The best of two worlds: toward large-scale monitoring of biodiversity combining COI metabarcoding and optimized parataxonomic validation

Benoit Penel, Christine N Meynard, Laure Benoit, Axel Boudonne, Anne-Laure Clamens, Laurent Soldati, Alain Migeon, Marie-pierre Chapuis, Sylvain Piry, Gael Kergoat, et al.

ECOGRAPHY

Research article

The best of two worlds: toward large-scale monitoring of biodiversity combining COI metabarcoding and optimized parataxonomic validation

Benoit Penel[1], Christine N. Meynard[1], Laure Benoit[2], Axel Boudonne[1], Anne-Laure Clamens[1], Laurent Soldati[1], Alain Migeon[1], Marie-Pierre Chapuis[2], Sylvain Piry[1], Gael Kergoat[1] and Julien Haran[2]

[1]CBGP, INRAE, CIRAD, IRD, Institut Agro, Univ Montpellier, Montpellier, France
[2]CBGP, CIRAD, INRAE, IRD, Institut Agro, Univ Montpellier, Montpellier, France

Correspondence: Penel Benoit (penelbenoit@gmail.com)

Ecography
2025: e07699
doi: 10.1111/ecog.07699
Subject Editor: Andres Baselga
Editor-in-Chief:
Dominique Gravel
Accepted 16 December 2024

In a context of unprecedented insect decline, it is critical to have reliable monitoring tools to measure species diversity and their dynamic at large-scales. High-throughput DNA-based identification methods, and particularly metabarcoding, were proposed as an effective way to reach this aim. However, these identification methods are subject to multiple technical limitations, resulting in unavoidable false-positive and false-negative species detection. Moreover, metabarcoding does not allow a reliable estimation of species abundance in a given sample, which is key to document and detect population declines or range shifts at large scales. To overcome these obstacles, we propose here a human-assisted molecular identification (HAMI) approach, a framework based on a combination of metabarcoding and image-based parataxonomic validation of outputs and recording of abundance. We assessed the advantages of using HAMI over the exclusive use of a metabarcoding approach by examining 492 mixed beetle samples from a biodiversity monitoring initiative conducted throughout France. On average, 23% of the species are missed when relying exclusively on metabarcoding, this percent being consistently higher in species-rich samples. Importantly, on average, 20% of the species identified by molecular-only approaches correspond to false positives linked to cross-sample contaminations or mis-identified barcode sequences in databases. The combination of molecular methodologies and parataxonomic validation in HAMI significantly reduces the intrinsic biases of metabarcoding and recovers reliable abundance data. This approach also enables users to engage in a virtuous circle of database improvement through the identification of specimens associated with missing or incorrectly assigned barcodes. As such, HAMI fills an important gap in the toolbox available for fast and reliable biodiversity monitoring at large scales.

Keywords: biodiversity assessment, Coleoptera, molecular operational taxonomic units, morphological validation, recognizable taxonomic units, taxonomic impediment

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Insects represent 61–71% of the known eukaryotic biodiversity (Adler and Foottit 2017) and despite their critical importance in ecosystem functioning (Goulson 2019), they remain far less well monitored than other organisms such as vertebrates (Goulson 2019). This situation arises mostly from the complexity associated with species recognition in general, and limited availability of taxonomic expertise (Hoagland 1996, McGill et al. 2016, Engel et al. 2021). This trend is also evident within insect taxa (Troudet et al. 2017).

Over the last decades, two main strategies have been developed to face the taxonomic impediment and limit taxonomic biases associated with large-scale monitoring of insects. The first strategy is to simplify species identification by assigning individuals into recognizable taxonomic units (RTU; Oliver and Beattie 1993). This simple categorization of individuals into morphologically uniform groups, on the basis of external characteristics, is fast, easy to implement, and has therefore been widely used in ecological studies (Krell 2004). It produces relatively accurate data on species richness, but the resulting datasets are generally study-dependent and have little or no other application, since the specific composition of samples is not inferred (Krell 2004). The second strategy arose with the development of DNA-based identification methods such as DNA barcoding (Hebert et al. 2003). Its evolution into metabarcoding – a high-throughput approach – enables the simultaneous identification of multiple species in pooled samples, based on the amplification and sequencing of universal markers acting as molecular barcodes (Epp et al. 2012). These methods, notably those relying on free environmental DNA (eDNA), are now proposed as a theoretically effective way of rapid, and large-scale biodiversity monitoring (Taberlet et al. 2018, Dornelas et al. 2019). When based on well-curated reference DNA-barcode libraries, metabarcoding is fast, highly accurate for species-level identification and can be applied to any development stage, tissue fragment or free DNA available for an insect community (Liu et al. 2020, Chua et al. 2023). However, this approach is technologically demanding and faces multiple intrinsic limitations impacting their generalization for biodiversity monitoring. First, abundance data cannot be reliably inferred with this method. While several studies show a positive correlation between biomass or DNA quantity in a sequenced specimen/sample and the number of sequenced reads (Luo et al. 2023), this alone is insufficient to estimate how DNA degradation state and primer biases affect DNA amplification patterns within and between samples (Elbrecht and Leese 2015, Lamb et al. 2019). Yet, providing reliable abundance data is at least as critical as estimating species richness in biodiversity monitoring. Abundance provides valuable insights on population variation across locations and taxa (van Klink et al. 2022), which are crucial parameters for implementing management and conservation measures (Lacasella et al. 2017, Callaghan et al. 2024). Abundance is also required to detect population declines and range contractions before species extinctions. Second, factors such as DNA quality (Hawthorne et al. 2023), cross-contaminations (Drake et al. 2022), and PCR primer efficiency (Elbrecht and Leese 2015) can cause false positive or false negative identifications that are difficult to control, especially when samples are not verified afterwards (destructive methods, but see Batovska et al. 2021) or are not available (eDNA approaches). Finally, molecular approaches remain at a relatively early stage of development, with no bioinformatic consensus on best practices for avoiding errors (e.g. data filtering, Creedy et al. 2022), and reference DNA-barcode libraries are incomplete for most insect groups.

Despite the complementary nature of human verification of samples and abundance recovery on one hand, and fast and accurate species-level barcoding identification on the other, these two approaches are currently always used separately. Pereira et al. (2021) proposed an automatic pipeline combining identification data produced in parallel with metabarcoding and taxonomic approaches. Though this method provided reliable and cross-validated identifications, it still relies on taxonomic expertise, and it is therefore difficult to apply on a large scale. In this study, we introduce the human-assisted molecular identification (HAMI) framework for insect monitoring and macro-arthropods in general. We developed and tested HAMI on a large dataset of relatively diverse, but also well-known temperate beetle communities (ca 500) from agricultural field margins in continental France. This framework takes full advantage of the speed and reliability of DNA metabarcoding approaches and combines them with key parataxonomic checks to act as safeguard for molecular identification and to recover reliable abundance data. Here we describe the HAMI framework, highlight its strengths over traditional approaches, and discuss its applicability in various contexts.

Material and methods

Sample collection

Beetle communities were collected as part of the 500-ENI network, a national, standardized effort to monitor biodiversity in ca. 500 agricultural field margins across the French mainland (Andrade et al. 2021, Supporting information). Agricultural lands occupy 34% of the Earth's emerged surface (Ramankutty et al. 2008), and field margins represent an important dispersal and refuge habitat for species of conservation interest (Marshall and Moonen 2002), making it an interesting case-study to implement large-scale biodiversity monitoring strategies. In this context, field margin beetles were collected three times a year in spring using sweeping nets along two 10 m transects, which were positioned 30 m apart in the same field margin for a given site. Live specimens were then collected from the nets using a mouth aspirator, killed directly and preserved in 96% ethanol. Beetles collected on the same day at the same site from both transects were pooled and constitute a single sample in the following analyses. Details of the spatio-temporal characteristics of the samples can be found in the Supporting information.

The HAMI framework

The aim of the HAMI framework is to leverage metabarcoding to identify rapidly large volumes of mixed insect samples, in combination with parataxonomic expertise to increase the accuracy of results and incorporate abundance data. The framework includes four main steps: pre-sorting of specimens into RTUs and high-definition imaging (Fig. 1A), DNA amplification and sequencing of mixed samples using a standard metabarcoding approach (Fig. 1B), bioinformatics processing of sequence data for accurate molecular species identification (Fig. 1C), and a final step of visual data reconciliation between molecular outputs and RTUs (Fig. 1D).

RTU sorting and imaging

This step relies on parataxonomic expertise and can be carried out by one or several people with general knowledge in beetle

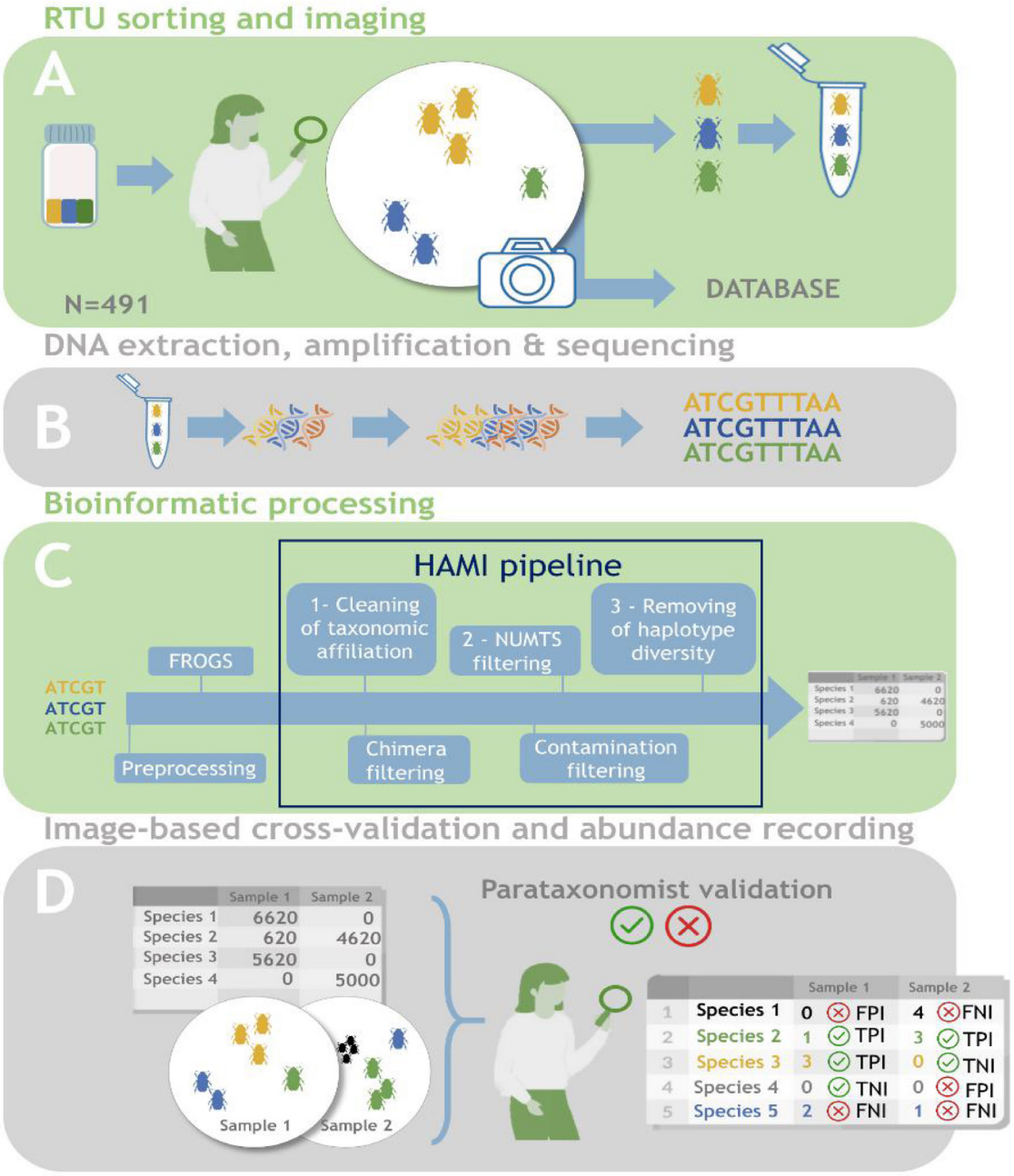

Figure 1. Description of the four stages of the HAMI methodology applied to our beetle samples (n = 492). Human iconography highlights steps where a parataxonomist is involved. The values shown in the occurrence matrix in 1C are the read numbers associated with each identification; values in the occurrence matrix in 1D represent the corresponding abundance data filled by matching the sequencing results with the stored photography of the same sample.

taxonomy. The goal of this first step is to separate individuals into groups of putative species, but being able to identify species is not needed. Training, based on pre-sampling of already identified specimens from collections, could be considered to guarantee the best sorting resolution, but was not used here. However, if several parataxonomists with heterogeneous taxonomic skills participate in the same effort, training is recommended to ensure consistency.

For each sample, specimens were first removed from alcohol, dried at room temperature and sorted into RTUs on a standard Petri dish under a binocular magnifier. Mixed samples of beetles from the same site generally consisted of an assemblage of species from divergent lineages, and a sample rarely contained closely related species belonging to the same genus. In case of doubt (intraspecific variation, sexual dimorphism or closely related species), distinct morphs were separated into different groups to obtain only homogeneous RTUs. Once sorted, the Petri dish was photographed using a Panasonic DMC-LX100 camera, mounted on a fixed base with a light maximizing device to maintain high-definition. Photographs were also taken on a millimetric paper to retain information on specimens' size (Supporting information).

Then, in line with our subsidiary objective of enriching the COI databases, and given that the RTUs often consisted of a single specimen, a protocol was implemented in order to conserve DNA-grade material for each species. Thus, as far as possible, a sole leg was taken from each RTU and transferred to a single tube for DNA extraction (bulk samples). However, when the RTUs were specimen-rich or tied to a complex fauna with potential cryptic species, we sampled several individuals to avoid species lumping. In addition, to limit PCR competition during the amplification step (Veillat et al. 2023), the volume of tissue included for each RTU was balanced at this stage: for very large specimens (> 10 mm), only a part of a leg was taken, while for very small specimens (< 2 mm), whole bodies were transferred to extraction tubes. The remaining specimens were stored in 96% ethanol at 4°C. To control the risk of cross-contamination between samples, the forceps used were disinfected using a Bunsen burner between each sample processed. In total, 492 beetle samples were processed.

Non-destructive DNA extraction, amplification, and sequencing

Bulk samples were subjected to a two-step PCR metabarcoding strategy (as detailed in Galan et al. 2018). DNA from each bulk sample was extracted using a 96-well plate animal genomic DNA miniprep kit (Biobasic) following the manufacturer's instructions, with an overnight lysis step to ensure extraction of hard-bodied beetle DNA. Based on the comparative study of Elbrecht et al. (2019) and preliminary tests conducted on mock communities constructed from our samples, the primer pair BF3+BR2 (called hereafter 'BB'; CCHGAYATRGCHTTYCCHCG/TCDGGRTGNCCRAARAAYCA; Elbrecht and Leese 2017, Elbrecht et al. 2019) was identified as the most suitable for cytochrome *c* oxidase subunit I (COI) metabarcoding of beetle communities (targets a 418 bp COI fragment that is part of the 658 bp universal COI barcode fragment). For library construction, we used the protocol detailed in Galan et al. (2018) with slight modifications (Supporting information). PCR1 were duplicated for each sample to ensure repeatability, and in each plate two negative controls were systematically incorporated for extraction (NEC) and for PCR (NPCRC). The resulting libraries were checked by electrophoresis on a 1.5% agarose gel, pooled and cleaned using AMpure beads (Beckman Coulter, CA, USA) with a volume ratio of 0.8X. Libraries were then paired-end sequenced in two different runs with the MiSeq Reagent Kit ver. 2 (500 cycles).

Bioinformatic processing

The bioinformatic workflow used here combined standard denoising and filtering steps for Illumina amplicon sequencing data and three data reduction steps specific to the HAMI (Fig. 1C): 1) cleaning of BOLD taxonomic affiliations; 2) filtering of nuclear mitochondrial pseudogenes (NUMTs); and 3) reduction of intraspecific haplotype diversity.

Reads were processed using FROGS, a user-friendly pipeline for analyses of large sets of DNA amplicons (Escudié et al. 2018), with the exception of the preprocessing step that fails for data produced following the dual index method of Kozich et al. (2013). Instead, we used a Shell script, available at https://doi.org/10.5061/dryad.sj6mf40 (Sow et al. 2019), which merges paired end reads into contigs (using Flash – ver. 1.2.6, Magoč and Salzberg 2011) and trims primers (with Cutadapt – ver. 1.8.3; Martin 2011). As a first step, FROGS retained reads with the expected length while accommodating for potential indel events (418 ± 3, 6 or 9 bp). In a second step, the SWARM algorithm (Mahé et al. 2014), was used to perform read clustering, leading to the formation of molecular operational taxonomic units (MOTUs), with a maximum sequence difference set at $d=1$ in order to obtain the finest partition. In a third step, MOTUs from chimeric sequences, or artificial DNA sequences resulting from the erroneous assembly of several DNA sequences produced during the two PCR steps, were removed using *VSEARCH* with de novo UCHIME method (Rognes et al. 2016). MOTUs were then subjected to molecular identification on the BOLD database (all arthropod sequences extracted from this database in August 2023, Supporting information). Using BLAST+, alignments between each MOTU and the database were produced (Camacho et al. 2009). Only the best hits with the same score are reported. If taxonomic affiliation does not provide a single identification among the best hits, this step returns multi-affiliation outputs to highlight possible conflicts and uncertainties in the databases. Identifications were also associated with an identification quality index (percentage of identity) for each taxonomic affiliation performed. All metrics for molecular identification were summarized in a table containing the number of reads associated with each affiliation in each sample. After FROGS processing,

following the two-step approach of Chapuis et al. (2023), we applied the more sensitive '*isBimeraDenovo*' R function (from DADA2 ver. 1.28.0; Callahan et al. 2016) to the FROGS output to remove residual chimeras (Supporting information). In parallel, we cleaned up BOLD taxonomic nomenclature errors in taxonomic ranks (Fig. 1C, (1)). This is based on the deletion of special characters, numbers, excessive number of words, and the standardization of unknown names and uncertainties in taxonomic affiliation (Fig. 1C, (1)).

Then, we carried out a contamination filtering step following the strategy proposed by Galan et al. (2016) and automated in Chapuis et al. (2023) with R scripts available at https://doi:10.18167/DVN1/D31UAV. Noise due to index switching during sequencing was removed by excluding occurrences in each sample whose read count was less than 0.002% of the cumulative total.

MOTUs with incongruous occurrences in duplicated PCR samples or with counts below the maximum for negative control samples (NEC and NPCRC) were eliminated as they were considered contaminations (Supporting information).

Particular attention was paid to removing divergent NUMT sequences (mitochondrial pseudogenes; Fig. 1C, (2)) which may introduce bias in molecular identification due to their old nuclear origin and divergence with the functional COI copy (Song et al. 2008). Divergent NUMTs generally result in non-functional amino acid sequences or stop codons or include frame shifts. Sequences with these characteristics were filtered out (Supporting information) as recommended by Song et al. (2008) and Creedy et al. (2022).

Finally, in order to simplify the output file, redundant MOTUs resulting from intraspecific variability were merged (Fig. 1C, (3)). MOTUs sharing the same taxonomic affiliations and falling within the standard percentage identity range of [97–100%] were merged together, while those with a percentage of identity below 97% were merged into a separate unique MOTU. MOTUs above the 97% arbitrary threshold were considered 'reliable' identifications because they were closely related to a reference COI barcode. Other MOTUs could either represent species for which barcode or divergent haplotypes are missing from the COI databases or recalcitrant NUMTs (Schultz and Hebert 2022) that have passed the filters. They were automatically discarded if, and only if, they were totally redundant with 'reliable' MOTUs, i.e. associated to the same taxonomic reference and occurring in the same samples (=NUMTs). Otherwise, they could be informative of biological variation, and for this reason were specifically 'flagged' MOTUs and retained in the final abundance table for morphological verification during the cross-validation step. Based on the recommendations of Creedy et al. (2022), the post-FROGS bioinformatics steps were designed in a user-friendly pipeline (https://github.com/BenoitPenel/HAMI) written in Python (van Rossum 1995) according to the Snakemake rules (Mölder et al. 2021) to ensure reproducibility.

Image-based cross-validation of metabarcoding identifications

In this final key step, the parataxonomist reconciliate MOTUs with the pre-sorted RTUs by examining the high-definition images of the samples (Fig. 1D). For each sample, when MOTUs affiliations aligned with delimited RTUs (true positive identification; TPI), records were associated with a visual count of their actual abundance in the samples and recorded in the occurrence table. MOTUs associated with a species absent in a given sample (false positive identification; FPI) were excluded. Conversely, RTUs that were not recovered by metabarcoding (false negative identification; FNI) were recorded and identified to the lowest taxonomic level (family, tribe, genus) based on the closest molecular taxonomic affiliation, the skills of the parataxonomist, and available resources (Supporting information).

The beetle fauna of western Europe is quite well characterized, both in terms of COI barcode libraries and images available in books or on the web. As such, an accurate identification of RTUs, according to MOTUs, can be obtained for the large majority of species, and an overview of their appearance can be easily verified, thus facilitating the reconciliation step by the parataxonomist. The composition of a beetle sample from a single locality (high phylogenetic diversity, rarely containing closely related species) also facilitates this step. In the case of complex speciose groups (e.g. Curculionoidea), the reconciliation step was validated by a second parataxonomist. Closely related species found together in a sample and indistinguishable on images were recorded at the genus level only (two species complexes in our case: *Captation seniculus* with *C. meieri* and *Oulema duftschmidi with O. melanopus*).

Comparing metabarcoding alone versus HAMI

The performance of HAMI, in terms of the diversity of species recorded in samples, was evaluated by a comparison with outputs of the metabarcoding protocol alone. Visual verification of samples in HAMI enables calculating the number and rate of true positive, false negative and false positive identifications in metabarcoding data alone.

TPI was calculated as the number of species identified by the molecular approach (with a threshold of 97%) and validated by the parataxonomist. The TPI rate, or the ability of molecular approaches alone to provide a correct species identification, i.e MOTU is the same as RTU in a sample, was calculated using the following formula:

$$\text{TPI rate} = \text{TPI}/(\text{TPI} + \text{FNI}_{\text{total}}) \quad (1)$$

The total number of FNI ($\text{FNI}_{\text{total}}$) was calculated as the sum of two types of FNI. These cases refer to the species identified by the parataxonomist but either: 1) not recorded at all by metabarcoding, linked to a failure of DNA amplification (hereafter referred to as FNI_{fail}), 2) or discarded because identification exceeded the 3% divergence threshold with a reference barcode sequence, in accordance with standard COI

metabarcoding practices (hereafter referred to as $FNI_{database}$). The FNI_{total} rate, or the rate of RTUs recovered by morphological cross-validations of HAMI in a sample, was calculated as:

$$FNI_{total}\ \text{rate} = FNI_{total}/(FNI_{total} + TPI) \quad (2)$$

FPI was calculated as the number of species identified with metabarcoding alone (percentage of identity higher than 97%) while it is not present in a sample. The FPI rate, or the rate of unfounded MOTUs that were rejected by morphological cross-validations of HAMI in a sample, was calculated as follows:

$$\text{FPI rate} = FPI/\left(FPI + TPI\right) \quad (3)$$

Finally, we used a GAM model (*gam* function from 'mgcv' ver. 1.9-0; Wood 2011) to see if we could find a linear or non-linear relationship between the raw species richness estimated by metabarcoding alone, versus the one estimated with HAMI.

Results

Of the 492 samples processed, over 11 700 specimens were examined and preliminary sorted into 4800 RTUs (i.e. the sum of the number of RTUs present in each sample – 10 ± 6 RTU/sample). Of all the sequences produced during the two sequencing runs, 5 775 155 and 3 061 634 sequences were retained, representing 33 678 and 77 424 MOTUs, respectively.

The dechimerization step associated to *isBimeraDenovo* functions lead to an exclusion of 99 284 and 116 628 sequences respectively and diminish the number of MOTUs to 25 322 and 55 152. The data filtering step associated to PCR duplicate and negative control has led to a slight reduction in the number of sequences but leading to a sharp fall in the number of MOTUs (Supporting information). At this stage, 6046 (85.7%) of the MOTUs in the first sequencing run and 5882 (81.1%) in the second have species-level affiliation, of which 3751 (53.2%) and 3805 (52.5%) corresponded to beetle species with more than 97% similarity to a sequence in the BOLD database. This represented 457 valid species once the HAMI cross-validation process was completed.

On average, samples consisted of 24 ± 30 specimens, corresponding to 9 ± 6 species (maximum: 32). One nitidulid species *Brassicogethe aenus* accounted for 18.6% of the specimens encountered (2193 specimens). Conversely, 143 of the beetle species recorded were only represented by a single specimen, and 59 species by only two specimens in the entire dataset.

Metabarcoding alone compared to HAMI

The average sensitivity of molecular approaches (correct identification of species), was estimated at 76% of species in a given sample (Fig. 2). A total of 360 samples out of 492 (73%) were affected by FNIs. FNIs represented an average of 24% of the species composition of each sample, or just over two species per sample (2.6 spp/sample). 40% of FNI cases stemmed from low alignment (< 97%) with reference linked to the lack of barcode and/or haplotype diversity in the national database, while 60% were due to the absence of amplification of the specimen's DNA (Fig. 2). Four families were particularly impacted by the $FNI_{database}$, accounting for more than 50% of the $FNI_{database}$ cases (Chrysomelidae 20.8%, Curculionidae 16.8%, Staphylinidae 12.0% and Latridiidae 8.8%). As for FNI_{fail} cases, i.e. species whose DNA was not amplified, more than 50% of the cases were associated with three families (Chrysomelidae 23.2%, Curculionidae 20.9% and Coccinellidae 10.50%).

FPIs were observed in 367 of 492 samples (74.7%) when using the molecular approach alone. On average, 22% of the specific composition identified by the molecular approach, or around two species per sample (2.2 spp./sample), were false positives. Interestingly, the genus *Leptomias* (Curculionidae) was identified in 76 of the 492 samples analyzed by the molecular approach, with perfect or near-perfect [100%; 97.12%], match between the sequenced DNA sequence and the reference COI barcode. This genus is known from Asia and was never encountered during the morphological examination by the parataxonomist, nor is it a species studied in our laboratory. This taxonomic misidentification corresponds to a *Wolbachia* endosymbiont widely found in beetles and erroneously published under the name *Leptomias* (OM830079.1 and OM830078.1). Another instance of FPI relates to the detection of a species of tropical weevil studied in the laboratory *Elaeidobius kamerunicus* in 10 out of 492 samples, probably due to contaminants during the PCR amplification process.

Overall, when considering the species richness of samples (thus excluding composition), the correlation between the estimates provided by metabarcoding alone and the actual richness validated by the parataxonomist is high but not perfect (Person's correlation: 0.79, p value < 0.001). The GAM model shows a non-linear relationship between metabarcoding estimates and HAMI estimates of species richness, where species-poor samples (ca < 20 species) have a better linear correlation, and where metabarcoding underestimated species richness in species-rich samples (Poisson GAM, k = 3, AIC = 2466, p-value < 0.01 – Fig. 3).

Discussion

Ecologists are increasingly using metabarcoding approaches (Creedy et al. 2022, Chua et al. 2023). But, as with many fields in the era of Big Data, a major challenge now lies in our ability to refine the quality of data produced (Hortal et al. 2015). In this study, we propose a first formal metabarcoding framework reintegrating the expertise of parataxonomists to produce qualitatively and quantitatively reliable data on biodiversity. This approach takes the best of both worlds, speeding up species identification via metabarcoding, while

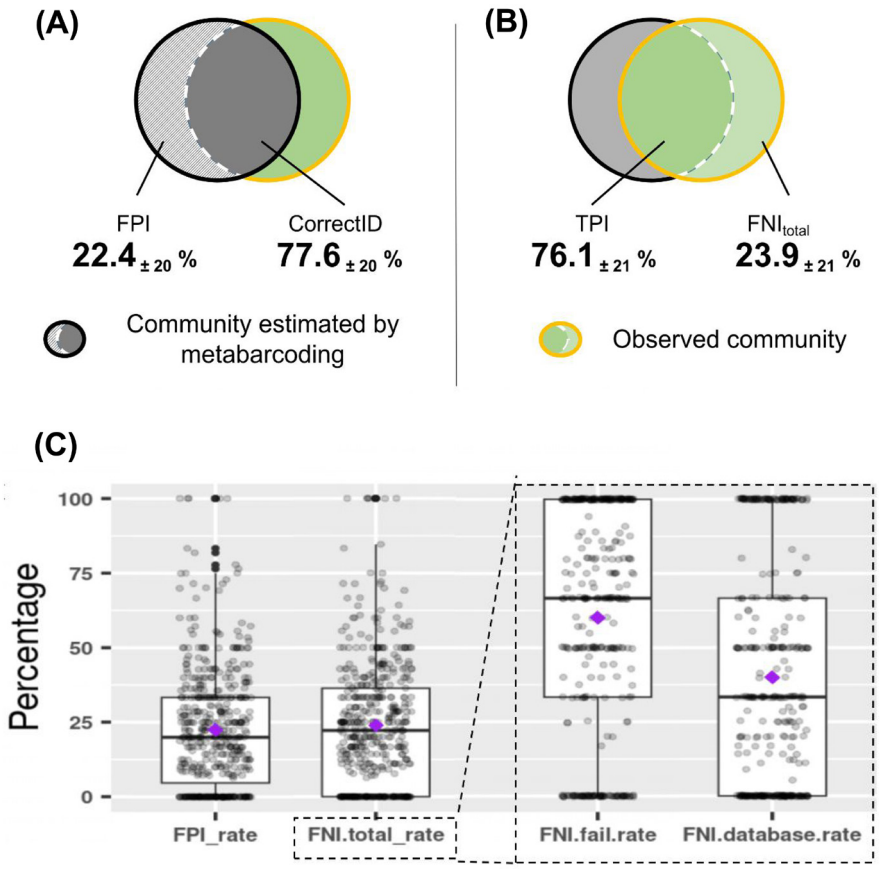

Figure 2. Performance of the metabarcoding approach to identify field margin beetle communities of 500-ENI network agricultural plots (n = 491). (A) Representation of the average proportion of identification produced by molecular approaches that are fictitious (FPI = False positive identification) and reliable (CorrectID = 1-FPI). (B) Representation of the average proportion of species correctly identify by molecular approaches on a given sample (TPI = true positive identification) and the proportion of species that have been missed (FNI_{total} = total number of false negative identification). (C) Boxplot of the FPI rate and FNI_{total} rate across samples with a decomposition of FNI_{total} according to the two possible types of FNI ($FNI_{database}$ = false negative identification linked to database incompleteness and FNI_{fail} = false negative identification linked to species not amplified).

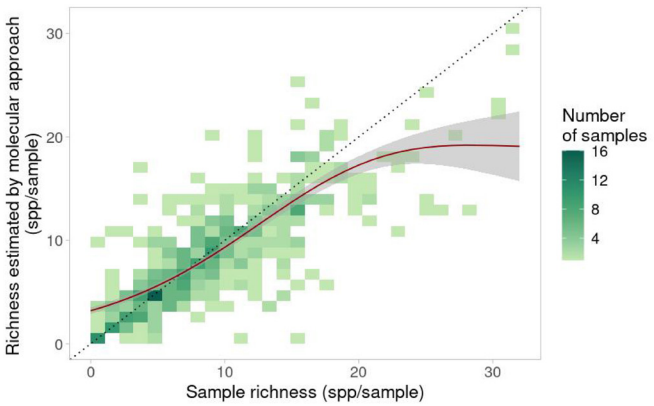

Figure 3. Relation between richness estimated by molecular approach from the one observed with HAMI framework. The black dotted line represents the expected x = y relation, while the full red is the predictive evolution of molecular richness estimation according to our GAM model and the number of species on a given sample (Poisson GAM, n = 492, k = 3, AIC = 2466, p-value < 0.001).

reintroducing human expertise into key steps to validate molecular identification and provide reliable abundance data. The advantages and limitations of this approach are detailed below.

Advantages and constraints of metabarcoding in a biodiversity monitoring context

Traditionally, morphology-based techniques have been used to monitor insects. However, these identification approaches are time-consuming, labor-intensive and require skilled entomologists, who are progressively disappearing (Engel et al. 2021). Over the past two decades, we have witnessed a gradual shift away from morphology-based approaches to DNA-based techniques (Piper et al. 2019). Applied to biodiversity monitoring, the development of high-throughput sequencing technologies has revolutionized the field due to their speed and cost-effectiveness. These attributes enabled species-level identification of thousands of beetle specimens from a set of 492 samples in just two months in our study-case. Such data

would have required hundreds of additional hours and a vast network of taxonomists to be produced based on morphology alone.

However, our results show that when applied to natural and highly diverse insect communities, metabarcoding is less effective than had been envisioned when this method was first proposed (Epp et al. 2012). In our dataset of 492 beetle samples with an average diversity of 9 ± 6 species, many species were missed (on average 2.6 species) by the molecular approach alone (type II error or FNI_{total}). The sensitivity estimated here (76%) is lower than what can be observed for insect mock communities using COI (e.g. 100% in Batovska et al. 2021 or 98% in Kocher et al. 2017). However, estimates similar to those inferred here have also been made on arthropod samples (Mata et al. 2021 [82.5% for 'mixture' sample] – calculated using their supplementary data and formula 1). The later estimate is also based on artificial communities (i.e. the 'mixture' sample which excluded singleton species), but includes a higher species diversity (40 species in the 'mixture' sample of Mata et al. 2021 versus less than 10 species in Kocher et al. 2017 and Batovska et al. 2021). These observations suggest that studies evaluating the effectiveness of molecular approaches with mock communities are not necessarily unrealistic when mock communities are sufficiently rich. These observations are in line with our metabarcoding sensitivity results and the conclusion of Duke and Burton (2020): the more species-rich a sample is, the greater the chances to miss species. To explain this, we can list the biases of the PCR primers that amplify some species more than others (Elbrecht and Leese 2015), variations in DNA quality across specimens and samples (Hawthorne et al. 2023) and database completeness (Keck et al. 2023).

Our results also show that the false-negative rate is highly dependent on database completeness, bioinformatics tasks and the user's choice of parameters. Despite a clear trend towards increased use of COI metabarcoding, the field remains at a relatively early stage of development (Creedy et al. 2022). This is particularly true when it comes to accepting or rejecting molecular identification. Indeed, molecular approaches rely on barcode database completeness and the barcoding gap hypothesis that posits that intraspecific genetic distance is smaller than interspecific distance, and that this gap could be used as a threshold to discriminate between species (Fig. 4, case 1 – Hebert et al. 2003). The standard threshold for the COI fragment in arthropods, including beetles, is 3% (e.g. Coccinellidae; Huang et al. 2020). Yet, intraspecific genetic distance can vary considerably across taxa (Fig. 4. case 2 – Avtzis et al. 2019; or case 3 – Ma et al. 2022) and may even overlap with interspecific genetic distance (Fig. 4. case 4 – Wiemers and Fiedler 2007). As such, a species with high intraspecific genetic divergence (Fig. 4, case 3) but for which few divergent haplotypes are available in barcode databases may be rejected by using metabarcoding alone (Meier et al. 2008). Although in the future we expect to see a reduction of such FNI cases because barcode databases will become more complete, the continued use of a single DNA fragment and threshold value for species identification remains inherently problematic (Wizenberg et al. 2023). This is

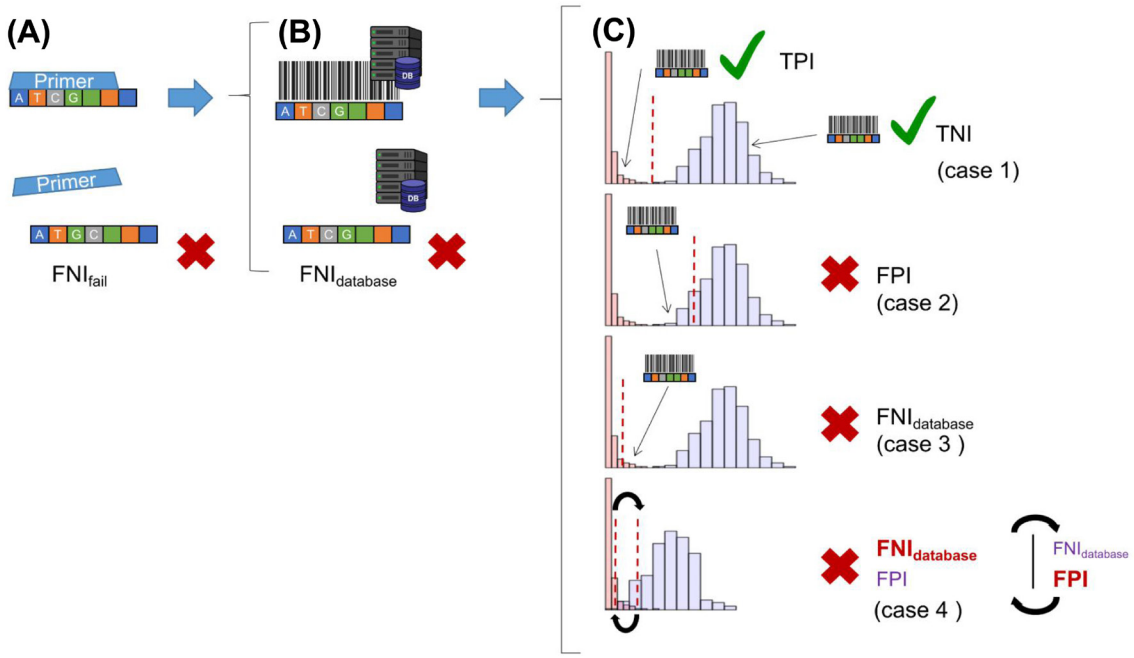

Figure 4. Iconography of the main source factors at the origin of false negative identification (FNI) and false positive identification (FPI). (A) Representation of the primer-bias issue leading to non-amplification of some specific DNA sequences. (B) Representation of the incompleteness of barcode databases compromising taxonomic identification of amplicons. (C) Representation of the relationship between FNI and FPI ratio according to the haplotype availability and the arbitrary threshold of 3% (illustrated by the red dotted line) used to distinguish intraspecific from interspecific distance in a molecular identification protocol (histograms).

why alternative, but also more complex, approaches using multiple thresholds (Arribas et al. 2021) or multiple fragments (Wizenberg et al. 2023) are emerging. In the meantime, the visual cross-validation implemented in HAMI also overcomes this flaw of traditional COI metabarcoding by 'rescuing' RTUs (here 420) with divergent haplotypes during the image cross-validation phase.

Furthermore,despite the use of a standardized protocol for specimen sampling and preservation, common and large-bodied (around 10 mm) beetle specimens, for which no amplification bias is known, were sporadically undetected by molecular approaches (13% of the FNI_{fail}). Our results further illustrate that there is a level of heterogeneity in DNA degradation between samples that poses a significant challenge to the metabarcoding approach when used alone (Hawthorne et al. 2023). In such cases, small variations in DNA quality, associated with the stochastic process of PCR amplification of mixed samples, are a plausible explanation. However, in contrast to the previous types of FNI, no long-term improvement is expected in reducing this type of FNI, despite it being the largest source (accounting for 60% of FNIs here). We therefore emphasize that even common species playing a major role in ecosystems (Winfree et al. 2015) can be overlooked in such assessments. This is a second critical aspect of the use of COI metabarcoding approaches alone for biodiversity monitoring, favoring the idea of a systematic morphological control of outputs. These types of shortcomings are also common in environmental DNA (eDNA) approaches for biodiversity monitoring. Efforts to compensate for the absence of specimens are very rarely implemented in this non-invasive method (Lyet et al. 2021, Mirimin et al. 2021, both using eDNA and camera/video trap) but may introduce significant biases in diversity estimates.

While several factors lead to a significant fraction of FNI, our study also highlights the prevalence of false positives (FPI), i.e. species that are not in the sample but are detected by the molecular methods, when metabarcoding is used alone. Here, 22% of the specific composition identified by the molecular approach, or approximately two species per sample, were false positives. In this study, the vast majority of beetle species are phytophagous or predators of non-beetle groups. We also mainly used legs for DNA extraction, so it is unlikely that these FPIs correspond to co-amplified gut content of the specimens. These fictitious identifications are likely the result of unavoidable cross-contaminations that occur between samples during the molecular laboratory protocol (Drake et al. 2022) despite the implementation of best practices to avoid them (i.e. single-use consumables, disinfection of forceps with a Bunsen burner and separate pre-PCR and post-PCR workspaces; Taberlet et al. 2018). To a lesser extent, contamination may also come from the working environment (Drake et al. 2022), as illustrated by the low prevalence of FPI due to contaminations from other beetle species studied in the same laboratory (1.3% of FPI). When contamination involves species from distant biomes, these cases can be easily filtered out. But when the FPI involves species from the same region, identification of a potential FPI is impossible without morphological verification of the specimens. FPIs also often result from errors in the taxonomic affiliation associated with the reference barcode sequences. In our case, a reference barcode assigned to a tropical beetle species was in fact the result of *Wolbachia* co-amplification and inadequate quality control during barcode production and submission. This type of bias persists for a significant fraction of molecular databases (Mioduchowska et al. 2018) and shows the importance of incorporating morphological controls into such approaches, as well as curated reference databases. Finally, despite the filtering steps, mitochondrial pseudo-genes (i.e NUMTs) cannot be excluded as factors of FPIs. Recalcitrant NUMTs remain a limitation, even for insect communities which, with a few exceptions, are less affected by NUMTs (Hebert et al. 2023) than most other animal communities (Schultz and Hebert 2022). However, the implementation of more complex filters (Andújar et al. 2021, Noguerales et al. 2023) is a possible area for improvement, in addition to morphological control.

Benefits of integrating parataxonomic checkpoints in metabarcoding

Neither the species richness nor the species composition estimated by the molecular approach alone are totally consistent with what is observed in a given sample (Fig. 2–3). The implementation of a parataxonomic visual check in the metabarcoding protocol significantly improved the resolution of the biodiversity assessment and species identifications in the samples. It resulted in the exclusion of an artefactual specific richness (22% of the richness estimated by metabarcoding – FPI) and to rescue 24% of the sample's species richness (FNI) among the 492 samples studied. The improvement in data quality was achieved by image-based verification, which greatly reduced sample handling time. On average, the total handling time was estimated at an hour (40 min for step A; 20 min for step D – Fig. 1) per sample, for beetle communities with a high variability in specimen count, averaging 24 ± 30 specimens, but sometimes reaching up to hundreds of individuals, and including up to 32 species (mean 9 ± 6). In this respect, the HAMI framework echoes the call by Dornelas et al. (2019) to build a 'Macroscope', by combining complementary tools and sources of information (in this case DNA and morphological approaches) to improve biodiversity monitoring (Pereira et al. 2021, Keck et al. 2022). The HAMI framework also echoes Engel et al. (2021), who argued that there is an urgent need to reconsider the place of taxonomic expertise in the study of biodiversity rather than blindly relying on technology alone. As well as generating high-quality data that can be linked to large-scale biodiversity inventory initiatives, HAMI also speeds up identification of common species, enabling parataxonomists to pinpoint and focus on problematic cases. Because HAMI highlights FNI, it can encourage the production of barcode sequences for species absent in databases (identified by taxonomists, with voucher specimens deposited in a reference collection;

Bourret et al. 2023), thus initiating a virtuous circle of improvement of the approach overall (Supporting information). Importantly, during this study, the parataxonomist also greatly improved his own taxonomic expertise by being repeatedly exposed to images of insects and their corresponding names. This could enable and accelerate better detection of inconsistencies in metabarcoding outputs, but it could also become a strategy for training new taxonomists, especially in areas where this skill is in short supply.

Aside from species identifications, HAMI stands out from conventional molecular biomonitoring methods as it allows recovering reliable abundance data. This represents a major step forward in large-scale biodiversity monitoring based on molecular methods. Although recent studies have shown a positive correlation between the biomass or the amount of DNA in a sequenced specimen/sample and the number of sequenced reads (Luo et al. 2023), the level of uncertainty in the estimates prevents this correlation from being applied to complex field samples (reviewed by Lamb et al. 2019). By accumulating quantitative data rather than qualitative presence/absence data, HAMI has overcome the previous limitation of the molecular approach. It offers a significant advance to monitor differences in population dynamics across various taxa and locations on a large scale (van Klink et al. 2022), which would benefit several research fields. As an example, such improvement will greatly benefit noxious species biomonitoring (i.e. invasive and/or pest species) by making better estimates of suitable settlement areas and priority management areas through abundance data (Lacasella et al. 2017). Reliable abundance data will also greatly support the field of conservation (Callaghan et al. 2024) by enabling efforts to be focused on specific populations, based on their abundance trends and the definition of appropriate policies.

Perspectives and future challenges

To address the challenges of the Anthropocene, the scientific community requires quick and reliable tools to monitor the evolution of biodiversity in a changing environment. While metabarcoding has undeniably accelerated the taxonomic identification of complex community samples (Piper et al. 2019), the rapid generation of a large amount of error-laden data hinders the creation of robust and accurate inferences. HAMI, which takes full advantage of the speed of metabarcoding and key parataxonomic verifications, has shown promising results in enhancing large-scale insect biodiversity monitoring, while remaining fast, affordable, and including reliable estimates of abundance. Its first application here in moderately diverse but also well-known system, represents an encouraging and revelant first step, since agricultural landscapes are associated with major insect declines (Wagner et al. 2021). Therefore, implementing large-scale monitoring strategies such as HAMI to help early detection of population declines and range shifts linked to these habitats is fundamental to propose coherent conservation strategies in productive environments. With relevant adaptation of this approach (e.g. appropriate primers, alternative barcode thresholds), HAMI can also be applied to multiple other contexts, such as large-scale monitoring of various insect clades, and the consideration of abundance over large spatial and temporal scales. Applying the HAMI approach to lesser-known communities, such as tropical environments, or communities with small species, and frequent cryptic species complexes, could admittedly be challenging. Yet, in the first case, HAMI can be applied without a strict barcode threshold to focus on MOTUs and images alone, which provides a relevant first approach to target poorly known faunas. The virtuous circle of COI database improvement promoted by HAMI can also be seen as an opportunity to enhance biodiversity discovery and taxonomist training in such cases. In regard to small, species-rich and morphologically homogeneous insect communities, the application of robotically automated individual imaging and Sanger barcoding, such as the DiversityScanner of Wührl et al. (2022), can also be implemented in the HAMI framework to improve its applicability in such cases. Furthermore, in the current context of rapid development of image analysis using artificial intelligence (AI), this protocol could also be further improved with automated recognition and counting of RTU in the images taken, as AI alone is not yet reliable for arthropods (Badirli et al. 2023). However, the efficiency of photo-based AI methods is highly sensitive to image quality (Fujisawa et al. 2023). Implementing an image stacking procedure as part of HAMI framework could be a first step towards improving the quality of photographs and support the combination with AI.

Acknowledgements – The 500-ENI network is funded by the French Ministry of Agriculture under the Ecophyto framework. We would like to thank everyone that has collected data in the field, the farmers who provided access, and everyone involved in the coordination of the 500-ENI data network.
Funding – This work was supported by the ANR AgriBiodiv (ANR-21-CE32-006-01) and an Ecophyto II+ project: GTP 500 ENI (OFB-21-1642).
Permits – Our study was carried out by scientists based in the same country as the study itself, respecting national data collection and sharing rules. Our working group strives to have a gender balance and be inclusive in all dimensions. We also lead a working group with relevant stakeholders within the 500 ENI network to encourage exchange and share results from our research.

Author contributions

Gael Kergoat and **Julien Haran** contributed equally to this publication. **Benoit Penel**: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Software (lead); Writing – original draft (equal). **Christine N. Meynard**: Conceptualization (equal); Funding acquisition (lead); Project administration (lead); Supervision (equal); Writing – original draft (equal). **Laure Benoit**: Methodology (equal); Resources (equal); Writing – original draft (equal). **Axel Boudonne**: Methodology (equal); Resources (equal); Writing – original draft (equal). **Anne-Laure Clamens**: Methodology (equal); Resources (equal); Writing – original draft (equal). **Laurent**

Soldati: Methodology (equal); Resources (equal); Writing – original draft (equal). **Alain Migeon**: Methodology (equal); Resources (equal); Writing – original draft (equal). **Marie-Pierre Chapuis**: Data curation (supporting); Software (supporting); Writing – original draft (equal). **Sylvain Piry**: Software (supporting); Writing – original draft (equal). **Gael Kergoat**: Conceptualization (equal); Supervision (equal); Writing – original draft (equal). **Julien Haran**: Conceptualization (equal); Methodology (lead); Supervision (lead); Writing – original draft (equal).

Transparent peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ecog.07699.

Data availability statement

All the scripts associated with the HAMI pipeline, the raw sequencing data and the R scripts used to analyze the data obtained are available at the ZENODO repository: https://zenodo.org/records/13760159 (Penel et al. 2024).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Adler, P. H. and Foottit, R. G. 2017. Introduction. – In: Footit, R. G. and Adler, P. H. (eds), Insect biodiversity. John Wiley & Sons, Ltd, pp. 1–7.

Andrade, C., Villers, A., Balent, G., Bar-Hen, A., Chadoeuf, J., Cylly, D., Cluzeau, D., Fried, G., Guillocheau, S., Pillon, O., Porcher, E., Tressou, J., Yamada, O., Lenne, N., Jullien, J. and Monestiez, P. 2021. A real-world implementation of a nation-wide, long-term monitoring program to assess the impact of agrochemicals and agricultural practices on biodiversity. – Ecol. Evol. 11: 3771–3793.

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P. and Emerson, B. C. 2021. Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. – Mol. Ecol. Resour. 21: 1772–1787

Arribas, P., Andújar, C., Salces-Castellano, A., Emerson, B. C. and Vogler, A. P. 2021. The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. – Mol. Ecol. 30: 48–61.

Avtzis, D. N., Lakatos, F., Gallego, D., Pernek, M., Faccoli, M., Wegensteiner, R. and Stauffer, C. 2019. Shallow genetic structure among the European populations of the six-toothed bark beetle *Ips sexdentatus* (Coleoptera, Curculionidae, Scolytinae). – Forests 10: art. 2.

Badirli, S., Picard, C. J., Mohler, G., Richert, F., Akata, Z. and Dundar, M. 2023. Classifying the unknown: insect identification with deep hierarchical Bayesian learning. – Methods Ecol. Evol. 14: 1515–1530.

Batovska, J., Piper, A. M., Valenzuela, I., Cunningham, J. P. and Blacket, M. J. 2021. Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches. – Sci. Rep. 11: art. 1.

Bourret, A., Nozères, C., Parent, E. and Parent, G. J. 2023. Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional library and a public repository. – Metabarcoding Metagenomics 7: e98539.

Callaghan, C. T., Santini, L., Spake, R. and Bowler, D. E. 2024. Population abundance estimates in conservation and biodiversity research. – Trends Ecol. Evol. 39: 515–523.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. 2016. DADA2: high resolution sample inference from Illumina amplicon data. – Nat. Methods 13: 581–583.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. 2009. Blast+: architecture and applications. – BMC Bioinform. 10: 421.

Chapuis, M.-P., Benoit, L. and Galan, M. 2023. Evaluation of 96-well high-throughput DNA extraction methods for 16S rRNA gene metabarcoding. – Mol. Ecol. Resour. 23: 1509–1525.

Chua, P. Y. S., Bourlat, S. J., Ferguson, C., Korlevic, P., Zhao, L., Ekrem, T., Meier, R. and Lawniczak, M. K. N. 2023. Future of DNA-based insect monitoring. – Trends Genet. 39: 531–544.

Creedy, T. J., Andújar, C., Meramveliotakis, E., Noguerales, V., Overcast, I., Papadopoulou, A., Morlon, H., Vogler, A. P., Emerson, B. C. and Arribas, P. 2022. Coming of age for COI metabarcoding of whole organism community DNA: towards bioinformatic harmonisation. – Mol. Ecol. Resour. 22: 847–861.

Dornelas, M., Madin, E. M. P., Bunce, M., DiBattista, J. D., Johnson, M., Madin, J. S., Magurran, A. E., McGill, B. J., Pettorelli, N., Pizarro, O., Williams, S. B., Winter, M. and Bates, A. E. 2019. Towards a macroscope: leveraging technology to transform the breadth, scale and resolution of macroecological data. – Global Ecol. Biogeogr. 28: 1937–1948.

Drake, L. E., Cuff, J. P., Young, R. E., Marchbank, A., Chadwick, E. A. and Symondson, W. O. C. 2022. An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data. – Methods Ecol. Evol. 13: 694–710.

Duke, E. M. and Burton, R. S. 2020. Efficacy of metabarcoding for identification of fish eggs evaluated with mock communities. – Ecol. Evol. 10: 3463–3476.

Elbrecht, V. and 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: e0130324.

Elbrecht, V. and Leese, F. 2017. Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. – Front. Environ. Sci. 5: 11.

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N. and Steinke, D. 2019. Validation of COI metabarcoding primers for terrestrial arthropods. – PeerJ 7: e7745.

Engel, M. S. et al. 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. – Zool. J. Linn. Soc. 193: 381–387.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., Erséus, C., Gusarov, V. I., Edwards, M. E., Johnsen, A., Stenøien, H. K., Hassel, K., Kauserud, H., Yoccoz, N. G., Bråthen, K. A., Willerslev, E., Taberlet, P., Coissac, E. and Brochmann, C. 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. – Mol. Ecol. 21: 1821–1833.

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S. and Pascal, G. 2018. FROGS: find, rapidly, OTUs with galaxy solution. – Bioinformatics 34: 1287–1294.

Fujisawa, T., Noguerales, V., Meramveliotakis, E., Papadopoulou, A. and Vogler, A. P. 2023. Image-based taxonomic classification of bulk insect biodiversity samples using deep learning and domain adaptation. – Syst. Entomol. 48: 387–401.

Galan, M., Razzauti, M., Bard, E., Bernard, M., Brouat, C., Charbonnel, N., Dehne-Garcia, A., Loiseau, A., Tatard, C., Tamisier, L., Vayssier-Taussat, M., Vignes, H. and Cosson, J.-F. 2016. 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. – mSystems 1: e00032-16.

Galan, M., Pons, J.-B., Tournayre, O., Pierre, É., Leuchtmann, M., Pontier, D. and Charbonnel, N. 2018. Metabarcoding for the parallel identification of several hundred predators and their prey: application to bat species diet analysis. – Mol. Ecol. Resour. 18: 474–489.

Goulson, D. 2019. The insect apocalypse, and why it matters. – Curr. Biol. 29: R967–R971.

Hawthorne, B. S. J., Cuff, J. P., Collins, L. E. and Evans, D. M. 2023. Metabarcoding advances agricultural invertebrate biomonitoring by enhancing resolution, increasing throughput, and facilitating network inference [Preprint]. – Open Science Framework.

Hebert, P. D. N., Cywinska, A., Ball, S. L. and deWaard, J. R. 2003. Biological identifications through DNA barcodes. – Proc. R. Soc. B 270: 313–321.

Hebert, P. D. N., Bock, D. G. and Prosser, S. W. J. 2023. Interrogating 1000 insect genomes for NUMTs: a risk assessment for estimates of species richness. – PLoS One 18: e0286620

Hoagland, K. E. 1996 The taxonomic impediment and the convention on biodiversity. – Assoc. Syst. Biol. 24: 61–67

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M. and Ladle, R. J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. – Annu. Rev. Ecol. Evol. Syst. 46: 523–549.

Huang, W., Xie, X., Huo, L., Liang, X., Wang, X. and Chen, X. 2020. An integrative DNA barcoding framework of ladybird beetles (Coleoptera: Coccinellidae). – Sci. Rep. 10: art. 1.

Keck, F., Blackman, R. C., Bossart, R., Brantschen, J., Couton, M., Hürlemann, S., Kirschner, D., Locher, N., Zhang, H. and Altermatt, F. 2022. Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. – Mol. Ecol. 31: 1820–1835.

Keck, F., Couton, M. and Altermatt, F. 2023. Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. – Mol. Ecol. Resour. 23: 742–755.

Kocher, A., Gantier, J.-C., Gaborit, P., Zinger, L., Holota, H., Valière, S., Dusfour, I., Girod, R., Bañuls, A.-L., Murienne, J. and Murienne, J. 2017. Vector soup: high-throughput identification of Neotropical phlebotomine sand flies using metabarcoding. – Mol. Ecol. Resour. 17: 172–182.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. and Schloss, P. D. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. – Appl. Environ. Microbiol. 79: 5112–5120.

Krell, F.-T. 2004. Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of 'morphospecies' sorting. – Biodivers. Conserv. 13: 795–812.

Lacasella, F., Marta, S., Singh, A., Stack Whitney, K., Hamilton, K., Townsend, P., Kucharik, C. J., Meehan, T. D. and Gratton, C. 2017. From pest data to abundance-based risk maps combining eco-physiological knowledge, weather, and habitat variability. – Ecol. Appl. 27: 575–588.

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G. and Taylor, M. I. 2019. How quantitative is metabarcoding: a meta-analytical approach. – Mol. Ecol. 28: 420–430.

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J. and Burridge, C. P. 2020. A practical guide to DNA metabarcoding for entomological ecologists. – Ecol. Entomol. 45: 373–385.

Luo, M., Ji, Y., Warton, D. and Yu, D. W. 2023. Extracting abundance information from DNA-based data. – Mol. Ecol. Resour. 23: 174–189.

Lyet, A., Pellissier, L., Valentini, A., Dejean, T., Hehmeyer, A. and Naidoo, R. 2021. eDNA sampled from stream networks correlates with camera trap detection rates of terrestrial mammals. – Sci. Rep. 11: 11362.

Ma, Z., Ren, J. and Zhang, R. 2022. Identifying the genetic distance threshold for Entiminae (Coleoptera: Curculionidae) species delimitation via COI barcodes. – Insects 13: 261.

Magoč, T. and Salzberg, S. L. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. – Bioinformatics 27: 2957–2963.

Mahé, F., Rognes, T., Quince, C., de Vargas, C. and Dunthorn, M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. – PeerJ 2: e593.

Marshall, E. and Moonen, A. 2002. Field margins in northern Europe : their functions and interactions with agriculture. – Agric. Ecosyst. Environ. 89: 5–21.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. – EMBnet. J. 17: art. 1.

Mata, V. A., Ferreira, S., Campos, R. M., da Silva, L. P., Veríssimo, J., Corley, M. F. V. and Beja, P. 2021. Efficient assessment of nocturnal flying insect communities by combining automatic light traps and DNA metabarcoding. – Environ. DNA 3: 398–408.

McGill, B. J., Dornelas, M. and Field, R. 2016. A new year with a new leadership team at GEB – or how to guarantee your paper gets into GEB. – Global Ecol. Biogeogr. 25: 1–2.

Meier, R., Zhang, G. and Ali, F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. – Syst. Biol. 57: 809–813.

Mioduchowska, M., Czyż, M. J., Gołdyn, B., Kur, J. and Sell, J. 2018. Instances of erroneous DNA barcoding of metazoan invertebrates: are universal cox1 gene primers too "universal"? – PLoS One 13: e0199609.

Mirimin, L., Desmet, S., Romero, D. L., Fernandez, S. F., Miller, D. L., Mynott, S., Brincau, A. G., Stefanni, S., Berry, A., Gaughan, P. and Aguzzi, J. 2021. Don't catch me if you can – using cabled observatories as multidisciplinary platforms for marine fish community monitoring: an in situ case study combining underwater video and environmental DNA data. – Sci. Total Environ. 773: 145351.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S. and Köster, J. 2021. Sustainable data analysis with Snakemake. – F1000Research 10: 33.

Noguerales, V., Meramveliotakis, E., Castro-Insua, A., Andújar, C., Arribas, P., Creedy, T. J., Overcast, I., Morlon, H., Emerson, B. C., Vogler, A. P. and Papadopoulou, A. 2023. Community metabarcoding reveals the relative role of environmental filtering and spatial processes in metacommunity dynamics of soil microarthropods across a mosaic of montane forests. – Mol. Ecol. 32: 6110–6128.

Oliver, I. and Beattie, A. J. 1993. A possible method for the rapid assessment of biodiversity. – Conserv. Biol. 7: 562–568.

Penel, B., Meynard, C. N., Benoit, L., Bourdonné, A., Clamens, A.-L., Soldati, L., Migeon, A., Chapuis, M. P., Piry, S., Kergoat, G. J. and Haran, J. 2024. Data from: The best of two worlds: toward large-scale monitoring of biodiversity combining COI metabarcoding and optimized parataxonomic validation. – Zenodo, https://zenodo.org/records/13760159.

Pereira, C. L., Gilbert, M. T. P., Araújo, M. B. and Matias, M. G. 2021. Fine-tuning biodiversity assessments: a framework to pair eDNA metabarcoding and morphological approaches. – Methods Ecol. Evol. 12: 2397–2409.

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C. and Blacket, M. J. 2019. Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. – GigaScience 8: giz092.

Ramankutty, N., Evan, A. T., Monfreda, C. and Foley, J. A. 2008. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. – Global Biogeochem. Cycles 22: GB1003.

Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. 2016. VSEARCH: a versatile open source tool for metagenomics. – PeerJ 4: e2584.

Schultz, J. A. and Hebert, P. D. N. 2022. Do pseudogenes pose a problem for metabarcoding marine animal communities? – Mol. Ecol. Resour. 22: 2897–2914.

Song, H., Buhay, J. E., Whiting, M. F. and Crandall, K. A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. – Proc. Natl Acad. Sci. USA 105: 13486–13491.

Sow, A., Brévault, T., Benoit, L., Chapuis, M.-P., Galan, M., Coeur d'acier, A., Delvare, G., Sembène, M. and Haran, J. 2019. Deciphering host–parasitoid interactions and parasitism rates of crop pests using DNA metabarcoding. – Sci. Rep. 9: 3646.

Taberlet, P., Bonin, A., Zinger, L. and Coissac, E. 2018. Environmental DNA: for biodiversity research and monitoring. – Oxford Univ. Press.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. and Legendre, F. 2017. Taxonomic bias in biodiversity data and societal preferences. – Sci. Rep. 7: 9132.

van Klink, R., Bowler, D. E., Gongalsky, K. B. and Chase, J. M. 2022. Long-term abundance trends of insect taxa are only weakly correlated. – Biol. Lett. 18: 20210554.

Van Rossum, G. 1995. Python reference manual (R 9525). – Art. R 9525.

Veillat, L., Boyer, S., Querejeta Coma, M., Magnoux, E., Roques, A., Lopez-Vaamonde, C. and Roux, G. 2023. Molecular biosurveillance of wood-boring cerambycid beetles using DNA metabarcoding. – ARPHA Preprints.

Wagner, D. L., Grames, E. M., Forister, M. L., Berenbaum, M. R. and Stopak, D. 2021. Insect decline in the Anthropocene: death by a thousand cuts. – Proc. Natl Acad. Sci. USA 118: e2023989118.

Wiemers, M. and Fiedler, K. 2007. Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). – Front. Zool. 4: 8.

Winfree, R., Fox, J. W., Williams, N. M., Reilly, J. R. and Cariveau, D. P. 2015. Abundance of common species, not species richness, drives delivery of a real-world ecosystem service. – Ecol. Lett. 18: 626–635.

Wizenberg, S. B., Newburn, L. R., Pepinelli, M., Conflitti, I. M., Richardson, R. T., Hoover, S. E. R., Currie, R. W., Giovenazzo, P. and Zayed, A. 2023. Validating a multi-locus metabarcoding approach for characterizing mixed-pollen samples. – Plant Methods 19: 120.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. – J. R. Stat. Soc. B 73: 3–36.

Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M., Schmidt, S., Cerretti, P. and Meier, R. 2022. DiversityScanner: robotic handling of small invertebrates with machine learning methods. – Mol. Ecol. Resour. 22: 1626–1638.