Title: Assessing intraspecific genetic diversity from community DNA metabarcoding data

Running Title (45 char max): DNA metabarcoding with haplotype accuracy

Word count: max 1779, methods 376

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

Affiliations:

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

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

3) Centre for Water and Environmental Research (ZWU) Essen, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany

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

Key words: metabarcoding, high-throughput sequencing, haplotyping, population genetics, ecosystem assessment

DNA metabarcoding provides species composition data for entire communities, yet information on intraspecific diversity is usually lost during data analysis. The capacity to infer intraspecific genetic diversity within whole communities would, however, represent a leap forward for ecological monitoring and conservation. We developed an amplicon-based sequence denoising approach that allows the identification of haplotypes from metabarcoding data sets and demonstrate its power with two freshwater macroinvertebrate data sets.

High-throughput analysis of DNA barcodes retrieved from environmental samples, i.e. DNA metabarcoding, allows for rapid and standardized assessment of community composition without the need for morphotaxonomy [1,2]. This new surge of data now enables biodiversity surveys at speeds and scales that were previously inconceivable in ecological and evolutionary studies.. While the approach has major strengths and is generally regarded as a game changer for ecological research [1], it still has limitations such as the fact that sequence variation is typically clustered into operational taxonomic units (OTUs, Fig. S1) thereby ignoring any intraspecific variation [2]. However, clustering is a crucial step to reduce the influence of PCR and sequencing errors that can otherwise generate false sequence variation [3]. This inability to detect intraspecific variation hampers e.g. our ability to detect environmental impacts at population level, long before

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

complete species or OTUs are lost [3]. Unfortunately, methods to extract haplotype information from metabarcoding data sets are generally not available. Some methods based on denoising algorithms capable of distinguishing between true haplotypes and sequencing noise were recently developed (e.g. [4-8]) and tested for microbial samples (e.g. [4,5,9]). Initial studies of species from individual samples [10], mock bulk [11] or environmental samples [12] did not use DNA metabarcoding for haplotype inference using real-world metazoan bulk samples at ecosystem level. For this study we used metabarcoding data collected as part of a governmental biomonitoring program for freshwater macroinvertebrates [13] and single species mock samples with known haplotypes [14] to explore and validate denoising strategies for metazoan bulk samples.

Our approach starts with denoising of quality filtered reads with unoise3 [7] followed by an additional threshold based filtering step which includes OTU clustering of denoised reads [15] and the removal of low abundant OTUs / haplotypes (See figure 1B). We validated this approach by using a single species mock community of known haplotype composition [11], in which we found 943 unexpected haplotypes (above 0.003% abundance) (Figure 1A). Filtering the raw sequence data with different quality thresholds (max ee [16]) reduced the number of unexpected haplotypes by only up to 10.22% (Fig S2). This consistency between the two independent sequencing replicates indicates that a major fraction of the detected haplotypes represent real biological signal (e.g. somatic mutations, numts or heteroplasmy [10,17]), which is difficult to differentiate from PCR and sequencing errors. Even after using different alpha values for the unoise3 algorithm some unexpected sequence variants remained (Fig S3). An error filtering of max ee = 0.5 in combination with an alpha of 5 was chosen for subsequent analysis (Figure 1C), as it offers the best tradeoff between expected and unexpected haplotypes (9 of 15 expected, 6 unexpected with low abundance), while retaining 67.08% (SD = 17.69%) of the original sequence data after quality filtering and before denoising.

For the denoising of our environmental biomonitoring samples, additional and more conservative filtering steps were introduced to ensure only true sequence variants are included in the analysis (discarding low abundant OTUs and haplotypes below 0.1% and 0.01%, as well as haplotypes below 5% abundance within each OTU of the respective sample, Figure 1C green line). Denoising of metabarcoding data from 18 macroinvertebrate samples of the Finnish routine stream monitoring, recovered 177 - 200 OTUs containing 534 - 646 haplotypes for the different primer pairs (Table S1). Figure 2 depicts some examples of haplotype diversity and geographic distribution revealed from our haplotyping approach. For *Taeniopteryx nebulosa* (Plecoptera) and *Hydropsyche pellucidula* (Trichoptera) we were able to find distinct latitudinal variation in haplotype composition (Figure 2A, B), while *Oulimnius tuberculatus* (Coleoptera) shows low genetic variation for all primer combinations (Figure 2C, Fig S3C). *Asellus aquaticus* (Isopoda) on the other hand shows very high genetic diversity for endemic haplotypes (Figure 2D). Some of these results are consistent with previous studies, e.g. earlier work reported high genetic diversity for *A. aquaticus* [18]. Oher published work, e.g. on *H. pellucidula* [19] and *O. tuberculatus* [20] is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

too limited in sampling size and region for proper comparison and discussion of the reasons for dispersal and potential leading edge colonization after the last glacial maximum.

Extracted haplotype patterns between replicates were highly reproducible (R^2 = 0.751, SD = 0.242), while at the same time recovering more sequence variants with longer amplicons (Figure S4). Taxon occurrence matched morphology based identifications [13] in most cases (only four false positive detections, Figure 2). The few inconsistencies between replicates in haplotypes and taxa occurrence are mostly affecting low abundance reads, likely as a result of stochastic effects and the percentage thresholds at which haplotypes and OTUs are discarded [21]. While the sequence alignments of all four primer sets shared most of the variable positions (Figure S4), additional effects of primer bias [11], tag switching [22,23], PCR and sequencing errors [24,25] can't be fully excluded. Additionally, many prospective true haplotypes might have been discarded by strict threshold based filtering, possibly underestimating the true diversity by missing rare sequence variants.

The extension of DNA metabarcoding through inclusion of intraspecific genetic variation of species communities, represents a paradigm shift in ecological and genetic research. We demonstrated that haplotypes can be successfully extracted from metazoan metabarcoding datasets and used for various purposes such as comparative landscape genetic or phylogeographic analysis. Even though DNA metabarcoding provides only single marker information and still shows the presence of some background noise, the approach holds enormous potential to generate hypotheses and explore patterns of environmental factors affecting genetic variation within species inhabiting an ecosystem. Interesting phylogeographic patterns could be explored by collecting taxa of interest and using population genetic markers with a higher resolution (e.g. microsatellites, ddRAD, [26]). Thus, our new haplotyping strategy can have substantial impact on ecosystem assessment, conservation and management. If applied within the framework of concerted international sampling campaigns [27] such monitoring datasets could be utilized to estimate population connectivity, migration events and population growth or bottlenecks, which is key to successful management and restoration projects [28].

Methods

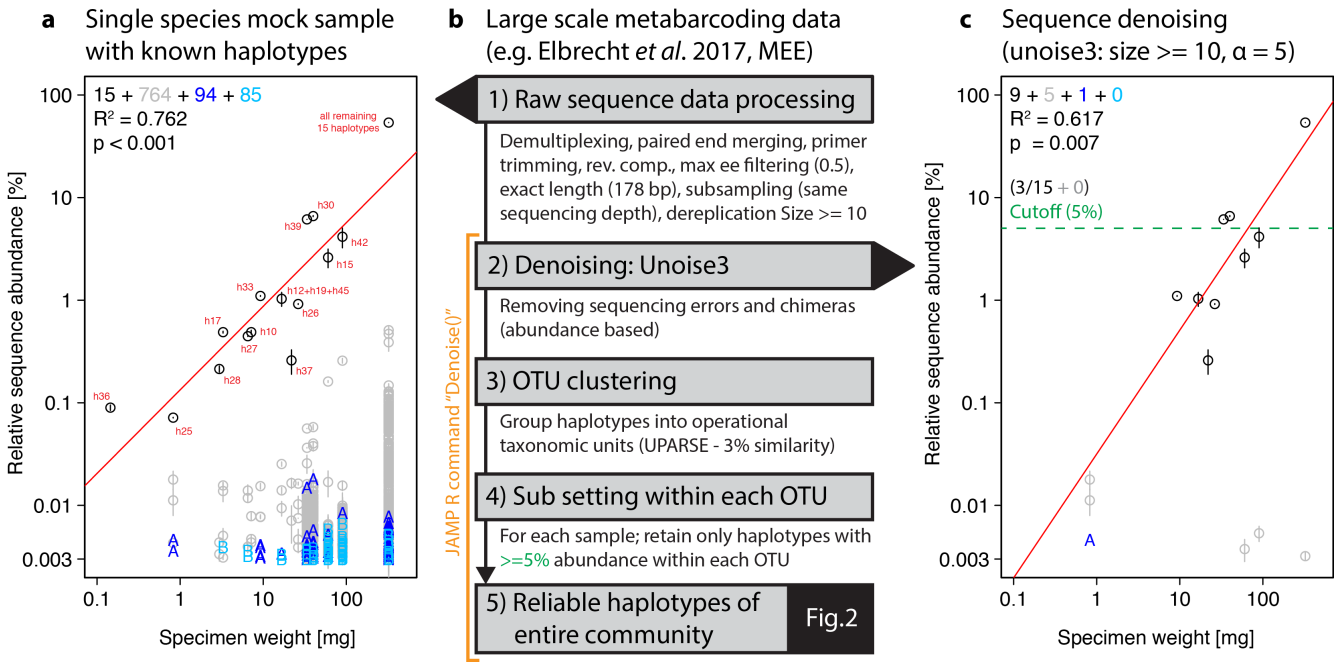
We tested our haplotyping strategy on two available DNA metabarcoding datasets, 1) a single species mock sample containing specimens with known haplotypes [14] and 2) a macroinvertebrate community dataset from the Finnish governmental stream monitoring program [13]. The samples were sequenced for a region nested within the classical Folmer COI fragment and with two replicates each. Hereby, the mock sample (1) was sequenced using a short primer set amplifying 178 bp, while the monitoring samples were amplified uising four different primer sets targeting a region of up to 421 bp [29]. Paired-end sequencing (250 bp) was performed on Illumina MiSeq and HiSeq systems with high sequencing depth (on average 1.53 million reads per sample, SD = 0.29).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

To extract individual haplotypes from the metabarcoding datasets, strict quality filtering followed by denoising (unoise3) was used, with additional threshold-based filtering steps (see Figure 1B). The full metabarcoding and haplotyping pipelines are available as an R package (https://github.com/VascoElbrecht/JAMP), which heavily relies on Usearch v10.0.240 [15]. The used pipeline commands are also available as supporting information (Figure S5, Scripts S1, JAMP v0.28). In short, pre-processing of reads involved sample demultiplexing, paired-end merging, primer trimming, generation of reverse complements where needed (to line all reads in the forward direction), max ee filtering (0.5 [16]), only keeping reads of exact length targeted by the respective primer set, subsampling to 1 and 0.4 million reads, respectively, to generate the same sequencing depth for the single species and monitoring samples. Read denoising was applied to all samples of a dataset using reads with >= 10 abundance in each dataset after dereplication. Different expected error cutoffs and alpha values were tested, with ee = 0.5 and alpha = 5 being used for the final analysis. For the single species mock sample, the denoised and quality filtered reads (prior to denoising) were then mapped against the expected 15 haplotype sequences using Vsearch (v2.4.3) [30]. The unoise3 implementation into the JAMP package allows for more quality filer reads, which we used for the Finnish monitoring samples in order to discard low abundant haplotypes and OTUs "Denoise(... , minhaplosize = 0.01, OTUmin = 0.1)". Additionally, within each OTU and sample site, only haplotypes with at least 5% abundance per sample were considered for generating haplotype maps and networks.

Data availability. Unprocessed raw sequence data are available from previous studies on the NCBI SRA archive (Singe species mock sample: SRR5295658 and SRR5295659 [14], monitoring samples: SRR4112287 [13]). The JAMP R package is available on GitHub (github.com/VascoElbrecht/JAMP) with the used R scripts (Script S1) and full haplotype tables (Table S1) available as supporting information.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017



115

116 **Figure 1:** Overview of DNA metabarcoding data of a single species mock sample containing specimens with 15 expected
117 haplotypes (black circles). Detected haplotypes (unexpected ones shown in grey and blue) plotted against specimen biomass
118 for the processed data (**A**) and followed by read denoising using unoise3 (**C**). Denoising was applied to both replicates
119 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
120 one replicate. For processing of larger scale samples (**B**, Fig. 2), all samples were pooled and denoised together, followed by
121 OTU clustering and read mapping then followed by discarding of haplotypes below a 5% threshold within each sample.

122

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

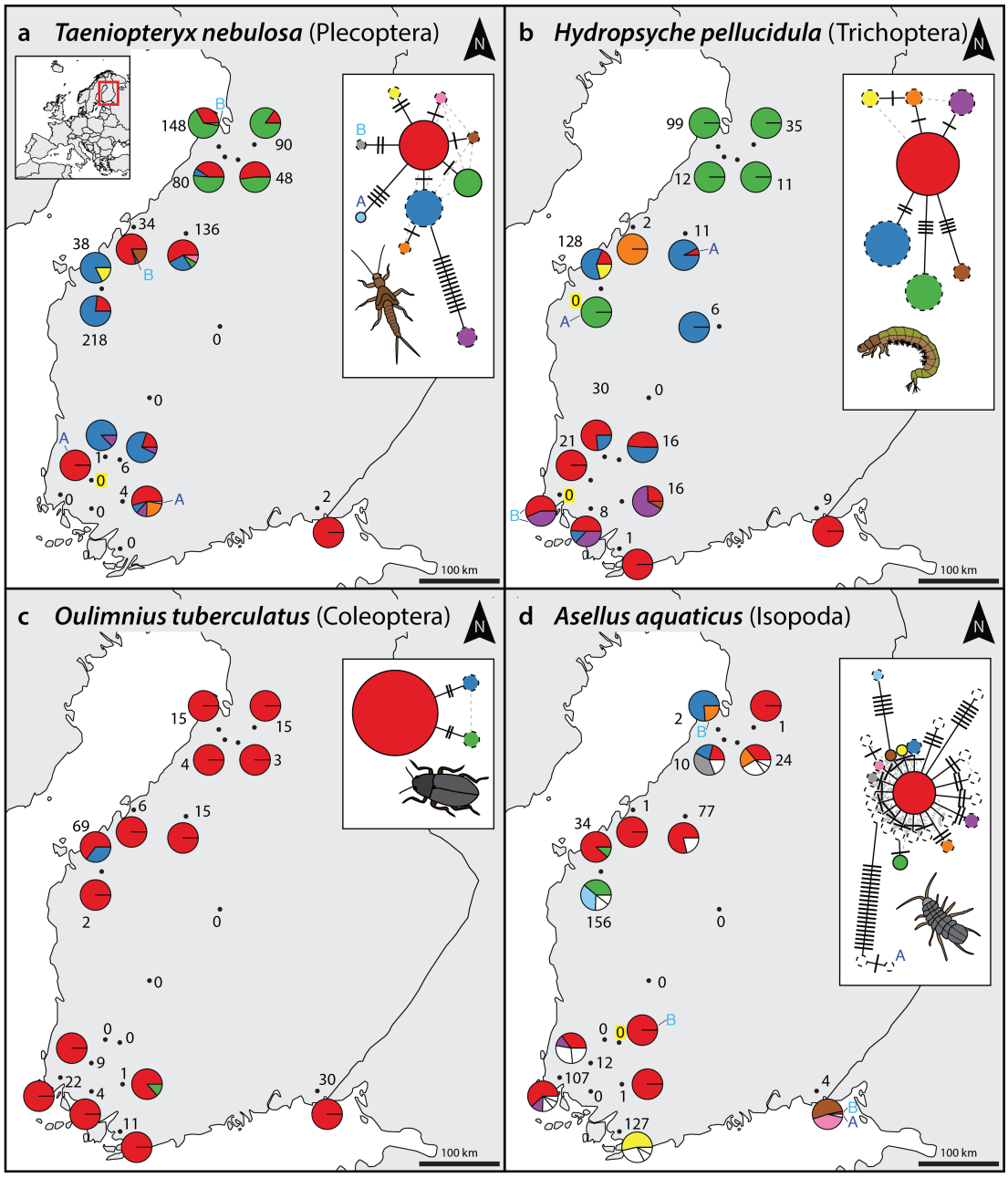


Figure 2: Haplotype maps and networks extracted from whole-community metabarcoding data sets for four abundant macroinvertebrate taxa (A = *Taeniopteryx nebulosa*, B = *Hydropsyche pellucidula*, C = *Oulimnius tuberculatus*, D = *Asellus aquaticus*). Numbers indicate sample size of the respective taxa based on morphological identification in a sample [13]. Conflicts between DNA and morphology based-detections are highlighted with a yellow background. Haplotype frequency composition per site is indicated with pie charts. For *A. aquaticus* only the 10 most common haplotypes are visualised with different colours (remaining ones in white). In the networks, each cross line represents one base pair difference between the respective haplotypes. An A or B next to a haplotype in the map or network indicates the presence of this haplotype in only in one replicate.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017

Acknowledgements

We would like to thank members of the leeselab for helpful discussions. This study is part of the European Cooperation in Science and Technology (COST) Action DNAqua-Net (CA15219).

Author contributions

V.E. developed the haplotyping concept, with contributions from E.E.V. and F.L., V.E. developed the bioinformatics and analysed the data, V.E., E.E.V. D.S., and F.L. wrote and revised the paper.

Supporting information

Figure S1: Schematic overview of errors affecting metabarcoding data and clustering / denoising strategies to reduce them.

Figure S2: Effect of different quality filtering (may ee) on reads of the singe species mock sample.

Figure S3: Effect of different alpha values in read denoising of the singe species mock sample.

Figure S4: Detailed plots of four example taxa from the denoised monitoring samples, showing haplotype maps & networks, similarity between replicates and sequence alignment for all BF/BR primer sets.

Figure S5: Overview of the haplotyping strategy used here and their implementation in the JAMP R package.

Table S1: Finland haplotype table (for all 4 different primer combinations).

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

1. Creer, S. et al. *Methods Ecol Evol* **7**, 1008–1018 (2016).
2. Callahan, B.J., McMurdie, P.J. & Holmes, S.P.1–5 (2017).doi:10.1038/ismej.2017.119
3. Bálint, M. et al. *Nature Climate Change* **1**, 1–6 (2011).
4. Eren, A.M. et al.**9**, 968–979 (2015).
5. Callahan, B.J. et al. *Nat Methods* **13**, 581–583 (2016).
6. Amir, A. et al. *mSystems* **2**, e00191–16–7 (2017).
7. Edgar, R.C. *bioRxiv* (2016).doi:10.1101/081257
8. Tikhonov, M., Leach, R.W. & Wingreen, N.S. *The ISME Journal* **9**, 68–80 (2015).
9. Needham, D.M., Sachdeva, R. & Fuhrman, J.A. *The ISME Journal* 1–16 (2017).doi:10.1038/ismej.2017.29
10. Shokralla, S. et al. *Mol Ecol Resour* n/a–n/a (2014).doi:10.1111/1755-0998.12236
11. Elbrecht, V. & Leese, F. *PLoS ONE* **10**, e0130324–16 (2015).
12. Sigsgaard, E.E. et al. *Nat. ecol. evol.* **1**, 0004–5 (2016).

13. Elbrecht, V., Vamos, E., Meissner, K., Aroviita, J. & Leese, F. *Methods Ecol Evol* 1–21 (2017).doi:10.7287/peerj.preprints.2759v2
14. Vamos, E.E., Elbrecht, V. & Leese, F. *PeerJ PrePrints* (2017).doi:10.7287/peerj.preprints.3037v1
15. Edgar, R.C. *Nat Methods* **10**, 996–998 (2013).
16. Edgar, R.C. & Flyvbjerg, H. *Bioinformatics* **31**, 3476–3482 (2015).
17. Bensasson, D., Zhang, D.X., Hartl, D.L. & Hewitt, G.M. *Trends Ecol Evol (Amst)* **16**, 314–321 (2001).
18. Sworobowicz, L. et al. *Freshwater Biol* **60**, 1824–1840 (2015).
19. Múrria, C., Zamora-Muñoz, C., Bonada, N., Ribera, C. & Prat, N. *Aquatic Insects* **32**, 85–98 (2010).
20. Čiampor, F., Jr & Kodada, J. *Zootaxa* (2010).
21. Leray, M. & Knowlton, N. *PeerJ* **5**, e3006–27 (2017).
22. Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. *Mol Ecol Resour* **15**, 1289–1303 (2015).
23. Esling, P., Lejzerowicz, F. & Pawlowski, J. *Nucleic Acids Res* **43**, 2513–2524 (2015).
24. Tremblay, J. et al. *Front. Microbiol.* **6**, 8966–15 (2015).
25. Nakamura, K. et al. *Nucleic Acids Res* **39**, e90 (2011).
26. Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. & Hoekstra, H.E. *PLoS ONE* **7**, e37135 (2012).
27. Leese, F., Altermatt, F., Bouchez, A. & Ekrem, T. *RIO* (2016).doi:10.3897/rio.2.e11321
28. Hughes, J.M., Schmidt, D.J. & FINN, D.S. *BioScience* **59**, 573–583 (2009).
29. Elbrecht, V. & Leese, F. *Front Fre Sci* (2017).doi:10.3389/fenvs.2017.00011
30. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. *PeerJ* **4**, e2584–22 (2016).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3269v1 | CC BY 4.0 Open Access | rec: 19 Sep 2017, publ: 19 Sep 2017