approaches via Real-Time PCR. These methods are time consuming and require suspicion of the identity of potential interfering microbe(s). The use of High Throughput Sequencing, however, offers the ability to apply a non-targeted approach for the detection of interfering spoilers in the applied case of spoilage yeasts. Here we used the 26S rDNA D1/D2 region of chromosome XII to verify the pureness of yeast cultures applied in brewing, wine and special beer fermentations. The results show that it is possible to detect differing species in supposedly pure yeast cultures by application of the new method. Some strains showed potential traits of intraspecific hybridization, horizontal gene transfer or syntrophic cultures, which interfered with the results. The 26S rDNA D1/D2 region showed to be discriminative for only some species, indicating the need to additionally apply more discriminative regions like ITS1. Furthermore, we propose a more comprehensive and powerful database, consisting of highly validated and identified cultures, that has to be built up to improve sequencing results.**

Descriptors: HTS, quality control, yeast, fermentation, pure cultures

## 1 Introduction

Pure and defined microbiological starter cultures, also called defined strain starters (DSS), are a valuable tool for predictive and controllable industrial fermentations [1]. DSS are defined as consisting of one or more strains of one or more species [1; 2]. Since the implementation of pure brewing yeast cultures by Emil Christian Hansen in 1883 [3], one of the first pure culture fermentation approaches, the global amount of fermentation volume over all industrial applications has increased rapidly. The total value of the global fermentation products market reached $ 149,469 million in 2016 and is predicted to grow further to $ 205,465 million by the year 2023 [4].

**https://doi.org/10.23763/BrSc19-05michel**

**Authors**

Maximilian Michel, Tim Meier-Dörnberg, Fritz Jacob, Mathias Hutzler, Research Center Weihenstephan for Brewing and Food Quality, TU München, Freising, Germany; Laura A. Hardulak, Bavarian State Collection of Zoology (SNSB-ZSM), Munich, Germany; Jérôme Morinière, Bavarian State Collection of Zoology (SNSB-ZSM), , Munich, Germany; AIM – Advanced Identification Methods GmbH, Munich, Germany; Axel Hausmann, Gerhard Haszprunar, Bavarian State Collection of Zoology (SNSB-ZSM), Munich, Germany; Department Biology II and GeoBioCenter, Ludwig-Maximilians-University, Planegg, Germany; Werner Back, TU München, Wissenschaftszentrum Weihenstephan, Freising, Germany; corresponding author: m.hutzler@tum.de

Pitching a pure culture of starter microorganism into a defined media results in a predictive process with a defined product [1]. Interference by other microorganisms, whether other species, or even just other strains from the same species will change the result of the fermentation. This will result in loss of product, inefficiency and potentially complete spoilage of batches, which in the worst case must be discarded [5]. Yeast strains as valuable fermenting microorganisms are used for many industrial fermentations. Purity of the applied yeast strains is important for processes such as wine, beer, and ethanol production, as well as to produce proteins in the biopharma industry [4; 6; 7]. For as long as DSS have been implemented, the ability to control the purity of yeast strains has been indispensable. In the last centuries and decades, the level of quality control of fermentations started to improve with the refinement of the microscope, became more sufficient with various selective enrichment cultivation media and physiological tests, and reached a high point with the implementation of Real-Time-PCR systems [8]. All these methods were developed to identify potentially differing microbes, which interfere with the purity of a culture or a culture fermentation. Every novel method has lowered the level of detection and increased the purity of fermentations.

As an example for potential detection of interfering spoilage yeasts, one may look at pure yeast cultures for beer fermentations. Analysis by microscope is restricted to a small amount of sample at a time, and high-level experience and knowledge is needed to identify

potentially different cell morphologies. Moreover, low levels of contamination by spoilage yeast in a crowded yeast sample can be detected only with great difficulty [9]. When using selective media to identify potential interfering yeasts in the fermentation process, the main culture is suppress by defined additives (e.g. antimycotics). If this additive also suppresses the interfering microorganism, or if the desired microorganism is resistant against the additive, a false negative or false positive result may be the outcome. Real-Time-PCR is tied to primers and probes, which are used to detect spoilers by targeted approaches. If a target is unknown, spoilage yeasts can potentially be detected by applying certain numbers of targeted approaches simultaneously in a process called multiplex systems. However, if the interfering spoilage yeast is not on the target list, high costs and a failed detection is the result [10].

In contrast, High Throughput Sequencing (HTS) can be applied as an untargeted approach of identifying potential spoilage yeasts in pure yeast starter cultures without prior knowledge of the identity of the potentially contaminating organism. For our application, the large subunit 26S rDNA D1/D2 region was chosen for the first approach as it contains variable species-specific sequences [11]. Prior to HTS, Sanger sequencing was the state of the art sequencing technique for many years. Now, HTS offers the ability to sequence multiple amplicons of one gene fragment, in contrast to classical Sanger sequencing. The latter sequenced the most frequently occurring fragment, which could then be linked to one species at a time. Less frequently occurring sequences were mostly not visible and could therefore go undetected [12]. HTS enables sequencing of differing amplicons, providing the opportunity to detect multiple species in the same sample at the same time [13; 14]. Recent results from medical research also show that it might be possible in the future to differentiate between strains of a single species as well [11; 15].

**Table 1 Applied yeast strains of this study, strain abbreviation, species and applied fermentations purpose**

| Strain abbreviation | Species | Fermentation purpose |
|---|---|---|
| TUM 193 | *Saccharomyces pastorianus* | Lager yeast |
| TUM 68 | *Saccharomyces cerevisiae* | Wheat beer yeast |
| TUM 127 | *Saccharomyces cerevisiae* | Wheat beer yeast |
| TUM 177 | *Saccharomyces cerevisiae* | Koelsch, Ale yeast |
| TUM 184 | *Saccharomyces cerevisiae* | Alt, Ale yeast |
| TUM 506 | *Saccharomyces cerevisiae* | Ale yeast |
| TUM 211 | *Saccharomyces cerevisiae* | Ale yeast |
| TUM 511 | *Saccharomyces cerevisiae* | Ale yeast |
| TUM 381 | *Saccharomyces cerevisiae* | Trappist, Ale yeast |
| TUM T 90 | *Torulaspora delbrueckii* | Special beers |
| TUM 523 | *Hanseniaspora uvarum* | Banana wine |
| TUM 536 | *Schizosaccharomyces pombe* | Special beers |
| TUM Bretta 1 | *Brettanomyces bruxellensis* | Lambic yeast |
| TUM SL 17 | *Saccharomycodes ludwigii* | Low alcohol wheat beer yeast |
| TUM V 1 | *Saccharomyces cerevisiae* | Wine fermentation |
| TUM V 12 | *Saccharomyces cerevisiae* | Wine fermentation |
| TUM V 2 | *Saccharomyces cerevisiae* | Wine fermentation |
| TUM 5-2-1 | *Kazachstania unispora* | Spoilage yeast |

The created fragment sequences are further processed to OTUs (Operational Taxonomic Units) at a length of about 150–250 bases and compared to publicly available databases (such as NCBI; see section Bioinformatics). As the large subunit 26S D1/D2 region is of interest due to its relatively high species/strain specificity, many reference sequences are already available. To generate a proof of this concept, HTS was applied as a new untargeted quality control tool for purity of yeast DSS. This HTS approach was applied to a total of 20 pure yeast samples and one pooled sample (nine commercially available pure *Saccharomyces* brewing yeast cultures, five pure non-*Saccharomyces* yeast cultures applied for special beers, three *S. cerevisiae* wine cultures and one isolated environmental spoilage yeast culture). The pooled sample was created to test the recovery of three species out of a pool of varying unknown species. The large subunit 26S rDNA D1/D2 region was used in this proof of principle test, but any other species-specific region such as ITS1 could also be applied [11].

## 2 Materials & Methods

### 2.1 Applied yeast strains

Table 1 lists the yeast strains that were used in this study. Pure strains were cultivated on wort agar slopes for 72 hours at 28 °C and stored in a sterile environment at 2–4 °C. The strains were subcultured at intervals of one month.

For the pooled sample, a mixture of TUM 211, TUM 523, TUM 5-2-1 and spontaneous growing yeast species was set up. For this purpose, samples of the pure cultures were added to a sample of 50 ml of spontaneously fermented wort at 12 °P (cool wort was left open prior to inoculation at an open window for one day). Fermentation was performed for 2 days at 28 °C. 1 ml of the sample was taken sterile, and DNA was extracted as described in paragraph 2.2..

### 2.2 DNA extraction & High-Throughput Sequencing

Samples were taken from wort agar slopes with sterile inoculation loops and transferred into 1.5 mL Eppendorf tubes. 1 ml of the pooled wort sample was taken sterile and transferred into a 1.5 mL Eppendorf tube. DNA extraction was performed using the Qiagen DNeasy Plant Mini kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Cleaned and extracted DNA was then used as a template to amplify the D1/D2 domain of the 26S rRNA gene using NL1 (5`-GCATATCAATAAGCGGAGGAAAAG-3`) and NL4 (5`-GGTCCGTGTTTCAAGACGG-3`) (95 °C/5 min; 35 cycles of 95 °C/30 s, 52 °C/60 s; 72 °C/60 s; 72 °C/10 min), primers were equipped with a binding region for Illumina index sequences [16]. In a second PCR reaction, successfully amplified samples were equipped with a unique combination of Illumina index sequences. For the construction of the Illumina libraries, we used the cleaned amplicons, which were pooled to equal molarity (100 ng). A preparative gel electrophoresis was used for size-selection of the Illumina libraries. High-Throughput Sequencing (HTS) was performed on

an Illumina MiSeq using v2 (2*250 bp, 500 cycles, maximum of 20 mio reads) chemistry.

### 2.3 Bioinformatics

Processing of sequences was performed with the VSEARCH v2.4.3 suite [17] and cutadapt v1.14 [18]. Only forward reads (approximately 234-bp long) were used for the analysis, due to low quality of the reverse reads preventing paired-end merging. Forward primers were removed with cutadapt, using the "discard_untrimmed" option to discard sequences for which the primer sequence was not reliably detected at >= 90% identity. Quality filtering was performed with the "fastq_filter" in VSEARCH, keeping sequences with zero expected errors ("fastq_maxee" 1). Sequences were dereplicated with "derep_fulllength", first at the sample level, and then as one entire fasta file. Chimeric sequences were filtered out using "uchime_denovo". "cluster_size" was used to cluster the remaining sequences into OTUs at 97% identity and create a contingency table of counts of reads per OTU per sample. To reduce noise, read counts were eliminated from the OTU table which were less than 0.01% of the total numbers of reads for their corresponding samples (see JAMP v3, https://github.com/VascoElbrecht/JAMP/). OTUs were BLASTed against the GenBank nucleotide database (nt) in Geneious (v9.1.7 – Biomatters, Auckland – New Zealand) program Megablast with default parameters. The resulting csv file, which included hit descriptions, taxonomy, Hit-%-ID value, and bit score, was exported from Geneious and combined with the OTU table. Graphs were created with OriginPro 2018b (OriginLab Corporation).

## 3 Results and Discussion

The following shows the results of the HTS sequencing of the 26S rDNA D1/D2 region of 20 pure yeast cultures and one pooled sample. The varying numbers of reads reflect varying levels of amplification success of the PCR reactions performed prior to sequencing. All percentage shares are calculated according to the number of reads for the corresponding sample. The results do not represent quantitative detection of species in the sample, they represent the amount of amplification and can therefore not give a defined rate of contamination. All applied OTU's were BLASTed against the NCBI GenBank nucleotide database (nt). Resulting species/strains can be found with respective accession number and percentage of pairwise identity in table 2. Accession number can be used to identify according sequences, projects and species names (https://blast.ncbi.nlm.nih.gov/Blast.cgi). As shown in table 2, some sequences of the database were labelled unspecific as uncultured fungus and uncultured *Saccharomycetaceae*, which indicates that an actual database for the here applied experiments needs to be augmented in order to produce reliable results. However, the database was sufficient for this first proof of principle.

**Table 2 OTU sequence results for according species and accession numbers**

| Pairwise identity [%] | Accession number | Species | OTU ID in this project |
|---|---|---|---|
| 100 | CP033481 | *S. cerevisiae* | 1 |
| 100 | MH443765 | *H. uvarum* | 4 |
| 100 | KY296084 | *Schizosaccharomyces pombe* | 7 |
| 100 | MK034127 | *T. delbrueckii* | 3 |
| 98.2 | JX409606 | *Uncultured Saccharomycetaceae* | 18 |
| 100 | KY109478 | *Saccharomycodes ludwigii* | 6 |
| 100 | KF908878 | *Dekkera anomala* | 11 |
| 97.9 | KY107593 | *Brettanomyces anomalus* | 99 |
| 98.3 | KF810069 | Uncultured yeast isolate ZB01142638 | 26 |
| 97.4 | KF810090 | Uncultured yeast isolate ZB05224068 | 28 |
| 99.1 | MG525064 | *Kazachstania unispora* | 15 |
| 100 | MG927742 | Uncultured fungus | 10 |
| 100 | MG773367 | *Wickerhamomyces pijperi* | 9 |
| 100 | KY558364 | *Kluyveromyces dobzhanskii* | 12 |

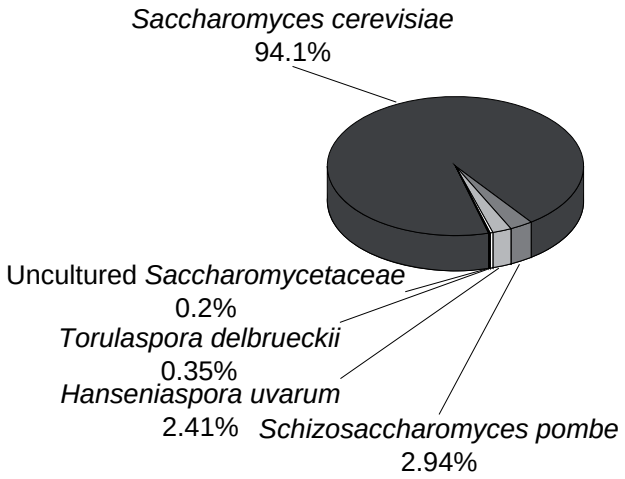


**Fig. 1 Composition of OTU's displayed in percent, detected for strain TUM 184 *Saccharomyces cerevisiae* culture (percentages result from total of 8946 reads)**

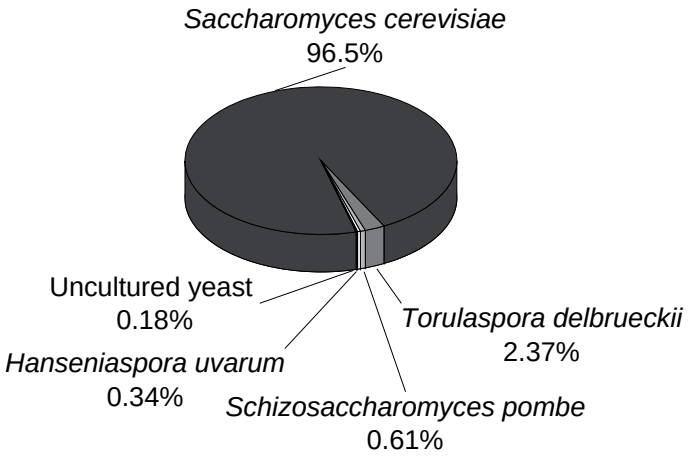


**Fig. 2 Composition of OTU's displayed in percent, detected for strain TUM 68 *Saccharomyces cerevisiae* culture (percentages result from total of 7251 reads)**

Completely pure cultures of single species were expected to show 100 % of one species. Sample TUM 211, TUM 381, TUM 177, TUM 506, TUM 511, TUM V12 and TUM V2 showed these expected results. A total of 100 % of OTU 1 (Table 2), *Saccharomyces cerevisiae* sequence were detected. TUM 211, TUM 381, TUM 177, TUM 506, TUM 511 are commercially available strains which are used to produce top fermented beers [19]. TUM V12 and TUM V2 are commercially available wine strains. The results indicate that the cultures were of pure *S. cerevisiae* without traces of other yeast species. The same positive result but with 100 % of *Torulaspora delbrueckii* (OTU 3) can be reported for the applied sample of TUM T 90. TUM T 90 is an alternative brewing yeast of the species *T. delbrueckii*, which was recently used for fermentation of novel beers [20].

As presented in figures 1 and 2, the reads of samples TUM 184 and TUM 68, showed 94.29 % respectively 96.5 % of *S. cerevisiae* OTU 1, but also 2.94 % respectively 0.61 % of *Schizosaccharomyces pombe* (OTU 7), 2.41 % respectively 0.34 % of *Hanseniaspora uvarum* (OTU 4), and 0.35 % respectively 2.37 % of *Torulaspora delbrueckii* (OTU 3). And in the case of TUM 184, 0.2 % of the reads corresponded to the class *Saccharomycetaceae* (OTU 18); whereas for TUM 68, 0.18 % of the reads corresponded to uncultured yeast (OTU 28).

Similar results as for TUM 184 and TUM 68 were detected with varying species for the supposed pure samples of TUM 127, TUM SL17, TUM 523, TUM 536, TUM Bretta 1 and TUM V1 visible in table 3.

These results can indicate different possible incidents with varying causes. Firstly, the culture may not have been as pure as thought prior to investigation. This could be due to contaminations of the pure cultures by other species. As the samples were not prepared from single colonies, it is likely that they were pure cultures of one strain of one species. The fact that more than one species was detected by HTS could potentially indicated the presence of syntrophic cultures. Syntrophic cultures are known to be very difficult to separate, as cells of different sizes adhere together [21]. The detected species are known to be able to grow in wort and might therefore be able to survive in close proximity with each other, making it hard to separate them on agar by traditional separation techniques. As *H. uvarum* represents a very fast-growing yeast species (generation time <1 h at 30 °C) [22], much faster than *Saccharomyces cerevisiae* strains (generation time <3 h at 30 °C), a prior negative detection of this contamination is very unlikely but cannot be definitively excluded. Other possible reasons are genetic changes on the ribosomal DNA of the *Saccharomyces* strains by differing impacts [23; 24]. One potential cause might be horizontal gene transfer as reported before by Xie et al. [25]. Another reason can be hybridization between the differing detected species [24]. As the 26S rDNA D1/D2 region is found in amounts of 100-200 tandem repeats on chromosome XII of many yeast species [26; 27], hybridization or horizontal gene transfer could lead to varying sequences during the tandem repeats [24; 25]. These variations may potentially be identified by whole genome sequencing which will be the task in further research on this topic.

The total reads of the fragments in the sample of strain TUM 193

**Table 3 Distribution of species results of the OTUs detected for six supposedly pure yeast culture samples**

| Strain | Supposed pure species | OUT ID | According species | Number of reads | Percentage of total reads [%] |
|---|---|---|---|---|---|
| TUM 127 | *S. cerevisiae* | 1 | *S. cerevisiae* | 5563 | 99.49 |
| | | 4 | *H. uvarum* | 19 | 0.33 |
| | | 7 | *S pombe* | 9 | 0.16 |
| TUM SL 17 | *Saccharomycodes ludwigii* | 6 | *S. ludwigii* | 5146 | 97.66 |
| | | 7 | *S. cerevisiae* | 123 | 2.44 |
| TUM 523 | *H. uvarum* | 4 | *H. uvarum* | 7017 | 99.84 |
| | | 1 | *S. cerevisiae* | 11 | 0.16 |
| TUM 536 | *S. pombe* | 7 | *S. pombe* | 2140 | 73.56 |
| | | 4 | *H. uvarum* | 769 | 26.43 |
| TUM Bretta 1 | *Brettanomyces bruxellensis* | 11 | *Dekkera anomala* | 1189 | 54.51 |
| | | 1 | *S. cerevisiae* | 599 | 27.46 |
| | | 6 | *Saccharomycodes ludwigii* | 299 | 13.7 |
| | | 99 | *B. anomalus* | 53 | 2.4 |
| | | 3 | *T. delbrueckii* | 30 | 1.37 |
| | | 4 | *H. uvarum* | 11 | 0.50 |
| TUM V1 | *S. cerevisiae* | 1 | *S. cerevisiae* | 5779 | 87.45 |
| | | 3 | *T. delbrueckii* | 668 | 10.10 |
| | | 26 | *Uncultured yeast* | 68 | 1.02 |
| | | 28 | *Uncultured yeast* | 56 | 0.84 |
| | | 7 | *S. pombe* | 25 | 0.37 |
| | | 4 | *H. uvarum* | 12 | 0.18 |

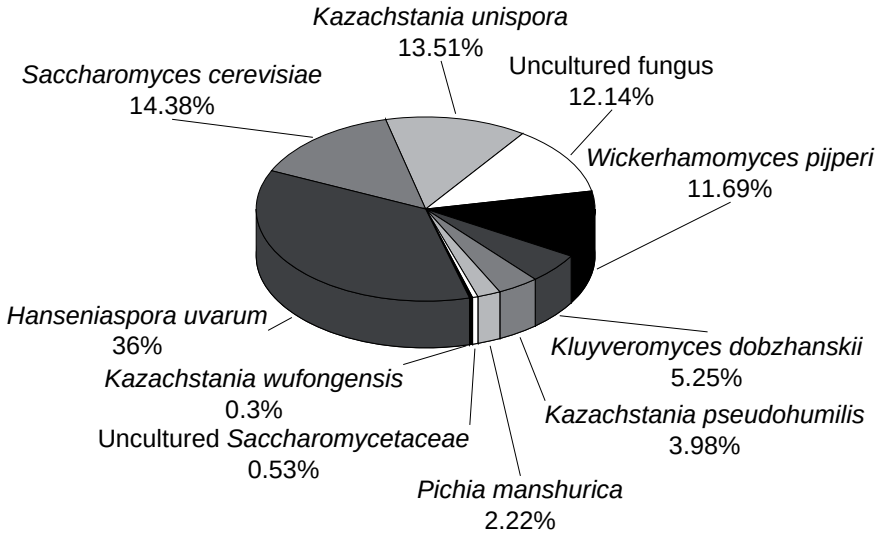


**Fig. 3 Composition of OTUs displayed in percent, detected for the pooled sample (percentages result from total of 6675 reads).**

showed 100 % of OTU 1 (*S. cerevisiae*). As this yeast strain is a bottom fermenting yeast of the species *S. pastorianus*, which is a hybrid of *S. cerevisiae* and *S. eubayanus* [28], it is concluded that the discovered OTU contained low or no amounts of the rDNA coming from *S. eubayanus* on this particular region. A small difference between these two species on the 26S rDNA D1/D2 region is possibly the case. When this OTU was BLASTed against *S. eubayanus* 26S rDNA D1/D2 it showed a similarity of 99 %. As mentioned above the use of the 26S rDNA D1/D2 region was only a first proof of principle. It will be followed by further investigations of the ITS1 region, which might be more discriminative and, in this case, more usable than the 26S rDNA D1/D2 region.

Figure 3 displays the results of the OTU distribution of the pooled sample. This sample was produced to create a mixture of unknown and known species. The sample contained inoculated strains of the species *S. cerevisiae*, *Kazachstania unispora* and *Hanseniaspora uvarum*. Further unknown interfering yeast species from the environment were present in the sample as it was spontaneously fermented by exposing the cold wort to the environment for one day. It is apparent that more than the inoculated species were detected (*K. pseudohumilis*, *Wickerhamomyces pijperi*, *Pichia manshurica*, *K. wufongenensis*) (Fig. 3). The inoculated species were reliably detected. The actual proof of detecting the pooled species was successful as visible in figure 3.

## 4 Conclusion/Summary

Defined strain starters (DSS) are one of the main factors contributing to the success of the fermentation industry. Most of the products created from fermentations are of high pureness and exceptional quality. The high quality is due to a predictable, efficient, controllable and reproducible fermentation performed with a defined starter [1]. Quality control of the purity of the starters is critical to the aussurance of high quality fermentation products. Over the last century, quality control has improved in sensitivity, speed, and reliability. New methods have always been the key to even higher quality. Over the last decades, molecular biological methods like Real-Time PCR and DNA-Fingerprinting have accelerated the improvements in quality control. HTS has proven to provide a major step forwards concerning detection of multiple species of yeast in one sample with unknown composition. Here, we demonstrated that the purity of yeast starters for beer, wine and special beers can be assured by HTS. As this was just a proof of principle, further adjustments to this method have to be performed in order to reliably detect interfering yeast species. Some results did not show the expected outcome, hinting at potential genetic variation in some yeast strains. As it cannot be completely excluded that the supposedly pure cultures had a certain amount of interfering yeasts, this will be analyzed in a further study. Despite these findings, the results indicate a new promising tool for non-targeted quality control. To improve further, other genetically diverse regions, such as internal transcribed spacers (e.g. ITS1) will be applied, as they promise to be more discriminative than the 26S rDNA D1/D region [29]. Furthermore, as the results of BLASTing against a public database in the present study have highlighted, a reference library for these specific regions is still needed in order for reliable identification at the species level to be fully achieved.

### Acknowledgements

This project was supported by grants from the Bavarian State Ministry of Science and the Arts (2009-2018: Barcoding Fauna Bavarica, BFB) and the German Federal Ministry of Education and Research (German Barcode of Life: 2012-2019, BMBF FKZ 01LI1101 and 01LI1501).

## 5 References

1. Speranza, B. (Ed.): Starter cultures in food production, Wiley Blackwell, Chichester (West Sussex), 2017. pp. 4-5.
2. Mozzi, F.; Raya, R. R. and Vignolo, G. M. (Eds.): Biotechnology of Lactic Acid Bacteria, Wiley-Blackwell, Oxford, UK, 2010. p. 7.
3. Hansen, E. C. and Klöcker, A.: Gesammelte theoretische Abhandlungen über Gärungsorganismen, G. Fischer, Jena, 1911.
4. Akhila Prasannan: Fermentation Products Market by Type (Alcohols, Amino Acids, Organic Acids, Biogas, Polymers, Vitamins, Antibiotics, and Industrial Enzymes), Feedstock (Corn, Rice, Wheat, Sugar Cane, Cassava, Barley, Potatoes, Sorghum, Sugar Beet,& Tubers), Process (Batch Fermentation, Continuous Fermentation), and End-user Industry (Food & Beverages; Pharmaceutical; Agriculture; Personal Care; Animal Feed; Textile & Leather) – Global Opportunity Analysis and Industry Forecast, 2017-2023, Allied Market Research (2017), https://www.alliedmarketresearch.com/fermentation-products-market ,last accessed 16.01.2019
5. Meier-Dörnberg, T.; Kory, O. Ingo; Jacob, F.; Michel, M. and Hutzler, M.: *Saccharomyces cerevisiae* variety diastaticus friend or foe?-spoilage potential and brewing ability of different *Saccharomyces cerevisiae* variety *diastaticus* yeast isolates by genetic, phenotypic and physiological characterization, FEMS yeast research, **18** (2018), no. 4, DOI:

10.1093/femsyr/foy023.

6. Basso, L. C.; Amorim, H. V. de; Oliveira, A. J. de and Lopes, M. L.: Yeast selection for fuel ethanol production in Brazil, FEMS yeast research, **8** (2008), no. 7, pp. 1155-1163.
7. Fleet, G. H.: Wine yeasts for the future, FEMS yeast research, **8** (2008), no. 7, pp. 979-995.
8. Hutzler, M.: Entwicklung und Optimierung von Methoden zur Identifizierung und Differenzierung von getränkerelevanten Hefen, Dissertation, TU München, München, Lehrstuhl für Technologie der Brauerei II, 2009.
9. Tubia, I.; Prasad, K.; Pérez-Lorenzo, E.; Abadín, C.; Zumárraga, M.; Oyanguren, I. et al.: Beverage spoilage yeast detection methods and control technologies: A review of *Brettanomyces*, International Journal of Food Microbiology, **283** (2018), pp. 65-76.
10. Stephenson, F. H.: Real-Time PCR: Calculations for molecular biology and biotechnology, Elsevier, 2016, pp. 215-320.
11. Colabella, C.; Corte, L.; Roscini, L.; Bassetti, M.; Tascini, C.; Mellor, J. C. et al.: NGS barcode sequencing in taxonomy and diagnostics, an application in "Candida" pathogenic yeasts with a metagenomic perspective, IMA fungus, **9** (2018), no. 1, pp. 91-105.
12. Woo, P. C. Y.; Leung, S.-Y.; To, K. K. W.; Chan, J. F. W.; Ngan, A. H. Y.; Cheng, V. C. C. et al.: Internal transcribed spacer region sequence heterogeneity in *Rhizopus microsporus*: implications for molecular diagnosis in clinical microbiology laboratories, Journal of clinical microbiology, **48** (2010), no. 1, pp. 208-214.
13. Chimeno, C.; Morinière, J.; Podhorna, J.; Hardulak, L.; Hausmann, A.; Reckel, F. et al.: DNA Barcoding in Forensic Entomology – Establishing a DNA reference library of potentially forensic relevant arthropod species, Journal of forensic sciences, **64** (2018), no. 2, pp. 1-9.
14. Morinière, J.; Cancian de Araujo, B.; Lam, A. Wai; 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, **11** (2016), no. 5, e0155497.
15. Zhang, N.; Wheeler, D.; Truglio, M.; Lazzarini, C.; Upritchard, J.; McKinney, W. et al.: Multi-Locus Next-Generation Sequence typing of DNA extracted from pooled colonies detects multiple unrelated Candida albicans strains in a significant proportion of patient samples, Frontiers in microbiology, **9** (2018), p. 1179.
16. Kurtzman, C. and Robnett, C.: Phylogenetic relationships among yeasts of the *Saccharomyces* complex - Complex determined from multigene sequence analyses, FEMS yeast research, **3** (2003), no. 4, pp. 417-432.
17. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C. and Mahé, F.: VSEARCH: a versatile open source tool for metagenomics, PeerJ, **4** (2016), e2584.
18. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet j., **17** (2011), no. 1, p. 10.
19. Meier-Dörnberg, T.; Hutzler, M.; Michel, M.; Methner, F.-J. and Jacob, F.: The importance of a comparative characterization of *Saccharomyces cerevisiae* and *Saccharomyces pastorianus* strains for brewing, Fermentation, **3** (2017), no. 3, DOI:10.3390/fermentation3030041.
20. Michel, M.; Meier-Dörnberg, T.; Jacob, F.; Schneiderbanger, H.; Haslbeck, K.; Zarnkow, M. and and Hutzler M.: Optimization of beer fermentation with a novel brewing strain *Torulaspora delbrueckii* using response surface methodology, TQ, **54** (2017), no. 1, pp. 23-33.
21. Morris, B. E. L.; Henneberger, R.; Huber, H. and Moissl-Eichinger, C.: Microbial syntrophy: interaction for the common good, FEMS microbiology reviews, **37** (2013), no. 3, pp. 384-406.
22. Deák, T.: Handbook of food spoilage yeasts, 2. ed., CRC Press, Boca Raton, Fla., 2008.
23. Fitzpatrick, D. A.: Horizontal gene transfer in fungi, FEMS microbiology letters, **329** (2012), no. 1, pp. 1-8.
24. Morales, L. and Dujon, B.: Evolutionary role of interspecies hybridization and genetic exchanges in yeasts, Microbiology and molecular biology reviews: MMBR, **76** (2012), no. 4, pp. 721-739.
25. Xie, J.; Fu, Y.; Jiang, D.; Li, G.; Huang, J.; Li, B. et al.: Intergeneric transfer of ribosomal genes between two fungi, BMC evolutionary biology, **8** (2008), p. 87.
26. Montrocher, R.; Verner, M. C.; Briolay, J.; Gautier, C. and Marmeisse, R.: Phylogenetic analysis of the *Saccharomyces cerevisiae* group based on polymorphisms of rDNA spacer sequences, International journal of systematic bacteriology, **48** Pt 1 (1998), pp. 295-303.
27. Espinar, M. T. Fernández; Martorell, P.; Llanos, R. de and Querol, A.: Molecular methods to identify and characterize yeasts in foods and beverages. In: Querol, A., Fleet, G. (Eds.): Yeasts in food and beverages, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 55-82.
28. Monerawela, C. and Bond, U.: The hybrid genomes of *Saccharomyces pastorianus*: A current perspective, Yeast (Chichester, England), **35** (2018), no. 1, pp. 39-50.
29. Kurtzman, C. P. and Robnett, C. J.: Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences, Antonie van Leeuwenhoek, **73** (1998), no. 4, pp. 331-371.

*Received 22 January 2019, accepted 20 March 2019*

# Acknowledgements

I would like to thank the following people for their valuable support during the completion of my thesis:

- My supervisors Prof. Dr. Gerhard Haszprunar and Dr. Axel Hausmann for guiding me throughout my PhD and helping me with writing, even when I had doubts.
- Jérôme Morinière, for always being the go-to person for help throughout my years spent at the ZSM. Your positive outlook, encouragement, and belief in your students' potential have been invaluable in helping me to reach my professional goals.
- Dr. Bruno Cancian de Araujo for being a kind, helpful, supportive, and knowledgeable colleague, as well as everyone else I have had the pleasure of working with at the ZSM.
- Dr. Matthias Geiger for your dedication in coordinating the GBOL project, including holding conferences and workshops.
- Dr. Ingrid Huber, Dr. Gesche Spielmann, and Dr. Ulrich Busch at the Bavarian Ministry of Health and Food Safety for giving us the opportunity to apply our methods in the food quality control industry.
- Prof. Dr. Jörg Müller and Olaf Schubert for making it possible to make a biodiversity survey of the Bavarian Forest National Park.
- Dr.-Ing. Mathias Hutzler and Dr.-Ing. Tim Meier-Dörnberg from the TUM Weihenstephan research center for brewing and food quality, for your collaboration on another exciting project.
- Dr. Frank Reckel and Dr. Jan E. Grunwald of the Bavarian Police Department for collaborating with us in forensic entomology.
- Caroline Chimeno for being a dedicated and supportive colleague. I am grateful I got to work with you and attend conferences together.
- Dr. Marina Querejeta Coma for diligently performing massive amounts of lab work while always having a positive attitude
- All of the interns and undergraduate students who worked with me in the laboratory, particularly Simona Balherr.
- Alexander Hausmann and Alejandro Izquierdo López for not giving up on our experiment and being determined to see it through with me.
- My family for supporting me in all of my endeavors, wherever my path has taken me.
- My boyfriend Vedran for always being supportive, caring, and knowing things. None of this is taken for granted.