1   **Resource Article**

2   **Haplotype-level metabarcoding of freshwater macroinvertebrate species: a**
3   **prospective tool for population genetic analysis**

4   *Running Title: DNA metabarcoding-based haplotyping*

5   **Authors and Affiliations**

6   Joeselle M. Serrana and Kozo Watanabe

7   Center for Marine Environmental Studies (CMES), Ehime University, Bunkyo-cho 3, Matsuyama, Ehime 790-8577, Japan

8   **Correspondence**

9   Prof. Kozo Watanabe, Ph.D.; Address: Molecular Ecology and Health (MEcoH) Laboratory, Engineering Building No. 2, Ehime
10   University, Bunkyo-cho 3, Matsuyama, Ehime 790-8577, Japan; Email: watanabe.kozo.mj@ehime-u.ac.jp

11 ABSTRACT

12 The development and evaluation of DNA metabarcoding protocols for haplotype-level resolution
 13 require attention, specifically for population genetic analysis, i.e., parallel estimation of genetic
 14 diversity and dispersal patterns among multiple species present in a bulk sample. Further exploration
 15 and assessment of the laboratory and bioinformatics strategies are warranted to unlock the potential of
 16 metabarcoding-inferred population genetic analysis. Here, we assessed the inference of freshwater
 17 macroinvertebrate haplotypes from DNA metabarcoding data using mock samples with known Sanger-
 18 sequenced haplotypes. We also examined the influence of different DNA template concentrations and
 19 PCR cycles on detecting true haplotypes and the reduction of spurious haplotypes obtained from DNA
 20 metabarcoding. We tested our haplotyping strategy on a mock sample containing 20 specimens from
 21 four species with known haplotypes based on the 658-bp Folmer region of the mitochondrial
 22 cytochrome c oxidase gene. The read processing and denoising step resulted in 14 zero-radius
 23 operational taxonomic units (ZOTUs) of 421-bp length, with 12 ZOTUs having 100% match with 12 of
 24 the Sanger haplotype sequences. Quality passing reads relatively increased with increasing PCR
 25 cycles, and the relative abundance of each ZOTUs was consistent for each cycle number. This
 26 suggests that increasing the cycle number from 24 to 64 did not affect the relative abundance of
 27 quality passing filter reads of each ZOTUs. Our study demonstrated the ability of DNA metabarcoding
 28 to infer intraspecific variability while highlighting the challenges that need to be addressed before its
 29 possible applications to population genetic studies.

30 **Keywords:** DNA metabarcoding, freshwater macroinvertebrates, haplotype, intraspecific diversity,
 31 population genetics, zero-radius operational taxonomic units

32 INTRODUCTION

33 High throughput sequencing (HTS) of amplified markers, i.e., DNA metabarcoding, enables the
34 simultaneous multi-species identification of community samples containing DNA from different origins
35 (Taberlet et al., 2012). The advancements in sequencing platform technologies made it possible to
36 generate millions of sequences from bulk (i.e., mixed, mass-collected, or community) samples and
37 potentially identify most taxa, including the rare and inconspicuous ones (Creer et al., 2016). The
38 number of sequences varies depending on the specificity of the barcoding gene and the species
39 diversity present in the environment (Yu et al., 2012; Zinger et al., 2019).

40 The development and application of DNA metabarcoding for biodiversity surveys have been
41 considered a game-changer for ecological research (Creer et al., 2016). The ability of the method to
42 identify freshwater macroinvertebrates better than morphological identification has been demonstrated
43 in previous literature (e.g., Hajibabaei et al., 2012; Serrana et al., 2019). However, other than the
44 characterization of the communities' taxonomic composition, DNA metabarcoding data can be used to
45 infer biodiversity indices and is by now established as a robust method for environmental impact
46 assessments of freshwater ecosystems (e.g., Aylagas et al., 2016; Deiner et al., 2017; Pawlowski et
47 al., 2018; Serrana et al., 2018). For these applications, biodiversity is generally quantified with
48 taxonomic, functional, or phylogenetic diversity at the community level (interspecific diversity) (Gotelli
49 et al., 2017). However, intraspecific diversity or the genetic variation within species from community
50 samples is not typically explored.

51 The inclusion of intraspecific diversity assessment would be beneficial in ecological monitoring
52 and management strategies given that haplotype data are direct proxies for the spatio-temporal
53 dynamics of populations and can substantially differ compared to community-level assessments
54 (Taberlet et al., 2012). In particular, the assessment of changes in population size in response to
55 environmental stressors (e.g., Monaghan et al., 2005; Weiss & Leese, 2016) are key areas of basic
56 and applied ecological research (Sutherland et al., 2013). Quantifying gene flow between populations
57 can examine the magnitude and mechanisms of population connectivity.

58 Previous literature has proposed using DNA metabarcoding for population genetic analysis
59 (e.g., Bohmann et al., 2014; Adams et al., 2019; Arribas et al., 2020; Shum & Palumbi, 2021). Recent
60 studies in freshwater environments were able to infer haplotypes from DNA metabarcoding of
61 individually-tagged specimens (Shokralla et al., 2014), bulk macroinvertebrate samples (Elbrecht &
62 Leese, 2015; Pedro et al., 2017; Zizka, Weiss & Leese, 2020), and from environmental DNA collected
63 from water samples (Stat et al., 2017; Alam et al., 2020; Tsuji et al., 2020a; Tsuji et al., 2020b; Dugal
64 et al., 2021), all of which highlighted the possibility of extracting haplotype or sequence variant
65 information from HTS marker-gene sequencing data. DNA metabarcoding can be a faster and cost-
66 effective alternative to the conventional approach that repeats genetic analysis for multiple individuals
67 from different populations. Metabarcoding datasets may allow the comparison of population genetic
68 structures of multiple species with different life stages and dispersal modes or abilities (Shum &
69 Palumbi, 2021).

70 However, the application of DNA metabarcoding to population genetic analysis are highly
71 influenced by the high number of reads containing sequencing errors, which could have occurred at
72 different steps in the metabarcoding procedure, from polymerase chain reaction (PCR) amplification
73 bias and errors to bioinformatics pipelines (Elbrecht et al., 2018; Turon et al., 2019). DNA template
74 concentration and the number of PCR cycles introduce PCR bias and errors, mainly when applied to
75 bulk samples (Pawluczyk et al., 2015; Dopheide et al., 2017; Collins et al., 2019). This makes it
76 difficult to distinguish true haplotypes from erroneous sequencing noise (Turon et al., 2019). Thus,
77 further exploration and assessment of the laboratory and bioinformatics strategies are required to
78 unlock the potential of metabarcoding-based inference of haplotype information.

79 Here, we inferred haplotype information of freshwater macroinvertebrate species from DNA
80 metabarcoding data. Using a mixed-species mock sample with known Sanger-sequenced haplotypes
81 based on the 658-bp barcode region of the mitochondrial cytochrome oxidase I (mtCOI) gene, we
82 tested whether the same COI haplotypes will be identified from the DNA metabarcoding data. We also
83 examined the influence of varying DNA template concentration and PCR cycles on detecting true
84 haplotypes and reducing spurious haplotypes obtained from DNA metabarcoding using the denoising
85 parameters of unoise3 (Edgar, 2016). As DNA metabarcoding becomes more established and

laboratory protocols and bioinformatics pipelines are continuously being developed, this study aimed to demonstrate that the method could be used to infer intraspecific genetic variability, showing promise for possible applications in population genetic analysis.

MATERIALS AND METHODS

Study species and Sanger-sequence haplotypes

We tested our haplotyping strategy on a mock sample by pooling extracted DNA of the 20 specimens from four species with known haplotypes assessed from published population genetics studies (*Amphinemura decemseta*; Gamboa et al., 2019) or our current DNA barcoding projects (*Kamimuria tibialis*, *Eucapnopsis bulba*, and *Epeorus latifolium*) (**Table 1**). Twenty haplotypes were identified based on the Sanger sequence of the 658-bp Folmer region of the mitochondrial cytochrome c oxidase I (mtCOI) gene using DnaSP 6.0 (Rozas et al., 2017).

The genealogical relationship of the haplotypes of each species via median-joining network was determined using a haplotype network tree (**Figure 1a**) drawn using the PopART software (Full-Feature Software for Haplotype Network Construction) (Leigh & Bryant, 2015). To examine the influence of varying input DNA template concentration on the detection of true haplotypes from the mock sample, a concentration gradient from the extracted DNA of the haplotypes from each species was prepared, i.e., 0.01, 0.05, 0.1, 0.5, and 1.0 ng/μL referred to as haplotypes 1 to 5, respectively (**Table 1**). Additional two samples with different concentrations of 5.0 and 10.0 ng/μL for *Kamimuria tibialis* were included, referred to as haplotype 6 and 7 (**Table 1**). DNA concentration was quantified with the QuantiFluor dsDNA system (Promega, Madison, WI, USA) on the Quantus Fluorometer (Promega, Madison, WI, USA).

Library preparation and next-generation sequencing

Fifteen μl each of the 20 haplotype DNA extracts adjusted to their respective concentrations were pooled as one mock sample. Elbrecht and Steinke's (2019) fusion primers (BF2 and BR2; **Figure 1b**) for freshwater arthropods were used to construct amplicon libraries following a one-step polymerase

chain reaction (PCR) protocol. The PCR master mix consists of 0.25 µl Phusion, 0.75 µl DMSO, one µl dNTPs, 1.25 µl each of the fusion forward and reverse primers (10 µM), five µl HF Buffer (New England Biolabs), and 15.5 µl of PCR-grade water. PCR cycling conditions were 30 seconds of initial denaturation at 98°C, followed by 20 cycles of 10 seconds denaturation at 98°C, 30 seconds annealing at 55°C, 30 seconds extension at 72 °C, and a final extension step of 5 minutes at 72°C. The same PCR protocol was done for eleven other cycles from 24 to 64 (increments of 4) with triplicate PCR amplification, including negative controls for each PCR cycle (also in triplicates).

The resulting 72 PCR amplicons were pooled according to replicates and size selected via solid-phase reversible immobilization (SPRI) beads. Each replicate was quantified via qPCR with the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA). Before sequencing, quality was assessed via the DNA 1000 assay using the Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA, USA). The replicate amplicon libraries were normalized to 2nM before pooling to ensure even read output distribution between replicates. Paired-end sequencing of the 2nM pooled library spiked with 20% PhiX was performed in paired-end 300 cycles sequencing on the Illumina MiSeq system.

Sequence processing and haplotype inference

The raw Illumina MiSeq paired-end reads were demultiplexed according to sample tags via the R package JAMP v.0.67 (<http://github.com/VascoElbrecht/JAMP>) (Elbrecht et al., 2018) and were quality-checked with FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). We demultiplexed 14.7 million reads assigned to each of the 72 samples. The paired-end reads were then merged via the JAMP pipeline, primer stripped, truncated at 421-bp, quality filtered with a maximum expected error filtering value of 2, and dereplicated excluding singletons using the USEARCH v11.0.667 pipeline (Edgar, 2013). To extract individual haplotypes from the dereplicated sequences, we employed a denoising strategy using unoise3 (Edgar, 2016). The zero-radius operational taxonomic unit (ZOTUs) sequences were mapped against the 658-bp mtCOI Sanger sequences of the mock samples using the UPARSE-REF algorithm (Edgar, 2013). The command is designed for validating mock community sequencing experiments where the set of biological sequences in the

sample is known. Note that amplicon sequence variants (ASVs), exact sequence variants (ESVs), sub-OTUs, OTUs with 100% sequence similarity, and ZOTUs are synonymous (Porter & Hajibabaei, 2018), but we used the latter term in this study since we employed unoise3's algorithm (Edgar, 2016). We also performed phylogenetic inference via neighbor-joining analysis of the Sanger and the ZOTU sequences employing the Jukes-Cantor substitution model with bootstrapping (1000) via the online MAFFT multiple sequence alignment software version 7 (https://mafft.cbrc.jp/alignment/server/large.html) (Katoh, Rozewicki & Yamada, 2019). Prior to the analysis, the 658-bp Sanger sequences of the mock samples were truncated to the 421-bp length of the BF2 and BR2 barcode.

Data visualization (i.e., boxplots and bubble plot) and statistical analysis were performed in R v.4.0.2 (R Core Team, 2019). The default Wilcoxon T-test analysis was performed to compare the read abundances of the raw, merged, and ZOTU-assigned sequences between the DNA template and negative controls for each cycle using the function `stat_compare_means()`.

RESULTS

Sanger sequence haplotypes

The mock sample consisted of four different species, i.e., *Amphinemura decemseta* (Nemouridae), *Kamimuria tibialis* (Perlidae), *Eucapnopsis bulba* (Capniidae), and *Epeorus latifolium* (Heptageniidae) from two freshwater insect orders (Table 1). However, upon trimming the Sanger sequences into the 421-bp length of the BF2 and BR2 barcode region to complement with the amplified region in the DNA metabarcoding data, three of the *K. tibialis* Sanger haplotypes (i.e., KT1, KT3, and KT4) were grouped into one, similar to the two haplotypes of *A. decemseta* (i.e., AD1 and AD3). Hence, the 20 haplotypes from the 658-bp long fragment of the mtCOI barcoding region (Folmer et al., 1994) were reduced to 17 haplotypes after trimming to 421-bp length.

Read abundance, reference sequence match, and phylogenetic inference

Except for the raw sequence and paired-end merging step at cycle 20, the read abundances of the DNA template samples and the negative controls were significantly different (**Figure 2a**). From the dereplicated reads, 462,665 were assigned into 14 ZOTUs (also referred to as "DNA metabarcoding haplotypes" in this study) (**Table S1**) of 421-bp length. For assessing the different PCR cycle numbers, all of the ZOTUs were detected from cycles 24 to 64. For cycle 20, only 8 of the 14 ZOTUs were represented, all of which match haplotypes with high concentrations (i.e., KT7, KT6, KT5, EL5, EL4, EB5, EB4, and EB3) (**Figure 2b**). Quality passing reads relatively increased with increasing cycle numbers, and the relative abundance of each ZOTUs was relatively consistent across the cycle numbers (**Table S1**). Notably, four ZOTUs were detected in the negative samples from different cycles. ZOTU04 was present on nine negative samples from cycles 28 to 64, ZOTU09 and ZOTU11 on two negative samples, and ZOTU08 and ZOTU10 on one, i.e., cycle 60 and 64, respectively. However, it should be noted that most of these occurrences in the negative samples were singleton or doubleton reads, with one sample (i.e., 40BR1) having the highest detection of only 22 reads.

After mapping the ZOTU sequences against the 20 Sanger sequence-haplotypes with 658-bp, 12 ZOTUs comprised of 450,793 (97%) matched-to-ZOTU reads had 100% sequence match against 12 of the Sanger haplotypes of the mock reference sequence (**Table S2**). The remaining eight Sanger haplotypes that were not detected from the DNA metabarcoding dataset were all the *A. decemseta* samples (AB1: 0.01, AB2: 0.05, and AB3: 0.10 ng/μL), two *E. bulba* (EB1: 0.01, and EB2: 0.05 ng/μL), *E. latifolium* (EL1: 0.01 ng/μL), and two *K. tibialis* (KT1:0.01, and KT3: 0.10 ng/μL). Based on the neighbor-joining tree (**Figure 3**), the remaining two ZOTUs (i.e., ZOTU10 and ZOTU14) without a 100% taxonomic match against the Sanger sequences clustered with the *A. decemseta* sequences.

DISCUSSION

Using a multi-species mock sample with known haplotypes based on Sanger sequencing of the highly variable mtCOI gene region, we presented the feasibility and limitations of using DNA metabarcoding data to extract intraspecific genetic diversity information by denoising the sequences into ZOTUs. Most recent studies inferred intraspecific diversity from macroinvertebrate DNA metabarcoding data were

also based on denoising algorithms (e.g., Elbrecht et al., 2018; Turon et al., 2019; Brandt et al., 2021). Before denoising, our read processing steps, i.e., paired-end merging, stripping adapter and primer sequences, barcode length truncation, quality filtering, and dereplication, were performed mainly following default and stringent settings. It is also important to note that we truncated the sequence reads to the entire length of the BF2 and BR2 barcode (i.e., 421-bp) to prevent the generation of shorter-than-barcode ZOTUs. We also discarded singletons, i.e., ZOTUs containing only one sequence read.

The four species selected in the mock sample have known haplotypes assessed from a published population genetics study (*Amphinemura decemseta*; Gamboa et al., 2019) or our current DNA barcoding projects through the sequencing of the 658-bp region of the mtCOI gene. Upon truncating the mock sequences to the 421-bp length of the BF2 and BR2 barcode, two species, i.e., *A. decemseta* and *K. tibialis* were reduced from 3 to 2 and 7 to 5 haplotypes, respectively. This means that some of the polymorphic sites that distinguished the haplotypes from each other were discarded from shortening the 658-bp barcoding region to 421-bp. Twelve of the 14 ZOTUs inferred from the DNA metabarcoding sequences had perfect sequence matches against 12 of the 20 (658-bp) or 17 (421-bp) Sanger haplotypes. The *K. tibialis* haplotype with a relatively high input DNA template concentration (i.e., 0.10 ng/μL) was likely not a false negative detection. Its reads were present in the metabarcoding data but merged with the other *K. tibialis* haplotype.

We report that initial DNA template concentration significantly influenced the detection of individuals from a mock community sample. This observation is in accordance with previous studies which reported that samples or taxa with low DNA template concentrations had lower detection probability (Martins et al., 2020). Accordingly, abundant taxa or samples with high biomass tend to have higher detection probabilities than those rare, smaller, or have low biomass from mixed-community samples (Carew, Coleman & Hoffmann, 2018; Erdozain et al., 2019; Serrana et al., 2019). The difference in biomass affects haplotypes' detection since most large specimens would be retained after read processing. These factors need to be addressed when metabarcoding-based haplotyping is used to infer abundance-based analysis for population genetics applications.

On the other hand, all of the ZOTUs were detected from cycles 24 to 64. For cycle 20, only 8 of the 14 ZOTUs were represented, all of which match haplotypes with high concentrations. Quality passing reads relatively increased with increasing cycle numbers, and the relative abundance of each ZOTUs was consistent for each cycle number. This suggests that increasing the PCR cycle from 24 to 64 did not affect the relative abundance of quality passing filter reads of each ZOTUs. Our findings align with previous studies that reported small or no effect of PCR cycle number on amplification bias (Sipos et al., 2007; Krehenwinkel et al., 2017). This contrasts with other reports that increasing PCR cycles reduces the proportion of sequences with low starting DNA or less well-amplified species in the mock sample (Piñol et al., 2015).

Moreover, higher PCR cycles have been reported to increase the formation of chimeric sequences and amplification bias (Aylagas et al., 2016); that is why some metabarcoding protocols discourage increasing the PCR cycles above 30. However, a literature survey on bulk metabarcoding studies presented that 73% of reports used more than 30 PCR cycles to circumvent primer annealing issues or to amplify samples with low amounts of DNA (van der Loos & Nijland, 2020). Our findings proved otherwise wherein low template samples were undetected or not amplified even after increasing the cycle number to 64. The relative abundances of the samples detected were relatively consistent for each cycle. However, we note that the sequence diversity in our mock sample was relatively low (i.e., four species with 20 haplotypes). A more diverse community might present a different pattern in PCR cycle effects, hence, warrants further evaluation.

The proper selection of the DNA marker in DNA metabarcoding assay is crucial because all of the downstream analyses, e.g., species detection and identification, rely on the marker's ability to amplify and discriminate the representative taxa of the target organisms (Piper et al., 2019). The mtCOI gene has a relatively high mutation rate and is capable of detecting intraspecific variation. Thus, its widespread use for population genetics studies (e.g., Hajibabaei et al., 2007; Curry et al., 2018) and has been widely used for DNA metabarcoding macroinvertebrates to date. With this, intraspecific variation can be extracted from community-based samples using various algorithms for sequence clustering and phylogenetic rates (Giusti et al., 2017; Vamos, Elbrecht & Leese, 2017;

Elbrecht et al., 2019). However, many HTS platforms to date still have strict limitations in sequence length, not utilizing the full-length of the mtCOI barcode (Piper et al., 2019).

As we have observed in this study, some species lost nucleotide information to distinguish the haplotypes from each other due to shortening the barcode region. Hence, the inability of shorter marker regions to differentiate haplotypes of certain species requires further development and evaluation. On the other hand, the advancements in sequencing technologies to generate longer sequences could allow the full-length generation of longer DNA barcodes from community samples (Baloğlu et al., 2021), resolving the limitations of short amplicon sequencing in DNA metabarcoding. Longer reads would allow the generation of high-resolution mitochondrial haplotype data, with potential applications for demographic history and selection analyses (Sigsgaard et al., 2019). However, given the high raw read error rate from these long-read sequencing platforms that range from 10-22% (i.e., the Oxford Nanopore MinION™), further assessment and exploration of library preparation and error-correction methods are recommended (Baloğlu et al., 2021) for its utilization for DNA metabarcoding studies, more so for haplotype-level inference.

The two DNA metabarcoding haplotypes (i.e., ZOTU10 and ZOTU14) that failed to have a perfect sequence match against the Sanger sequences clustered with the *A. decemseta* samples. Although we could identify them as *A. decemseta* sequences based on the phylogenetic approach, metabarcoding failed to detect a 100% match of the Sanger haplotypes of this species from the mock sample. We could not rule out PCR amplification, primer bias, or sequencing errors as the reason for these false positive detections or spurious haplotypes obtained from the metabarcoding data, even if we performed relatively strict read quality filtering parameters. Elbrecht et al. (2018) reported that artificial or false haplotypes could not be fully excluded even with stringent filtering settings due to undetected chimeric sequences or systematic sequencing errors that might persist across replicates.

Additionally, nine blank samples had sequences for ZOTU04, ZOTU08, ZOTU09, ZOTU11, and ZOTU10. Although most of these occurrences were singleton or doubleton reads, the presence of these reads in the negative controls might be due to tag-jumps or the amplification of sequences carrying false combinations of used nucleotide tags that is common for dual indexed libraries (Esling, Lejzerowicz & Pawlowski, 2015; Schnell, Bohmann & Gilbert, 2015; Zizka et al., 2019). Hence, the

challenges with tag jumping and contamination between libraries require attention to alleviate false read-to-sample assignments that would be problematic once the method is employed to environmental samples.

CONCLUSION

We demonstrated that haplotype information could be extracted from mixed community samples of benthic macroinvertebrates. We recovered 14 zero-radius operational taxonomic units (ZOTUs) of 421-bp length, with 12 ZOTUs having 100% match with 12 of the mock haplotype sequences. Moreover, quality passing reads relatively increased with increasing PCR cycle numbers. The relative abundance of each ZOTUs was consistent across the cycle numbers suggesting that increasing the cycles did not affect the relative abundance of quality passing filter reads. As DNA metabarcoding becomes more established and laboratory protocols and bioinformatics pipelines are continuously being developed, our study demonstrated that the method could be used to infer intraspecific variability, showing promise for possible applications and highlighting the challenges that need to be addressed before haplotype-level metabarcoding can entirely be used for population genetic studies.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (Grant No. 19K21996 and 19H02276). We thank Dr. Dávid Murányi, Dr. Maribet Gamboa and Dr. Sakiko Yaegashi for the EPT samples and their molecular data. We also thank Dr. Naohito Tokunaga of the Division of Analytical Bio-Medicine for his assistance in performing high-throughput sequencing on the Illumina MiSeq platform of the Advanced Research Support Center (ADRES), Ehime University.

Author contributions

K.W. and J.M.S. conceptualized and designed the study; J.M.S. performed laboratory work and bioinformatics analyses. J.M.S. and K.W. interpreted the data and wrote the manuscript.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data availability statement

The raw sequence data were deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRR14631855). The Sanger (GenBank accession numbers also listed in Table 1) and ZOTU sequences (Fasta format), other input data, and the R markdown document implementing the analyses contained in this manuscript are available in the figshare data repository at <https://doi.org/10.6084/m9.figshare.15090180.v1> (Serrana & Watanabe, 2021).

ORCID

Joeselle M. Serrana: <https://orcid.org/0000-0002-6967-5407>; Kozo Watanabe: <https://orcid.org/0000-0002-7062-595X>

References

- Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G. J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: can environmental DNA (eDNA) cut it as a population genetics tool? *Genes*, 10(3), 192.
- Alam, M. J., Kim, N. K., Andriyono, S., Choi, H. K., Lee, J. H., & Kim, H. W. (2020). Assessment of fish biodiversity in four Korean rivers using environmental DNA metabarcoding. *PeerJ*, 8, e9508.
- Arribas, P., Andújar, C., Salces-Castellano, A., Emerson, B. C., & Vogler, A. P. (2021). The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. *Molecular Ecology*, 30(1), 48-61.
- Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Frontiers in Marine Science*, 3, 96.
- Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2021). A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods in Ecology and Evolution*, 12(5), 794-804.
- 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.
- Bohmann, K., Gopalakrishnan, S., Nielsen, M., Nielsen, L. D. S. B., Jones, G., Streicker, D. G., & Gilbert, M. T. P. (2018). Using DNA metabarcoding for simultaneous inference of common vampire bat diet and population structure. *Molecular Ecology Resources*, 18(5), 1050-1063.
- Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, <https://doi.org/10.1111/1755-0998.13398>.
- Carew, M. E., Coleman, R. A., & Hoffmann, A. A. (2018). Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? *PeerJ*, 6, e4980.

Collins, R. A., Bakker, J., Wangenstein, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., ... & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, 10(11), 1985-2001.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., ... & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, 7(9), 1008-1018.

Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., & Baird, D. J. (2018). Identifying North American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? *Freshwater Science*, 37(1), 178-189.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... & Pfrender, M. E. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872-5895.

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., & Newcomb, R. D. (2019). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution*, 10(1), 120-133.

Dugal, L., Thomas, L., Reinholdt Jensen, M., Sigsgaard, E. E., Simpson, T., Jarman, S., ... & Meekan, M. (2021). Individual haplotyping of whale sharks from seawater environmental DNA. *Molecular Ecology Resources*, <https://doi.org/10.1111/1755-0998.13451>.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10(10), 996-998.

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257.

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.

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, 6, e4644.

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

Elbrecht, V., Braukmann, T. W., Ivanova, N. V., Prosser, S. W., Hajibabaei, M., Wright, M., ... & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7, e7745.

Erdozain, M., Thompson, D. G., Porter, T. M., Kidd, K. A., Kreutzweiser, D. P., Sibley, P. K., ... & Hajibabaei, M. (2019). Metabarcoding of storage ethanol vs. conventional morphometric identification in relation to the use of stream macroinvertebrates as ecological indicators in forest management. *Ecological Indicators*, 101, 173-184.

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, 43(5), 2513-2524.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology Biotechnology*, 3: 294-9.

Gamboa, M., Muranyi, D., Kanmori, S., & Watanabe, K. (2019). Molecular phylogeny and diversification timing of the Nemouridae family (Insecta, Plecoptera) in the Japanese Archipelago. *PloS one*, 14(1), e0210269.

Giusti, A., Tinacci, L., Sotelo, C. G., Marchetti, M., Guidi, A., Zheng, W., & Armani, A. (2017). Seafood identification in multi-species products: assessment of 16SrRNA, cytb, and COI Universal Primers' efficiency as a preliminary analytical step for setting up metabarcoding next-generation sequencing techniques. *Journal of Agricultural and Food Chemistry*, 65(13), 2902-2912.

Gotelli, N. J., Shimadzu, H., Dornelas, M., McGill, B., Moyes, F., & Magurran, A. E. (2017). Community-level regulation of temporal trends in biodiversity. *Science Advances*, 3(7), e1700315.

Hajibabaei, M., Singer, G. A., Hebert, P. D., & Hickey, D. A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *TRENDS in Genetics*, 23(4), 167-172.

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.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 20(4), 1160-1166.

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), 1-12.

Leigh, J. W., & Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, 6(9), 1110-1116.

Martins, F. M., Porto, M., Feio, M. J., Egeter, B., Bonin, A., Serra, S. R., ... & Beja, P. (2020). Modelling technical and biological biases in macroinvertebrate community assessment from bulk preservative using multiple metabarcoding markers. *Molecular Ecology*, DOI: <https://doi.org/10.1111/mec.15620>.

Monaghan, M. T., Robinson, C. T., Spaak, P., & Ward, J. V. (2005). Macroinvertebrate diversity in fragmented Alpine streams: implications for freshwater conservation. *Aquatic Sciences*, 67(4), 454-464.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéoz-Perret-Gentil, L., Beja, P., Boggero, A., ... & Feio, M. J. (2018). The future of biotic indices in the ecogenomic era: Integrating (e) DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, 637, 1295-1310.

Pawluczyk, M., Weiss, J., Links, M. G., Aranguren, M. E., Wilkinson, M. D., & Egea-Cortines, M. (2015). Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, 407(7), 1841-1848.

Pedro, P. M., Piper, R., Bazilli Neto, P., Cullen Jr, L., Dropa, M., Lorencao, R., ... & Turati, D. T. (2017). Metabarcoding analyses enable differentiation of both interspecific assemblages and intraspecific divergence in habitats with differing management practices. *Environmental Entomology*, 46(6), 1381-1389.

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (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), 819-830.

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

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, 27(2), 313-338.

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

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, 34(12), 3299-3302.

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated—reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, 15(6), 1289-1303.

Serrana, J. M., Yaegashi, S., Kondoh, S., Li, B., Robinson, C. T., & Watanabe, K. (2018). Ecological influence of sediment bypass tunnels on macroinvertebrates in dam-fragmented rivers by DNA metabarcoding. *Scientific Reports*, 8(1), 1-10.

Serrana, J. M., Miyake, Y., Gamboa, M., & Watanabe, K. (2019). Comparison of DNA metabarcoding and morphological identification for stream macroinvertebrate biodiversity assessment and monitoring. *Ecological Indicators*, 101, 963-972.

Serrana, J. M. & Watanabe, K. (2021). Data from: Haplotype-level metabarcoding of freshwater macroinvertebrate species: a prospective tool for population genetic analysis. *figshare*, <https://doi.org/10.6084/m9.figshare.15090180.v1>.

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, 14(5), 892-901.

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

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, 13(2), 245-262.

- Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., ... & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, 7(1), 1-11.
- Sutherland, W. J., Freckleton, R. P., Godfray, H. C. J., Beissinger, S. R., Benton, T., Cameron, D. D., ... & Hails, R. S. (2013). Identification of 100 fundamental ecological questions. *Journal of Ecology*, 101(1), 58-67.
- 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.
- Tsuji, S., Maruyama, A., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020a). Environmental DNA analysis shows high potential as a tool for estimating intraspecific genetic diversity in a wild fish population. *Molecular Ecology Resources*, 20(5), 1248-1258.
- Tsuji, S., Shibata, N., Sawada, H., & Ushio, M. (2020b). Quantitative evaluation of intraspecific genetic diversity in a natural fish population using environmental DNA analysis. *Molecular Ecology Resources*, 20(5), 1323-1332.
- van der Loos, L. M., & Nijland, R. (2020). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, <https://doi.org/10.1111/mec.15592>.
- Vamos, E., Elbrecht, V., & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625.
- Weiss, M., & Leese, F. (2016). Widely distributed and regionally isolated! Drivers of genetic structure in *Gammarus fossarum* in a human-impacted landscape. *BMC Evolutionary Biology*, 16(1), 153.
- 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.
- Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., ... & De Barba, M. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, 28(8), 1857-1862.
- Zizka, V. M., Elbrecht, V., Macher, J. N., & Leese, F. (2019). Assessing the influence of sample tagging and library preparation on DNA metabarcoding. *Molecular Ecology Resources*, 19(4), 893-899.
- Zizka, V. M. A., Weiss, M., & Leese, F. (2020). Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. *Metabarcoding and Metagenomics*, 4, e51925.

Supporting information

The supplementary data to this article is provided as Additional file 1 (DOCX).

449 **FIGURES**

450

451 **Figure 1** Median-joining haplotype networks showing the level of intraspecific variation within each of
452 the four species based on the 658-bp mtCOI Sanger sequence (a). The number of mutations is
453 represented as hatch marks or numbers. Position of the sequenced amplicon (BF2 & BR2) along the
454 mtCOI barcode region (b).

Figure 2 Comparison of read abundance (log-transformed) at different read processing steps per cycle and between the DNA template samples and the negative controls (a). Significance code: "***" associated with a variable at $p < 0.01$, and "*" at $p < 0.05$. Bubble plot showing the relative abundance of ZOTUs per PCR cycle (b). Circle size represents the Mean values of three replicates, and the gray shadow as standard deviation (SD). "-X" are samples with DNA template; "-B" are negative controls.

461

462 **Figure 3** Neighbor-joining tree of the mock and the zero-radius operational taxonomic units (ZOTU)
463 sequences. Sequences are highlighted based on species. The red bar represents a ZOTU without a
464 taxonomic match, and the text in red represents a Sanger haplotype without a ZOTU match.

TABLES

Table 1 Sample and haplotype information table.

Order	Family	Species	GenBank Accession	Sample Code	Conc. ng/uL [†]	Haplotype Information					
						658-bp [‡]			421-bp [§]		
						<i>h</i>	<i>Hd</i>	<i>Eta</i>	<i>h</i>	<i>Hd</i>	<i>Eta</i>
Plecoptera	Nemouridae	<i>Amphinemura decemseta</i>	MK132341	AD1	0.01						
			MK132323	AD2	0.05	3	1.00	6	2	0.67	5
			MK132340	AD3	0.10						
	Perlidae	<i>Kamimuria tibialis</i>	MZ543957	KT1	0.01						
			MZ543958	KT2	0.05						
			MZ543959	KT3	0.10						
			MZ543960	KT4	0.50	7	1.00	245	5	0.86	169
			MZ543961	KT5	1.00						
			MZ543962	KT6	5.00						
			MZ543963	KT7	10.00						
	Capniidae	<i>Eucapnopsis bulba</i>	Unr. Seq.	EB1	0.01						
			Unr. Seq.	EB2	0.05						
			Unr. Seq.	EB3	0.10	5	1.00	104	5	1.00	71
			Unr. Seq.	EB4	0.50						
			Unr. Seq.	EB5	1.00						
Ephemeroptera	Heptageniidae	<i>Epeorus latifolium</i>	MZ543952	EL1	0.01						
			MZ543953	EL2	0.05						
			MZ543954	EL3	0.10	5	1.00	19	5	1.00	14
			MZ543955	EL4	0.50						
			MZ543956	EL5	1.00						

Notes: [†]DNA concentration in the mock sample; [‡]Haplotypes based on the 658-bp Folmer region of the mtCOI gene; [§]421-bp internal COI fragment of the BF2 + BR2 primers. "*h*" refers to the number of haplotypes; "*Hd*" haplotype diversity; "*Eta*" Total number of mutations. "Unr. Seq." are unreleased sequences (Fasta files are provided in the supplementary materials).