Estimating intraspecific genetic diversity from community DNA metabarcoding data

Vasco Elbrecht [Corresp., 1, 2] , **Ecaterina Edith Vamos** [1] , **Dirk Steinke** [2] , **Florian Leese** [1, 3]

[1] Aquatic Ecosytem Research, University of Duisburg-Essen, Essen, NRW, Germany

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

[3] Centre for Water and Environmental Research (ZWU) Essen, University of Duisburg-Essen, Essen, NRW, Germany

Corresponding Author: Vasco Elbrecht
Email address: luckylion07@googlemail.com

Background. DNA metabarcoding is used to generate species composition data for entire communities. However, sequencing errors in high throughput sequencing instruments are fairly common, usually requiring reads to be clustered into operational taxonomic units (OTU), losing information on intraspecific diversity in the process. While COI haplotype information is limited in resolution, it is nevertheless useful in a phylogeographic context, helping to formulate hypothesis on taxon dispersal.

Methods. This study combines sequence denoising strategies, normally applied in microbial research, with additional abundance-based filtering to extract haplotypes from freshwater macroinvertebrate metabarcoding data sets. This novel approach was added to the R package "JAMP" and can be applied to Cytochrome c oxidase subunit I (COI) amplicon datasets. We tested our haplotyping method by sequencing i) a single-species mock community composed of 31 individuals with different haplotypes spanning three orders of magnitude in biomass and ii) 18 monitoring samples each amplified with four different primer sets and two PCR replicates.

Results. We detected all 15 haplotypes of the single specimens in the mock community with relaxed filtering and denoising settings. However, up to 480 additional unexpected haplotypes remained in both replicates. Rigorous filtering removes most unexpected haplotypes, but also can discard expected haplotypes mainly from the small specimens. In the monitoring samples, the different primer sets detected 177 - 200 OTUs, each containing an average of 2.40 to 3.30 haplotypes per OTU. Population structures were consistent between replicates, and similar between primer pairs, depending on the primer length. A closer look at abundant taxa in the data set revealed various population genetic patterns, e.g. *Taeniopteryx nebulosa* and *Hydropsyche pellucidula* with a difference in north-south haplotype distribution, while *Oulimnius tuberculatus* and *Asellus aquaticus* display no clear population pattern but differ in genetic diversity.

Discussion. We developed a strategy to infer intraspecific genetic diversity from bulk invertebrate monitoring samples using metabarcoding data. It needs to be stressed that at this point metabarcoding-informed haplotyping is not capable of capture the full diversity present in such samples, due to variation in specimen size, primer bias and loss of sequence variants with low abundance. Nevertheless, for a high number of species intraspecific diversity was recovered, identifying potentially isolated populations and potential taxa for further more detailed phylogeographic investigation. While we are currently lacking large-scale metabarcoding data sets to fully take advantage of our new approach, metabarcoding-informed haplotyping holds great promise for biomonitoring efforts that not only seek information about biological diversity but also underlying genetic diversity.

1 **Title: Estimating intraspecific genetic diversity from community DNA**

2 **metabarcoding data**

4 **Running Title (45 char max):** Extracting haplotypes from metabarcoding data

5 **Authors:** Vasco Elbrecht[1,2*], Ecaterina Edith Vamos[1], Dirk Steinke[2], Florian Leese [1,3]

7 Affiliations:

8 1) Aquatic Ecosystem Research, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 5, 45141

9 Essen, Germany

10 2) Centre for Biodiversity Genomics, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1,

11 Canada

12 3) Centre for Water and Environmental Research (ZWU) Essen, University of Duisburg-Essen, Universitätsstraße 2,

13 45141 Essen, Germany

14 **Corresponding author: Vasco Elbrecht (vasco.elbrecht@uni-due.de),

16 **Abstract:**

17 **Background.** DNA metabarcoding is used to generate species composition data for entire communities. However,

18 sequencing errors in high throughput sequencing instruments are fairly common, usually requiring reads to be

19 clustered into operational taxonomic units (OTU), losing information on intraspecific diversity in the process. While

20 COI haplotype information is limited in resolution, it is nevertheless useful in a phylogeographic context, helping to

21 formulate hypothesis on taxon dispersal.

22 **Methods.** This study combines sequence denoising strategies, normally applied in microbial research, with

23 additional abundance-based filtering to extract haplotypes from freshwater macroinvertebrate metabarcoding data

24 sets. This novel approach was added to the R package "JAMP" and can be applied to Cytochrome c oxidase subunit

25 I (COI) amplicon datasets. We tested our haplotyping method by sequencing i) a single-species mock community

26 composed of 31 individuals with different haplotypes spanning three orders of magnitude in biomass and ii) 18

27 monitoring samples each amplified with four different primer sets and two PCR replicates.

28   **Results.** We detected all 15 haplotypes of the single specimens in the mock community with relaxed filtering and

29   denoising settings. However, up to 480 additional unexpected haplotypes remained in both replicates. Rigorous

30   filtering removes most unexpected haplotypes, but also can discard expected haplotypes mainly from the small

31   specimens. In the monitoring samples, the different primer sets detected 177 - 200 OTUs, each containing an

32   average of 2.40 to 3.30 haplotypes per OTU. Population structures were consistent between replicates, and similar

33   between primer pairs, depending on the primer length. A closer look at abundant taxa in the data set revealed various

34   population genetic patterns, e.g. *Taeniopteryx nebulosa* and *Hydropsyche pellucidula* with a difference in north-

35   south haplotype distribution, while *Oulimnius tuberculatus* and *Asellus aquaticus* display no clear population pattern

36   but differ in genetic diversity.

37   **Discussion.** We developed a strategy to infer intraspecific genetic diversity from bulk invertebrate monitoring

38   samples using metabarcoding data. It needs to be stressed that at this point metabarcoding-informed haplotyping is

39   not capable of capture the full diversity present in such samples, due to variation in specimen size, primer bias and

40   loss of sequence variants with low abundance. Nevertheless, for a high number of species intraspecific diversity was

41   recovered, identifying potentially isolated populations and potential taxa for further more detailed phylogeographic

42   investigation. While we are currently lacking large-scale metabarcoding data sets to fully take advantage of our new

43   approach, metabarcoding-informed haplotyping holds great promise for biomonitoring efforts that not only seek

44   information about biological diversity but also underlying genetic diversity.

45

46

47   **Introduction**

48   High-throughput analysis of DNA barcodes retrieved from environmental samples, i.e. DNA metabarcoding, allows

49   for the rapid and standardized assessment of community composition without the need for morpho-taxonomy

50   (Taberlet et al., 2012a; Creer et al., 2016). This new surge of data enables biodiversity surveys at speeds and scales

51   that were previously inconceivable in ecological and evolutionary studies. While the approach has major strengths

52   and is generally regarded as a game changer for ecological research (Creer et al., 2016), it still has limitations such

53   as the fact that sequences are typically clustered into operational taxonomic units (OTUs, Fig. S1) thereby ignoring

54   any intraspecific sequence variation (Callahan, McMurdie & Holmes, 2017). However, clustering is often used to

55   reduce the influence of PCR and sequencing errors that can otherwise generate false OTUs (Edgar, 2013). The

56   inability to detect sequence variation within OTUs hampers our ability to detect impacts at population level.

57   Simultaneous assessment of inter- and intraspecific diversity, however, represents a leap forward in ecological

58   research and management because haplotype data are direct proxies for spatio-temporal dynamics of populations

59   and both parameters can differ substantially (Taberlet et al., 2012b). In particular the assessment of fragmentation

60   (e.g. Weiss & Leese 2016) or changes in population size in response to environmental impacts are key areas of basic

61   and applied ecological research (e.g. Sutherland et al. 2012). For management, this parameter is also important

62   because genetic variation is typically lost long before species or OTUs disappear (Bálint et al., 2011). Unfortunately,

63   methods to extract haplotype information from metabarcoding data sets are generally not widely available and thus

64   most studies are based on single-specimen analyses. Some of those are based on denoising algorithms capable of

65   distinguishing between true haplotypes and sequencing noise (e.g. (Tikhonov, Leach & Wingreen, 2015; Eren et al.,

66   2015; Edgar, 2016; Callahan et al., 2016; Amir et al., 2017) and have been tested for microbial samples (e.g. (Eren

67   et al., 2015; Callahan et al., 2016; Needham, Sachdeva & Fuhrman, 2017). Wares & Pappalardo (2016) suggested

68   that haplotype information in metazoan datasets can be used to, for instance, improve taxa abundance estimates,

69   which was successfully demonstrated with freshwater fish fecal samples (Corse et al., 2017). Recent studies were

70   also able to infer haplotypes with metabarcoding for single specimens (Shokralla et al., 2014), arthropod bulk

71   samples (Elbrecht & Leese, 2015; Pedro et al., 2017) and environmental water samples (Sigsgaard et al., 2016), all

72   highlighting the possibility to extract sequence variant information within OTUs when targeting metazoan taxa.

73   We here further explore bioinformatics strategies in order to unlock the potential of metabarcoding based

74   haplotyping of entire and complex metazoan communities. We combined stringent quality filtering of reads with the

75   recently developed *unoise3* denoising strategy (Edgar, 2016) and calibrated this approach using a previously

76   characterized single-species mock sample composed of specimens with known haplotypes (Elbrecht & Leese, 2015;

77   Vamos, Elbrecht & Leese, 2017). Subsequently, we collected multi-species metabarcoding data from 18 sample

78   sites as part of a governmental freshwater macroinvertebrate biomonitoring program (Elbrecht et al., 2017). These

79   were denoised with the developed strategy and we tested the potential to detect intraspecific variation over a broad

80   geographic gradient across multiple taxa.

81

82   **Materials & Methods**

83   We tested our haplotyping strategy on two available DNA metabarcoding datasets, 1) a single-species mock sample

84   containing 31 specimens with known haplotypes from an earlier population genetics project (Elbrecht et al., 2014;

85   Vamos, Elbrecht & Leese, 2017) and 2) a multi-species macroinvertebrate community dataset from the Finnish

86   governmental stream monitoring program (Elbrecht et al., 2017). Haplotypes were determined by bidirectional

87   sanger sequencing for the single species mock samples (Elbrecht et al., 2014), while the multi-species sample was

88   metabarcoded on Illumina systems using several primer sets (Elbrecht & Leese, 2015; 2017; Vamos, Elbrecht &

89   Leese, 2017). Resulting OTU centroids were assembled into haplotypes as described in Elbrecht & Leese (2017).

90   The samples were sequenced for a region nested within the classical Folmer COI region (Folmer et al., 1994) with

91   two replicates each. The single-species sample was sequenced using a short primer set amplifying 178 bp, while the

92   multi-species monitoring samples were amplified using four different primer sets targeting a region of up to 421 bp

93   (Elbrecht & Leese, 2017). Paired-end sequencing (250 bp) was performed on Illumina MiSeq and HiSeq systems

94   with high sequencing depth (on average 1.53 million reads per sample, SD = 0.29).

95   To extract individual haplotypes from the metabarcoding datasets, we used strict quality filtering followed by

96   denoising (unoise3 Edgar, 2016, with additional threshold-based filtering steps, see Fig. 1B). The full metabarcoding

97   and haplotyping pipelines are available as part of the "**J**ust **A**nother **M**etabarcoding **P**ipeline" (JAMP) R package

98   (https://github.com/VascoElbrecht/JAMP), which uses Usearch v10.0.240 (Edgar, 2013), Vsearch v2.4.3 (Rognes et

99   al., 2016) and Cutadapt 1.9 (Martin, 2011) for most of the data processing. The advantage of the JAMP wrapper is

100   its modularity and the automated generation of additional summary statistics and extended quality filtering options.

101   All pipeline commands used are also available as supporting information (Fig. S2, Scripts S1, JAMP v0.28). In short,

102   pre-processing of reads involved sample demultiplexing, paired-end merging, primer trimming, generation of

103   reverse complements where needed (to align all reads in the forward direction), maximum expected error (ee)

104   filtering = 0.5 (Edgar & Flyvbjerg, 2015), only keeping reads of exact length targeted by the respective primer set,

105   subsampling to 1 and 0.4 million reads, respectively, to generate the same sequencing depth for the single species

106   and monitoring samples. To further reduce the amount of sequences affected by sequencing errors we discarded

107   sequences below 10 reads or 0.001% abundance in each sample and applied read denoising with unoise3 after

108   pooling all samples as implemented in Usearch (Edgar, 2016) using only reads with >= 10 abundance in each

109   sample after dereplication. Different expected error cutoffs and alpha values were tested, with ee = 0.5 and alpha = 5

110   being used for the final analysis of the 18 monitoring samples. With lower ee values, more low quality sequences

111 were discarded (Edgar & Flyvbjerg, 2015). Similarly, lower alpha values led to more strict denoising with unoise3

112 (Edgar, 2016).

113 For the single-species mock sample, the denoised and quality filtered reads (prior to denoising) were mapped against

114 the expected 15 haplotype sequences using Vsearch (Rognes et al., 2016). The unoise3 implementation in the JAMP

115 package adds additional threshold-based filtering after the denoising step, which we used for the Finnish multi-

116 species monitoring samples in order to discard haplotypes with less than 0.01% abundance in at least one sample

117 and OTUs with less than 0.1% abundance in at least one sample ("Denoise(... , minhaplosize = 0.01, OTUmin =

118 0.1)"). All read mapping steps of denoised data were done with Vsearch. Additionally, within each OTU and sample

119 site, only haplotypes with at least 5% abundance per sample were considered for generating haplotype maps and

120 networks, in order to exclude low abundance OTUs which can be difficult to separate from PCR artifacts and

121 sequencing errors (withinOTU = 5). The Denoise function also includes presence based filtering for larger datasets,

122 requiring a specific haplotype or OTU being present in a minimum number of samples (minHaploPresence=1 or

123 minOTUPresence=1). However, as we had only 18 sample sites available this filtering was not applied to the dataset.

124

125

126 **Results**

127 Our approach starts with denoising of quality filtered reads using unoise3 (Edgar, 2016) followed by an additional

128 threshold-based filtering step which includes OTU clustering of denoised reads (Edgar, 2013) and the removal of

129 low abundant OTUs / haplotypes (see Fig. 1B). We validated this approach by using a single species mock

130 community of known haplotype composition (Elbrecht & Leese, 2015), in which we found 943 unexpected

131 haplotypes above 0.003% abundance with no expected error filtering applied (Fig. 1A). Filtering the raw sequence

132 data with different quality thresholds (max ee, Edgar & Flyvbjerg, 2015) reduced the number of unexpected

133 haplotypes by only up to 10.22% (Fig S3). The consistency between the two independent sequencing replicates

134 indicates that a major fraction of the detected haplotypes represent in fact, real biological signal (e.g. somatic

135 mutations, numts or heteroplasmy, (Bensasson et al., 2001; Shokralla et al., 2014), which is difficult to differentiate

136 from PCR and sequencing errors. Even after using different alpha values for the unoise3 algorithm some unexpected

137 sequence variants remained (Fig S4). An error filtering of max ee = 0.5 in combination with an alpha of 5 was

138 chosen for subsequent analysis (Fig. 1C), as it offers the best trade-off between expected and unexpected haplotypes

139 (9 of 15 expected, 6 unexpected with low abundance), while retaining 67.08% (SD = 17.69%) of the original

140 sequence data after quality filtering and before denoising.

141   For the denoising of our multi-species monitoring samples, additional and more conservative filtering steps

142 were introduced to ensure only true sequence variants are included in the analysis (discarding low abundant OTUs

143 and haplotypes below 0.1% and 0.01%, as well as haplotypes below 5% read abundance within each OTU of the

144 respective sample, Fig. 1C green line). Denoising of metabarcoding data from 18 macroinvertebrate samples of the

145 Finnish routine stream monitoring, recovered 177 - 200 OTUs containing 534 - 646 haplotypes (on average 2.40 -

146 3.30 haplotypes per OTU, SD = 2.13 - 3.26) for the different primer pairs (Table S1). Most OTUs were only present

147 in a few sample locations, allowing for only limited population genetic analysis (Fig. S5, see also Fig. S7 in Elbrecht

148 et al., 2017). Fig. 2 depicts some examples of haplotype diversity and geographic distribution for more common and

149 widely distributed taxa in this study. For *Taeniopteryx nebulosa* (Plecoptera) and *Hydropsyche pellucidula*

150 (Trichoptera) we found distinct patterns of latitudinal variation in haplotype composition (Fig. 2A, B), while

151 *Oulimnius tuberculatus* (Coleoptera) showed low genetic variation across all primer combinations (Fig. 2C, Fig.

152 S3C). *Asellus aquaticus* (Isopoda) on the other hand showed very high genetic diversity for endemic haplotypes (Fig.

153 2D).

154 Extracted haplotype patterns between replicates were highly reproducible (R^2 = 0.751, SD = 0.242), while at the

155 same time recovering more sequence variants with longer amplicons (Fig. S6). Taxon occurrence for the four taxa

156 analyzed in detail matched morphology based identifications (Elbrecht et al., 2017) in most cases (only four false

157 positive detections, Fig. 2). The few inconsistencies between replicates in haplotypes and taxa occurrence are mostly

158 affecting low abundance reads. In the sequence alignments, all four primer sets shared most of the variable positions

159 (Fig. S6).

160

161 **Discussion**

162 In this case study, we developed and demonstrated a bioinformatic strategy to process metabarcoding data first using

163 a controlled single-species approach, in order to extract intraspecific genetic diversity information from complex

164 multi-species metazoan environmental samples. While our multi-species dataset was limited to only 18 sampling

165 sites, and many taxa were not widely distributed (Elbrecht et al., 2017), we could still infer potential population

166 genetic patterns for some of the abundant and more widespread taxa. Where available, observed population genetic

167   patterns were also consistent with previous studies, e.g. earlier work reported high genetic diversity for *A. aquaticus*

168   (Sworobowicz et al., 2015). Other published work, e.g. on *H. pellucidula* (Múrria et al., 2010) and *O. tuberculatus*

169   (Čiampor & Kodada, 2010) was too limited in sampling size and region for proper comparison.

170   Deriving haplotypes from metabarcoding data does not require specialized field or laboratory protocols, as existing

171   data is analyzed. And while our dataset is very limited with just 18 sample sites, there are efforts underway to

172   implement DNA metabarcoding-based monitoring of stream water quality in Europe, potentially generating HTS

173   data for thousands of sample sites every year (Leese et al., 2016). Such haplotype data, even though limited in

174   resolution and based only on a single gene marker, could be used to formulate hypotheses about taxa dispersal at an

175   unprecedented scale (Hughes, Schmidt & FINN, 2009), which would be highly beneficial for the renaturation and

176   management of aquatic ecosystems.

177   While the detection of haplotypes from bulk samples was demonstrated in this and other studies (Sigsgaard et al.,

178   2016; Corse et al., 2017; Pedro et al., 2017), the limitations of metabarcoding-based haplotyping remain relatively

179   unexplored. Metabarcoding data sets can be affected by primer bias (Elbrecht & Leese, 2015), tag switching (Esling,

180   Lejzerowicz & Pawlowski, 2015; Schnell, Bohmann & Gilbert, 2015), as well as PCR and sequencing errors

181   (Nakamura et al., 2011; Tremblay et al., 2015). Such issues can lead to artificial haplotypes, which are usually

182   sufficiently different to distinguish them from actual haplotypes in the samples, especially if they are less abundant

183   and thus likely influenced by stochastic effects (Leray & Knowlton, 2017). We applied very strict quality filtering in

184   our pipeline, and cautiously discarded all haplotypes below 5% abundance within an OTU. This is necessary, as low

185   abundant haplotypes can not be separated from sequencing errors (Nakamura et al., 2011; Tremblay et al., 2015),

186   somatic mutations (Shokralla et al., 2014) and other noise in the data, as we have shown for the single species mock

187   samples. Strict filtering will remove rare and low abundant haplotypes, but it is necessary to reduce the amount of

188   false positive artificial sequences that result from the currently rather high error rates of HTS instruments. Even with

189   such strict filtering settings, we can not be fully confident that all false haplotypes were excluded e.g. as the result of

190   undetected chimeric sequences (Edgar et al., 2011) or systematic sequencing errors (Nakamura et al., 2011;

191   Schirmer et al., 2015; Schirmer, 2016) that likely persist across replicates. Approaches relying on the comparison of

192   replicate samples could be an appropriate strategy in particular when working with unicellular organisms (Lange et

193   al., 2015). However, for our metazoan communities many variants occur within both replicates (Fig. 1).

194   Macroinvertebrate communities can vary considerably in biomass, which means rare and small specimens will be

195   underrepresented when extracting DNA from bulk samples (Elbrecht, Peinert & Leese, 2017). Thus, taxa in the

196   sample are sequenced at different sequencing depth, which likely has an influence on the amount of false haplotypes

197   detected within each OTU. Additionally, differences in specimen biomass can skew the detection of haplotypes, as

198   only those of large specimens will be retained in bioinformatics analysis (haplotypes of small specimens are likely

199   below 5% abundance). Such uncertainties need to be considered when doing population genetic analysis, which is

200   usually done at specimen level, with the exact number of specimens and haplotypes known for each sampling site. It

201   has to be emphasized that at this point metabarcoding-based haplotyping only provides very limited information of

202   genetic diversity and phylogeography of a given taxon. However, interesting patterns emerging from such studies

203   can be subsequently explored by collecting taxa of interest and using standard population genetic markers with a

204   higher resolution (e.g. microsatellites, ddRAD Peterson et al., 2012). Our study demonstrates the feasibility and

205   potential of metabarcoding data for the investigation of population genetic patterns of entire complex environmental

206   communities. The shortcomings and the level of resolution of this novel approach need to be carefully tested (e.g. by

207   constructing mock samples using synthesized DNA). Additionally, more bioinformatics approaches suited for the

208   analysis of metazoan bulk samples need to be developed, especially with respect to variation in specimen biomass

209   (Elbrecht, Peinert & Leese, 2017). Furthermore, most software currently used in this field was developed for

210   microbial samples and should therefore be further tested and benchmarked for its feasibility in studies involving

211   eukaryotes. Despite the clear limitations of this haplotyping approach, we are confident that it will be useful in

212   future large-scale studies of genetic diversity. While metabarcoding studies will remain affected by sequencing

213   errors (potentially leading to false haplotypes), we expect that most of these issues can be mitigated by increasing

214   the number of sampling sites to several hundred or even thousands. For large-scale efforts such as routine

215   monitoring using metabarcoding (Baird & Hajibabaei, 2012; Gibson et al., 2015; Elbrecht et al., 2017), this might

216   soon become a feasible option if not standard. Additionally, references databases should be further completed and

217   extended to cover a large geographic range in order to assign species names and ground truth the detected

218   haplotypes (Carew et al., 2017; Curry et al., 2018).

219

Conclusions

221   Our study demonstrates that haplotypes can be extracted from complex metazoan metabarcoding datasets. This

222   proof of concept work already shows emerging population genetic patterns for a few species, but more large-scale

223   validation studies are needed to explore the limitations and the potential of metabarcoding-based haplotyping. While

224   some shortcomings such as occasional false positive detections and loss of rare and small taxa are difficult to

225   overcome per sample for such complex communities, these can be partly offset by studying comparative patterns of

226   intraspecific variation across many taxa and sites. As metabarcoding becomes more accessible and larger DNA-

227   based biodiversity assessment and monitoring initiatives emerge, sampling and extracting haplotypes from hundreds

228   of sites might become a feasible path of future research.

229

230

231

232   **Data availability.** Unprocessed raw sequence data are available from previous studies on the NCBI SRA archive.

233   Single species mock sample: SRR5295658 and SRR5295659 (Vamos, Elbrecht & Leese, 2017), monitoring samples:

234   SRR4112287 (Elbrecht et al., 2017). The JAMP R package is available on GitHub

235   (github.com/VascoElbrecht/JAMP) with the used R scripts (Script S1) and full haplotype tables (Table S1) available

236   as supporting information.

237

238

239

240

241 **Figures**

242

243 **Figure 1:** Overview of DNA metabarcoding data of a single-species mock sample containing specimens with 15

244 distinct haplotypes (black circles). Detected haplotypes (unexpected ones shown in grey and blue) plotted against

245 specimen biomass for the processed data (**A**) and followed by read denoising using unoise3 (**C**). Denoising was

246 applied to both replicates individually, with a circle if the read was detected in both samples (error bar = SD) and A

247 or B if the read was found in only one replicate. For processing of large-scale samples (**B**, Fig. 2), all samples were

248 pooled and jointly denoised, followed by OTU clustering and read mapping then followed by discarding of

249 haplotypes below a 5% threshold within each sample.

250

251 **Figure 2**: Haplotype maps and networks extracted from multi-species monitoring metabarcoding datasets amplified

252 with the BF2+BR2 primer set for four abundant macroinvertebrate taxa (A = *Taeniopteryx nebulosa*, B =

253 *Hydropsyche pellucidula*, C = *Oulimnius tuberculatus*, D = *Asellus aquaticus*). Numbers next to each sampling site

254 indicate sample size of the respective taxa based on morphological identification in a sample (Elbrecht et al., 2017).

255 Conflicts between DNA and morphology-based detections are highlighted in yellow. Haplotype frequency

256 composition per site is indicated by pie charts. For *A. aquaticus* only the 10 most common haplotypes are visualised

257 with different colours (remaining ones in white). Each crossline in a network represents one base pair difference

258 between the respective haplotypes. Dashed lines around a circle indicate novel haplotypes that were not available in

259 the BOLD reference database. An A or B next to a haplotype in the map or network indicates the presence of this

260 haplotype in only in one replicate. Shapefile-data © OpenStreetMap contributors, licensed under Creative Commons

261 2.0 (CC BY-SA).

263   **Acknowledgements**

264   We would like to thank members of the leeselab for helpful discussions.

265

266

267   **Supporting information**

268   **Figure S1:** Schematic overview of errors affecting metabarcoding data and clustering / denoising strategies to

269   reduce them.

270   **Figure S2:** Overview of the haplotyping strategy used here and their implementation in the JAMP R package.

271   **Figure S3:** Effect of different quality filtering (max ee) on reads of the single species mock sample.

272   **Figure S4:** Effect of different alpha values in read denoising of the single-species mock sample.

273   **Figure S5:** Bar plots of haplotype distribution within each OTU.

274   **Figure S6:** Detailed plots of four example taxa from the denoised multi-species monitoring samples, showing

275   haplotype maps & networks, similarity between replicates and sequence alignment for all BF/BR primer sets.

276   **Table S1:** Finland haplotype table (for all four different primer combinations).

277   **Scripts S1:** Metabarcoding and denoising pipeline, and additional scripts used to produce the figures.

278

279
280
281   Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde
282   ER, Gonzalez A, Knight R 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.
283   *mSystems* 2:e00191–16–7. DOI: 10.1128/mSystems.00191-16.
284   Baird DJ, Hajibabaei M 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-
285   generation DNA sequencing. 21:2039–2044.
286   Bálint M, Domisch S, Engelhardt CHM, Haase P, Lehrian S, Sauer J, Theissinger K, Pauls SU, Nowak C 2011.
287   Cryptic biodiversity loss linked to global climate change. *Nature Climate Change* 1:1–6. DOI:
288   10.1038/nclimate1191.
289   Bensasson D, Zhang DX, Hartl DL, Hewitt GM 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses.
290   *Trends in ecology & evolution (Personal edition)* 16:314–321.
291   Callahan BJ, McMurdie PJ, Holmes SP 2017. Exact sequence variants should replace operational taxonomic units in
292   marker-gene data analysis. :1–5. DOI: 10.1038/ismej.2017.119.
293   Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP 2016. DADA2: High-resolution sample
294   inference from Illumina amplicon data. *Nature Methods* 13:581–583. DOI: 10.1038/nmeth.3869.
295   Carew ME, Nichols SJ, Batovska J, St Clair R, Murphy NP, Blacket MJ, Shackleton ME 2017. A DNA barcode
296   database of Australia's freshwater macroinvertebrate fauna. *Marine and freshwater research*:1–15. DOI:
297   10.1071/MF16304.
298   Corse E, MEGLÉCZ E, Archambaud G, Ardisson M, MARTIN J-F, Tougard C, Chappaz R, DUBUT V 2017. A
299   from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet
300   metabarcoding studies. *Molecular ecology resources* 17:e146–e159. DOI: 10.1111/1755-0998.12703.
301   Creer S, Deiner K, Frey S, Porazinska D, Taberlet P, Thomas WK, Potter C, Bik HM 2016. The ecologist's field

302   guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* 7:1008–1018. DOI:
303   10.1111/2041-210X.12574.
304 Curry CJ, Gibson JF, Shokralla S, Hajibabaei M, Baird DJ 2018. Identifying North American freshwater
305   invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? *Freshwater Science*
306   37:178–189. DOI: 10.1086/696613.
307 Čiampor F Jr, Kodada J 2010. Taxonomy of the Oulimnius tuberculatus species group (Coleoptera: Elmidae) based
308   on molecular and morphological data. *Zootaxa*.
309 Edgar RC 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*
310   10:996–998. DOI: 10.1038/nmeth.2604.
311 Edgar RC 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. DOI:
312   10.1101/081257.
313 Edgar RC, Flyvbjerg H 2015. Error filtering, pair assembly and error correction for next-generation sequencing
314   reads. *Bioinformatics* 31:3476–3482. DOI: 10.1093/bioinformatics/btv401.
315 Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R 2011. UCHIME improves sensitivity and speed of chimera
316   detection. *Bioinformatics* 27:2194–2200. DOI: 10.1093/bioinformatics/btr381.
317 Elbrecht V, Leese F 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer
318   Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PloS ONE*
319   10:e0130324–16. DOI: 10.1371/journal.pone.0130324.
320 Elbrecht V, Leese F 2017. Validation and development of freshwater invertebrate metabarcoding COI primers for
321   Environmental Impact Assessment. *Frontiers in Freshwater Science*. DOI: 10.3389/fenvs.2017.00011.
322 Elbrecht V, Feld CK, Gies M, Hering D, Sondermann M 2014. Genetic diversity and dispersal potential of the
323   stonefly Dinocras cephalotes in a central European low mountain range. *Freshwater Science* 33:181–192. DOI:
324   10.1086/674536.
325 Elbrecht V, Peinert B, Leese F 2017. Sorting things out: Assessing effects of unequal specimen biomass on DNA
326   metabarcoding. *Ecology and Evolution* 7:6918–6926. DOI: 10.1002/ece3.3192.
327 Elbrecht V, Vamos E, Meissner K, Aroviita J, Leese F 2017. Assessing strengths and weaknesses of DNA
328   metabarcoding based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and*
329   *Evolution*:1–21. DOI: 10.7287/peerj.preprints.2759v2.
330 Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML 2015. Minimum entropy decomposition:
331   Unsupervised oligotyping for sensitive partitioning of high- throughput marker gene sequences. 9:968–979.
332   DOI: 10.1038/ismej.2014.195.
333 Esling P, Lejzerowicz F, Pawlowski J 2015. Accurate multiplexing and filtering for high-throughput amplicon-
334   sequencing. *Nucleic acids research* 43:2513–2524. DOI: 10.1093/nar/gkv107.
335 Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R 1994. DNA primers for amplification of mitochondrial
336   cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and*
337   *biotechnology* 3:294–299.
338 Gibson JF, Shokralla S, Curry C, Baird DJ, Monk WA, King I, Hajibabaei M 2015. Large-Scale Biomonitoring of
339   Remote and Threatened Ecosystems via High-Throughput Sequencing. *PloS one* 10:e0138432–15. DOI:
340   10.1371/journal.pone.0138432.
341 Hughes JM, Schmidt DJ, FINN DS 2009. Genes in Streams: Using DNA to Understand the Movement of
342   Freshwater Fauna and Their Riverine Habitat. *BioScience* 59:573–583. DOI: 10.1525/bio.2009.59.7.8.
343 Lange A, Jost S, Heider D, Bock C, Budeus B, Schilling E, Strittmatter A, Boenigk J, Hoffmann D 2015.
344   AmpliconDuo: A Split-Sample Filtering Protocol for High-Throughput Amplicon Sequencing of Microbial
345   Communities. *PloS one* 10:e0141590–22. DOI: 10.1371/journal.pone.0141590.
346 Leese F, Altermatt F, Bouchez A, Ekrem T 2016. DNAqua-Net: Developing new genetic tools for bioassessment
347   and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes*. DOI: 10.3897/rio.2.e11321.
348 Leray M, Knowlton N 2017. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina
349   COI metabarcoding. *PeerJ* 5:e3006–27. DOI: 10.7717/peerj.3006.
350 Martin M 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*
351   17:10–12.
352 Múrria C, Zamora-Muñoz C, Bonada N, Ribera C, Prat N 2010. Genetic and morphological approaches to the
353   problematic presence of three Hydropsychespecies of the pellucidulagroup (Trichoptera: Hydropsychidae) in
354   the westernmost Mediterranean Basin. *Aquatic Insects* 32:85–98. DOI: 10.1080/01650424.2010.482939.
355 Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi
356   H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S 2011. Sequence-specific error profile of Illumina sequencers.
357   *Nucleic acids research* 39:e90. DOI: 10.1093/nar/gkr344.

358 Needham DM, Sachdeva R, Fuhrman JA 2017. Ecological dynamics and co-occurrence among marine
359   phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME Journal*:1–16. DOI:
360   10.1038/ismej.2017.29.
361 Pedro PM, Piper R, Bazilli Neto P, Cullen L Jr., Dropa M, Lorencao R, Matté MH, Rech TC, Rufato MO Jr., Silva
362   M, Turati DT 2017. Metabarcoding Analyses Enable Differentiation of Both Interspecific Assemblages and
363   Intraspecific Divergence in Habitats With Differing Management Practices. *Environmental Entomology*:1–9.
364   DOI: 10.1093/ee/nvx166.
365 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE 2012. Double Digest RADseq: An Inexpensive Method
366   for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PloS one* 7:e37135. DOI:
367   10.1371/journal.pone.0037135.t001.
368 Rognes T, Flouri T, Nichols B, Quince C, Mahé F 2016. VSEARCH: a versatile open source tool for metagenomics.
369   *PeerJ* 4:e2584–22. DOI: 10.7717/peerj.2584.
370 Schirmer M 2016. Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data. *BMC*
371   *bioinformatics*:1–15. DOI: 10.1186/s12859-016-0976-y.
372 Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C 2015. Insight into biases and sequencing errors for
373   amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research*:1–16. DOI:
374   10.1093/nar/gku1341.
375 Schnell IB, Bohmann K, Gilbert MTP 2015. Tag jumps illuminated - reducing sequence-to-sample
376   misidentifications in metabarcoding studies. *Molecular ecology resources* 15:1289–1303. DOI: 10.1111/1755-
377   0998.12402.
378 Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M 2014. Next-generation DNA
379   barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single
380   specimens. *Molecular ecology resources*:n/a–n/a. DOI: 10.1111/1755-0998.12236.
381 Sigsgaard EE, Nielsen IB, Bach SS, Lorenzen ED, Robinson DP, Knudsen SW, Pedersen MW, Jaidah MA, Orlando
382   L, Willerslev E, Møller PR, THOMSEN PF 2016. Population characteristics of a large whale shark aggregation
383   inferred from seawater environmental DNA. *Nature Ecology & Evolution* 1:0004–5. DOI: 10.1038/s41559-
384   016-0004.
385 Sworobowicz L, Grabowski M, Mamos T, Burzyński A, Kilikowska A, Sell J, Wysocka A 2015. Revisiting the
386   phylogeography of *Asellus aquaticus* in Europe: insights into cryptic diversity and spatiotemporal
387   diversification. *Freshwater biology* 60:1824–1840. DOI: 10.1111/fwb.12613.
388 Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E 2012a. Towards next-generation biodiversity
389   assessment using DNA metabarcoding. *Molecular Ecology* 21:2045–2050. DOI: 10.1111/j.1365-
390   294X.2012.05470.x.
391 Taberlet P, Zimmermann NE, Englisch T, Tribsch A, Holderegger R, Alvarez N, Niklfeld H, Coldea G, Mirek Z,
392   Moilanen A, Ahlmer W, Marsan PA, Bona E, Bovio M, Choler P, Cieślak E, Colli L, Cristea V, Dalmas J-P,
393   Frajman B, Garraud L, Gaudeul M, Gielly L, Gutermann W, Jogan N, Kagalo AA, Korbecka G, Küpfer P,
394   Lequette B, Letz DR, Manel S, Mansion G, Marhold K, Martini F, Negrini R, Niño F, Paun O, Pellecchia M,
395   Perico G, Piękoś-Mirkowa H, Prosser F, Puşcaş M, Ronikier M, Scheuerer M, Schneeweiss GM, Schönswetter
396   P, Schratt-Ehrendorfer L, Schüpfer F, Selvaggi A, Steinmann K, Thiel-Egenter C, van Loo M, Winkler M,
397   Wohlgemuth T, Wraber T, Gugerli F, IntraBioDiv Consortium 2012b. Genetic diversity in widespread species
398   is not congruent with species richness in alpine plant communities. *Ecology letters* 15:1439–1448. DOI:
399   10.1111/ele.12004.
400 Tikhonov M, Leach RW, Wingreen NS 2015. Interpreting 16S metagenomic data without clustering to achieve sub-
401   OTU resolution. *The ISME Journal* 9:68–80. DOI: 10.1038/ismej.2014.117.
402 Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG 2015. Primer and
403   platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology* 6:8966–15. DOI:
404   10.3389/fmicb.2015.00771.
405 Vamos EE, Elbrecht V, Leese F 2017. Short COI markers for freshwater macroinvertebrate metabarcoding.
406   *Metabarcoding and Metagenomics*. DOI: 10.7287/peerj.preprints.3037v1.
407 Wares J, Pappalardo P 2016. Can Theory Improve the Scope of Quantitative Metazoan Metabarcoding? *Diversity*
408   8:1–15. DOI: 10.3390/d8010001.
409 Weiss M & Leese F (2016). Widely distributed and regionally isolated! Drivers of genetic structure in *Gammarus*
410   *fossarum* in a human-impacted landscape. BMC Evolutionary Biology, 1–14. DOI: 10.1186/s12862-016-0723-
411   z

Figure 1

Overview of DNA metabarcoding data of a single-species mock sample containing specimens with 15 distinct haplotypes (black circles).

Detected haplotypes (unexpected ones shown in grey and blue) plotted against specimen biomass for the processed data (**A**) and followed by read denoising using unoise3 (**C**). Denoising was applied to both replicates individually, with a circle if the read was detected in both samples (error bar = SD) and A or B if the read was found in only one replicate. For processing of large-scale samples (**B**, Fig. 2), all samples were pooled and jointly denoised, followed by OTU clustering and read mapping then followed by discarding of haplotypes below a 5% threshold within each sample.
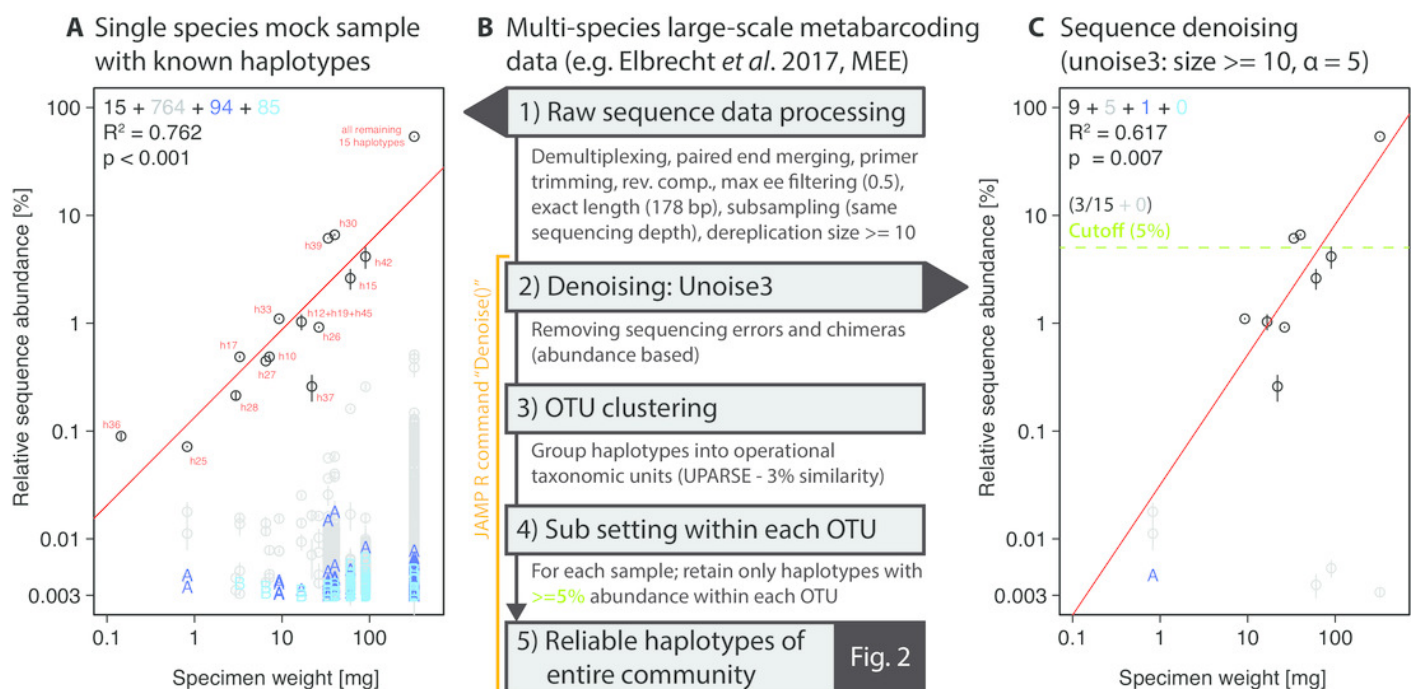
Figure 2

Haplotype maps and networks extracted from multi-species community metabarcoding datasets amplified with the BF2+BR2 primer set for four abundant macroinvertebrate taxa.

(A = *Taeniopteryx nebulosa*, B = *Hydropsyche pellucidula*, C = *Oulimnius tuberculatus*, D = *Asellus aquaticus*). Numbers next to each sampling site indicate sample size of the respective taxa based on morphological identification in a sample (Elbrecht et al., 2017) . Conflicts between DNA and morphology based-detections are highlighted in yellow. Haplotype frequency composition per site is indicated by pie charts. For *A. aquaticus* only the 10 most common haplotypes are visualised with different colours (remaining ones in white). Each crossline in a network represents one base pair difference between the respective haplotypes. Dashed lines around a circle indicate novel haplotypes that were not available in the BOLD reference database. An A or B next to a haplotype in the map or network indicates the presence of this haplotype in only in one replicate.

a *Taeniopteryx nebulosa* (Plecoptera)

b *Hydropsyche pellucidula* (Trichoptera)

c *Oulimnius tuberculatus* (Coleoptera)

d *Asellus aquaticus* (Isopoda)