Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos

Vera Marie Alida Zizka1,3, Martina Weiss1,2, Florian Leese1,2

1 Aquatic Ecosystem Research, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, North Rhine-Westphalia, Germany
2 Centre for Water and Environmental Research (ZWU) Essen, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, North Rhine-Westphalia, Germany
3 Present address: Zoological Research Museum Alexander Koenig, Leibniz Institute for Animal Biodiversity, Adenauerallee 160, 53113 Bonn, North Rhine-Westphalia, Germany

Corresponding author: Vera Marie Alida Zizka (v.zizka@leibniz-zfmk.de)

Abstract

Genetic diversity is the most basal level of biodiversity and determines the evolutionary capacity of species to adapt to changing environments, yet it is typically neglected in routine biomonitoring and stressor impact assessment. For a comprehensive analysis of stressor impacts on genetic diversity, it is necessary to assess genetic variants simultaneously in many individuals and species. Such an assessment is not as straightforward and usually limited to one or few focal species. However, nowadays species diversity can be assessed by analysing thousands of individuals of a community simultaneously with DNA metabarcoding. Recent bioinformatic advances also allow for the extraction of exact sequence variants (ESVs or haplotypes) in addition to Operational Taxonomic Units (OTUs). By using this new capability, we here evaluated if the analysis of intraspecific mitochondrial diversity in addition to species diversity can provide insights into responses of stream macrozoobenthic communities to environmental stressors. For this purpose, we analysed macroinvertebrate bulk samples of three German river systems with different stressor levels using DNA metabarcoding. While OTU and haplotype number were negatively correlated with stressor impact, this association was not as clear when studying haplotype diversity across all taxa. However, stressor responses were found for sensitive EPT (Ephemeroptera, Plecoptera, Trichoptera) taxa and those exceedingly resistant to organic stress. An increase in haplotype number per OTU and haplotype diversity of sensitive taxa was observed with an increase in ecosystem quality and stability, while the opposite pattern was detected for pollution resistant taxa. However, this pattern was less prominent than expected based on the strong differences in stressor intensity between sites. To compare genetic diversity among communities in river systems, we focussed on OTUs, which were present in all systems. As OTU composition differed strongly between rivers, this led to the exclusion of a high number of OTUs, especially in diverse river systems of good quality, which potentially diminished the increase in intraspecific diversity. To better understand responses of intraspecific genetic diversity to environmental stressors, for example in river ecosystems, it would be important to increase OTU overlap between compared sites, e.g. by sampling a narrower stressor gradient, and to perform calibrated studies controlling for the number of individuals and their haplotypes. However, this pioneer study shows that the extraction of haplotypes from DNA metabarcoding datasets is a promising source of information to simultaneously assess intraspecific diversity changes in response to environmental impacts for a metacommunity.

Key Words
denoising, environmental impact, genetic erosion, haplotypes, pollution
Introduction

Degradation, pollution, and exploitation of freshwater ecosystems have resulted in a drastic decline of biodiversity (Vörösmarty et al. 2010; WWF 2018). The magnitude of biodiversity loss depends on stressor intensities as well as on resistance and resilience of biotic communities (Elmqvist et al. 2003; Dobson et al. 2006; Vörösmarty et al. 2010). So far, degradation and recovery processes have mostly been studied at the level of species diversity (alpha diversity). However, the underlying genetic diversity within species is an essential variable to consider in this context, as it determines the evolutionary capacity of a species to adapt to changing environments. A high level of intraspecific genetic variation is assumed to occur in intact and stable ecosystems, where effective population sizes are large and relatively constant over time. When comparing genetic diversity within the same species under stressor impact, the diversity is assumed to decline under stressor impact (‘genetic erosion hypothesis’) primarily due to reduced population sizes leading to enhanced genetic drift (Amos and Balmford 2001; van Straalen and Timmermans 2002; Reusch et al. 2005; Reynolds et al. 2012; Ribeiro and Lopes 2013). As the most basal level of biodiversity, genetic diversity within species is typically the first to decrease, and the last to regenerate, after stressor impact. It consequently provides a proxy for environmental impacts on communities long before, or even if never visible on species diversity level (Guttman 1994; Hughes et al. 2008; Reynolds et al. 2012). Such studies support the view that intraspecific genetic diversity and its distribution and turnover can be used as a measure for population and habitat stability. However, since the detailed assessment of intraspecific genetic variation is complex and elaborate, it is typically neglected at least in regulatory species assessment and monitoring programmes, or species diversity of a habitat is regarded as a proxy for intraspecific diversity (Vellend and Geber 2005; Vellend 2005; Laikre et al. 2020). The linkage between both diversity levels has been addressed for example by Basalgo et al. (2013) investigating water beetle communities across Europe and Taberlet et al. (2012), who showed that alpine plant species diversity is not correlated with intraspecific diversity.

Nowadays, alpha diversity can be assessed with great resolution using DNA metabarcoding (Hänfling et al. 2016; Deiner et al. 2016; Macher et al. 2018). With these data, stressor impacts can be analysed simultaneously for many taxa not distinguishable by morphological determination methods (Bagley et al. 2019; Beermann et al. 2018; Pfrender et al. 2010; Theissinger et al. 2019). With DNA metabarcoding, responses are typically inferred at Operational Taxonomic Unit (OTU) level. In most cases, distance-based thresholds are used to define OTUs with the aim that these reflect as closely as possible biological species, but other, more flexible approaches exist (Fujisawa and Barraclough 2013; Mahé et al. 2014). While OTU classification can drastically improve the taxonom-
in unrestored sewage channels. Heterogeneous conditions and stressor impacts are also present in the Ennepe, and include near-natural rural sites, urban sites stressed through occasional stormwater retention basin overflow, and sites with sewage treatment plant inflow. In comparison, the river Sieg is considered as a stable, near-natural river system with a good ecological and chemical status. Punctual stressor inflow is present through rainwater retention basins, but not immediately at sampling sites. By comparing communities between the different streams, we want to test if species and intraspecific diversity is correlated with the present stressor gradient. Following predictions from the ecological habitat concept, which links the presence and abundance of species over time to the available resources (see e.g. Van Dyck, 2012), we expect OTU diversity to be highest in river Sieg, moderate in river Ennepe, and lowest in river Emscher. According to the genetic erosion hypothesis, we predict haplotype diversity to covary with OTU richness. We expect highest haplotype number and diversity at the river Sieg due to its long-time stable good ecological conditions, supporting large and stable population sizes. Lower values are expected at the river Ennepe, where communities are regularly affected by organic stressor influx and thus recurrent population decline, resulting in higher genetic drift. The lowest values are expected for river Emscher due to the complete erasure of MZB diversity in the history of this river system caused by the usage as sewage transport system, and the still prevalent stress level in many parts. We predict different patterns for taxa sensitive (EPT—Ephemeroptera, Plecoptera, Trichoptera) or resistant (PR—‘Pollution Resistant’—Arhynchobdellida, Enchytraeida, Haplotaxida, Isopoda, Rhynchobdellida) to organic pollution. Specifically, we assume that OTU and haplotype diversity for EPT taxa will decline with increasing stress because of declining population sizes, and eventual local species extinction. In contrast, we expect that PR taxa will show opposing patterns because of their resistance to organic pollution and their ability to rather use organic pollutants as a resource, potentially facilitating large population sizes (Smith et al. 2007; Friberg et al. 2010; Ribeiro and Lopes 2013).

Material and methods

Sampling

Macroinvertebrates were sampled according to Water Framework Directive compliant protocol (Meier et al. 2006) at six sites in the rivers Emscher and Sieg in autumn 2016 and 2017, and spring 2017 and 2018 (Figure 1). In short, kick-net sampling of different habitats with 20 subsamples in the Sieg, and ten subsamples in the Emscher due to fewer available microhabitats, was executed. The seven sites at the Ennepe were sampled in autumn 2017 and spring 2017 and 2018 similarly with ten subsamples. Subsamples were pooled, large parts of substrate discarded (e.g. stones, leaves, small branches), and samples, including macrozoobenthic specimens and remaining substrate, were transferred to 1 l bottles filled up with ethanol. Approximately 1/3 of bottle volume was filled with the sample and 2/3 with 96 % technical ethanol. If volume of sampled material was too large, it was divided into multiple bottles. Samples were transported to the laboratory and old ethanol was replaced with new 96 % technical ethanol on the same day.

Laboratory protocols

Samples were examined under a binocular (Leica S6E) to separate individuals from substrate. Substrate was discarded and individuals were counted and separated into two size categories (size class A: ≤ 25 mm, size class B: ≥ 25 mm) (see Elbrecht et al. 2017b for the procedure). Individuals of the two size classes were dried in petri dishes overnight and homogenised to fine powder with an IKA Ultra Turrax Tube Disperser (BMT-20-S-M sterile tubes, full speed for 30 min). Two times half of a spatula (~20 mg) was transferred into two Eppendorf tubes, 600 µl TNES buffer and 15 µl Proteinase K (10 mg/ml) were added per tube, and incubated overnight at 36 °C shaking at 250 rpm in a Thermoshaker (ThermoMixer C, Eppendorf). A salt extraction protocol (after Sunnucks and Hales 1996, adjusted as in Weiss and Leese 2016) was used to isolate DNA from powder. After DNA extraction and subsequent RNA digestion (1 µl RNase A
Data analysis

Sequences were analysed with JAMP-0.67 (https://github.com/VascoElbrecht/JAMP) including demultiplexing of data, paired-end-merging, and primer trimming, following standard settings. For haplotype extraction only reads of expected fragment length (421 bp) were included. A strict quality filtering was applied (maximal expected error max_ee = 0.3) and reads with an abundance < 0.003 % in a sample were excluded from the dataset. The algorithm Unoise3 (Edgar 2016) implemented in JAMP, was used to denoise the dataset (alpha = 3) and to separate common haplotypes from chimeras and sequencing noise. A past experimental study on fish found Unoise3 to be particularly efficient for denoising (Tsuij et al. 2019). The denoising approach is based on the assumption that high abundant unique reads (centroids) are real sequences amplified from the biological template. Defined by distance (d), other unique sequences (neighbours) are grouped around these highly abundant sequences. Based on the Levenshtein distance and abundance (defined by α), neighbours showing a small difference and abundance compared to the centroid are predicted to be erroneous. Denoised reads were assigned to OTUs (clustered by 3 % distance) and the number of sequences per haplotype were determined in each sample. As a further filtering step, OTUs with an abundance below 0.01 % (OTU/min = 0.01) and haplotypes with an abundance below 0.003 % (minhaplotype = 0.003) in at least one sample were discarded (see Elbrecht et al. 2018 for detailed explanation). This step was included, to filter also low abundant unique sequences, which are not integrated in the filtering through alpha. Taxonomic assignment of haplotypes was conducted through a comparison with the database BOLD (Ratnasingham and Hebert 2007) with the programme BOLDigger (Buchner and Leese 2020). Haplotypes with similarity < 95 % to a deposited sequence in the database were excluded from further analysis to prevent incorrect assignments potentially leading to the assessment of erroneous diversity patterns. Read numbers per haplotype of technical PCR replicates were fused and the average was calculated. Further analyses were carried out with the average read number per haplotype. To assess haplotype richness per OTU, we used count data. However, in order to approximate also traditional population genetic measures, we calculated haplotype and nucleotide diversity per sample site and season with Arlequin 3.5 (Excoffier and Lischer 2010) using read depths as a proxy for haplotype abundance. Data were not normally distributed and therefore the non-parametric Kruskal-Wallis test was used to check for effects of river system on diversity variables. A post-hoc Dunn test (package dunn.test()), Diino 2017) was used to conduct pairwise comparisons for significant differences. All statistical analysis was conducted in R (R Development Core Team, 2008). Table modification and figure preparation were carried out using the packages vegan (Oksanen et al. 2019), tidyverse (Wickham et al. 2019) and ggplot2 (Wickham 2016) implemented in R.

Results

After denoising and abundance filtering, on average 29,063 reads were present in Emscher samples, 36,289 reads in Sieg samples, and 51,981 reads in Ennepe samples. Because filtering thresholds were based on relative abundances (see Material and methods, Data analysis), reads per sample were not adjusted to uniform numbers. Samples contained 228–694 haplotypes, which clustered into 70–155 OTUs. OTU and haplotype number was higher at river Ennepe and Sieg than at river Emscher (p < 0.01, Fig. 2). A high number of unique haplotypes per sample site was detected in all river systems (Suppl. material 7: Fig. S7). Splitting the dataset in EPT (Ephemeroptera, Plecoptera, Trichoptera) and PR (‘Pollution Resistant’) taxa revealed 13–273 haplotypes, clustered into 2–46 OTUs (zeros excluded, e.g. E5) for EPT taxa, and 8–257 haplotypes, clustered into 4–55 OTUs for PR taxa. As no plecopterans were found at river Emscher, only ET taxa could be analysed for this river. Total OTU and haplotype number of EPT taxa was affected by the river system with more OTUs and haplotypes at Ennepe and Sieg than at Emscher (p < 0.001). Sample sites E4 and E5 showed remarkably high OTU and haplotype numbers assigned to PR taxa compared to all other sample sites. However, no effect of the river system was detected on PR taxa (p = 0.09) (Fig. 2). The number of counted individuals before laboratory processing differed between all
river systems (p < 0.001) and seasons (p < 0.05) (Suppl. material 8: Table S1). Within streams, individual numbers did not significantly differ between sampling sites. However, no correlation was detected between total specimen number per site and season, and average haplotype number per OTU (Suppl. material 1: Fig. S1).

To compare average haplotype number per OTU and haplotype diversity between river systems, we searched for OTUs present in most samples. The five most common OTUs, all occurring at more than 50 % of the analysed samples, were: OTU 2 (Baetis rhodani, 63 %), OTU 9 (Asellus aquaticus, 52 %), OTU 12 (Stylodrilus heringianus, 63 %), OTU 65 (Esolus parallelepipedus, 55 %), and OTU 107 (Microtendipes pedellus, 52 %).

To further increase the number of shared OTUs between sites, samples collected at different seasons were merged (E1-E6, En1-En7, S1-S6). By this, we identified six OTUs occurring in more than 80 % of all sites (OTU 2: Baetis rhodani, 84 %; OTU 5: Orthocladius sp. A, 89 %, OTU 9: Asellus aquaticus, 84 %; OTU 45: Orthocladius sp. B, 84 %; OTU 67: Tanytarsus eminulus, 84 %, OTU 107: Microtendipes pedellus, 95 %). To increase the number of OTUs for analyses, all OTUs present in at least one of the samples per river system were included, resulting in four different datasets (Em-En-S: 78 shared OTUs, Em-En: 110 shared OTUs, Em-S: 125 shared OTUs, En-S: 155 shared OTUs). Per dataset > 47 % of shared OTUs were assigned to dipterans, of which the majority (> 90 %) were chironomids (Suppl. material 2: Fig. S2).

Comparisons of average haplotype number per OTU and haplotype diversity revealed no differences between river systems for all four datasets when all taxa were included (Suppl. material 3: Fig. S3). Dividing the datasets into OTUs assigned to EPT (pollution sensitive) and PR (pollution resistant) taxa revealed a significant effect of the river system on the average haplotype number per shared OTU when comparing all three river systems (EPT: p < 0.05, PR: p < 0.05) (Fig. 3). Ennepe and Sieg showed a higher average haplotype number per OTU for EPT taxa than the Emscher (En: 4; S: 3.3; Em: 2.7, p < 0.05), while the ratio for PR taxa was higher at Emscher (3.1) than at the other two rivers (En: 2.7, S: 2.6). When comparing only shared PR OTUs between Emscher and Sieg, more haplotypes per OTU were found at the Emscher (Em: 3.2; S: 2.4, p < 0.05). Observations on PR taxa also showed a higher haplotype diversity at the Emscher in comparison to both other streams (Em: 0.374, En: 0.303, S: 0.249). When comparing shared OTUs only between Emscher and Sieg, a significantly higher haplotype diversity of PR taxa was observed at the Emscher (Em: 0.3338; S: 0.1934, p < 0.05) (Fig. 3). Detailed information on average haplotype number per OTU, haplotype diversity and nucleotide diversity per sample site are illustrated in Suppl. material 4: Fig. S4 (average haplotype number per OTU), Suppl.
Figure 3. A) Average haplotype number per OTU for the four datasets of shared OTUs. Sensitive (EPT) and pollution resistant (‘PR’) taxa are shown. * indicates significant difference between regarded groups. B) Haplotype diversity for the four datasets of shared OTUs.

Discussion

We expected a general effect of stressors on alpha (OTU) diversity and intraspecific genetic diversity in the three river systems. In line with these expectations, the heavily impacted river Emscher showed the lowest OTU number compared with the other two systems. However, no significant differences between Ennepe and Sieg were detected, and an even higher number of highly pollution-sensitive stoneflies (Plecoptera) was found at the stronger impacted river Ennepe. Since strong stressor impact differences between river systems were expected, the lack of MZB community impacts between river Sieg and Ennepe seemed unexpected. However, further system or population specific factors, which were not specifically tested in the present study, could have influenced the observed pattern. These factors should be considered in the future (e.g. population history), also including individual numbers of investigated species. As expected, haplotype numbers per sample site were
lower in river Emscher than in Ennepe and Sieg, especially for pollution sensitive mayfly and caddisfly taxa. Vice versa, pollution resistant (‘PR’) taxa had higher haplotype numbers in the river Emscher. Genetic diversity estimates inferred via average haplotype number per OTU, haplotype diversity as well as nucleotide diversity revealed higher values at river Sieg and Ennepe compared to the Emscher, but again no differences between the former two rivers and therefore supported results based on OTU and haplotype number. In this case, the hypothesis that sustainable good ecological conditions and stability at river Sieg induced stable and large population size, favouring a high level of genetic diversity in sensitive MZB communities, was supported (Reynolds et al. 2012; Ribeiro and Lopes 2013). The Em-En-S dataset further underlines a strong correlation between average haplotype number per OTU and total number of ET taxa, further emphasising a linkage between species (OTU) diversity and intraspecific variability (Vellend and Geber 2005). Assuming a higher number of sensitive ET taxa in ecologically intact streams, the correlation also links local habitat conditions with genetic diversity, increasing from Emscher to Sieg and Ennepe. For the other datasets (Suppl. material 2: Fig. S2), no significant differences in intraspecific diversity was observed for E(P)T taxa. Beside actual biological signal, this is most likely due to the insufficiency of underlying datasets or methodological problems that will be discussed in the following paragraphs (from paragraph ‘OTU overlap’ on). Pollution resistant (PR) taxa showed a higher average haplotype number per OTU and haplotype diversity at the Emscher than at the other two systems when all three rivers were compared. This supports initial assumptions about low competition pressure and increased population growth of those taxa in stressed systems (Gaufin and Tarzwell 1952; Smith et al. 2007; Friberg et al. 2010). Detected patterns comparing all river systems are also supported by significant differences in genetic diversity between Emscher and Sieg and correlations between total number of PR taxa and intraspecific diversity.

The data presented, hold great potential to inform on metacommunity and metapopulation structure and processes. However, several limitations need to be considered when further testing and applying the approach:

**OTU overlap:** The fact, that no significant differences between average haplotype number per OTU and haplotype diversity were found for the other comparisons of E(P)T taxa (pairwise comparisons of all rivers), probably results from the small overlap of OTUs shared between rivers. The highly heterogeneous, site-specific stressor

---

**Figure 4.** Correlation analysis of total number of OTUs assigned to EPT/PR taxa and average haplotype (ht) number per OTU for those groups. A–D four datasets of shared OTUs for EPT taxa. E–H four datasets of shared OTUs for PR taxa.
levels and variable time since restoration at Emscher further inflated variation in OTU composition, thereby limiting statistical power for comparisons of genetic variation based on highly frequent OTUs. We decided to reduce our dataset to OTUs at least present in one sample site of compared river systems to circumvent differences in infraspecific genetic variation due to species specific traits and sensitivities (e.g. Stavrakakis et al. 1999; Machet et al. 2016). However, this approach excludes taxa unique to specific sites, limiting the analysis of metacommunities at highly diverse stream ecosystems (Sieg, Ennepel), and also limiting the number of metapopulations to study with respect to mitochondrial genetic diversity. An alternative approach would be to use the global infraspecific variation at higher taxonomic levels (e.g. genera, families). This would maximise the number of OTUs included, yet, such analyses based on higher taxonomic level, would have led to comparisons of parameters inferred from very different species for the different river systems, making it impossible to disentangle stressor impacts on intraspecific diversity from species-specific patterns.

Mitochondrial single-gene marker: Beside the limitations due to a small overlap in OTUs between river systems, the utility of the mitochondrial COI marker as a measure of infraspecific genetic diversity and variability could have deflated signal strength. Even if various studies have successfully applied this marker for even more specific population analyses, the use of only a single gene can be misleading – and mitochondrial genes are especially unique due to their haploid structure, the lack of recombination and purely maternal inheritance (Ballard and Whitlock 2004; Leese and Held 2011).

Bioinformatic haplotype extraction: Furthermore, as outlined in previous studies (Elbrecht et al. 2018; Tsuji et al. 2019; Turon et al. 2019), the main challenge in extracting haplotypes from metabarcoding datasets is the separation of ‘real’ environmental sequences from those produced by PCR or sequencing errors. New programmes enable the denoising of datasets with the optimisation of filtering steps, to efficiently separate real sequences from erroneous ones. However, a decision has to be made concerning the strictness of filtering. By using high filtering thresholds, erroneous sequences are excluded with a higher probability, but at the same time it is more likely to exclude real sequences of low abundance. In comparison, a lower filtering increases the number of rare real sequences, but also includes a higher number of erroneous sequences into diversity analysis. For the present study, we followed denoising as recommended in Elbrecht et al. 2018, where the genetic variability of benthic macroinvertebrates in Finland’s streams was investigated, implementing an α-value of 5, which was also applied in Turon et al. 2019. The additional percentual abundance threshold filtering for OTUs and haplotypes was set after suggestions in Elbrecht et al. 2018 and applied in Laini et al., in review. We found a high number of unique ESVs per sample site (exemplary networks of the two most frequently found EPT and PR taxa shown in Suppl. material 7: Fig. S7), which is similar to other metabarcoding studies (Laini et al. in review; Elbrecht et al. 2018), but exceeds those found in studies based on single specimen barcoding (Williams et al. 2006; Lucentini et al. 2011; Weiss and Leese 2016). Higher numbers of unique ESVs could be induced into the dataset through sequencing errors, which would emphasise the application of an even higher filtering threshold on metabarcoding datasets. However, the increased number of ESVs found through metabarcoding can also be real haplotype variants in one specimen, e.g. due to somatic mutations, which cannot be determined through single specimen barcoding due to the underlying sequencing method (Elbrecht et al. 2018; Tsuji et al. 2019). Evidence of erroneous ESVs due to pseudogenes was considered negligible in our dataset (e.g. 1.4 % of EPT taxa ESVs and 0.4 % of PR taxa ESVs showed stop codons). The additional filtering step, i.e. discarding all ESVs assigned to a deposited sequence in BOLD with <95 %, was applied to further exclude potentially erroneous sequences. For indicator EPT taxa this is reasonable since reference databases are very comprehensive (~89 %, Weigand et al. 2019). However, a bias might have been introduced through this approach for PR taxa, since reference databases are less complete for these taxa, potentially excluding ESVs that constitute an important biological signal. It should be noted, however, that of the excluded 75 PR taxa ESVs, none was considered for the detailed analysis of infraspecific variation due to a lack of OTU overlap between sites. In future studies, implementing a higher overlap of taxa between compared sites should re-evaluate this filtering threshold.

Individual sample size: Lastly, while the strength of the approach is that thousands of specimens were processed at once, there is little control about the individual number of specimens per species or OTU. Comparisons of genetic diversity rely on the number of specimens sampled and future studies should carefully control for this in order to test the reliability and robustness of the approach.

Conclusion

Using macrozoobenthic taxa from three differently impacted German river systems, our study shows that denoised metabarcoding data can provide valuable information to assess effects of environmental variables on infraspecific genetic diversity. Even if the choice of data filtering thresholds is still a trade-off between rare ‘real’ sequences and erroneous, artificially generated ones, we were able to link stressor effects to changes of infraspecific genetic variation in aquatic macroinvertebrate communities. Sites with good and stable ecological conditions showed higher infraspecific diversity than stressed sites, which is also coupled with higher OTU diversity. However, due to a low OTU overlap between river systems, genetic diversity analyses were based only on subsets, including all shared OTUs. This subsampling induced the exclusion of variability and ecological specialists,
especially at highly diverse sample sites and might have skewed actual differences. Due to these limitations in the underlying data, we cannot disentangle effects of stressors from e.g. population and colonisation and dynamics. Future studies need to address these aspects in more detail and should include several replicates of similar conditions, presuppose a threshold of overlapping OTUs in compared river systems and control for the number of individuals per OTU to allow for a statistically accurate comparison. However, in conclusion, our study adds to the growing number of studies that highlight the potential to extract haplotype information from metabarcoding datasets for more holistic biodiversity assessments.

Acknowledgements

This study was supported by the German Federal Ministry of Education and Research project “German Barcode of Life II” (GBOLII) to FL (FKZ 01II11101 and 01II1501). We want to thank the Emschergenossenschaft/Lippe-Verband and Emepe-Ruhr-Kreis for support during this study and Bianca Peinert for her help in the field and the laboratory. FL and VMAZ are supported by the COST Action CA15219 DNAqua-Net.

References

Buchner D, Leese F (2020) BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. MBMG 4: e53535. https://doi.org/10.3897/mbmg.4.53535


Williams HC, Ormerod SJ, Bruford MW (2006) Molecular systematics and phylogeography of the cryptic species complex Baetis rhodani
Metabarcoding and Metagenomics 4: e51925


Supplementary material 1
Figure S1. Total number of aquatic macroinvertebrate individuals per sample and season plotted against the average haplotype number per OTU. Different colours indicate the three river systems
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: genetic
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl1

Supplementary material 2
Figure S2. Four different datasets including shared OTUs between the different river systems (Emscher-Ennepe-Sieg, Emscher-Ennepe, Emscher-Sieg, Sieg-Ennepe). Number of OTUs is illustrated with taxonomic assignment on order level
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: genetic, occurrence
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl2

Supplementary material 3
Figure S3. Average haplotype number per OTU for the four different datasets of shared OTUs. Values are illustrated for all sample sites including all shared OTUs
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: genetic
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl3

Supplementary material 4
Figure S4. Average haplotype number per OTU for the four different datasets of shared OTUs. Datasets are split into EPT (Ephemeroptera, Plecoptera, Trichoptera) and PR (‘Pollution Resistant’) taxa
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: genetic
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl4

Supplementary material 5
Figure S5. Average haplotype diversity for all four datasets of shared OTUs separated according to sample sites and EPT (Ephemeroptera, Plecoptera, Trichoptera) and PR (‘Pollution Resistant’) taxa
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: (measurement/occurrence/multimedia/etc.)
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl5

Supplementary material 6
Figure S6. Average nucleotide diversity for all four datasets of shared OTUs separated according to sample sites and EPT (Ephemeroptera, Plecoptera, Trichoptera) and PR (‘Pollution Resistant’) taxa
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: genetic
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl6

Supplementary material 7
Figure S7 – part 1. Haplotype network of the two most frequent EPT (Ephemeroptera, Plecoptera, Trichoptera) and PR (‘Pollution Resistant’) taxa
Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese
Data type: genetic
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.4.51925.suppl7
Supplementary material 8

Table S1. Number of macroinvertebrate individuals per sample and season

Authors: Vera Marie Alida Zizka, Martina Weiss, Florian Leese

Data type: individual counts

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.4.51925.suppl8