Title: Sorting things out - assessing effects of unequal specimen biomass on DNA metabarcoding

Running Title: Influence of specimen biomass in metabarcoding

Authors: Vasco Elbrecht[1*], Bianca Peinert[1], Florian Leese [1,2]

Affiliations:

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

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

*Corresponding author: vasco.elbrecht@uni-due.de, phone: +49.201-1834053

Abstract

Environmental bulk samples often contain many taxa that vary several orders of magnitude in biomass. This can be problematic in DNA metabarcoding and metagenomic high-throughput sequencing approaches, as large specimens contribute disproportionately high amounts of DNA template. Thus, a few specimens of high biomass will dominate the dataset, potentially leading to smaller specimens remaining undetected. Sorting of samples by specimen size and balancing the amounts of tissue used per size fraction should improve detection rates, but this approach has not been systematically tested.

Here we explored the effects of size sorting on taxa detection using two freshwater macroinvertebrate monitoring samples, collected from a low-mountain stream in Germany. Specimens were morphologically identified and sorted into three size classes (body size < 2.5x5, 5x10 and up to 10x20 mm). Tissue from each size category was extracted individually, and pooled to simulate samples that were not sorted by biomass ("Unsorted"). Additionally, size fractions were pooled so that each specimen contributed approximately equal amounts of biomass ("Sorted"). Mock samples were amplified using four different DNA metabarcoding primer sets targeting the Cytochrome c oxidase I (COI) gene.

Sorting taxa by size and pooling them proportionately according to their abundance lead to a more equal amplification of taxa compared to the processing of complete samples without sorting. The sorted samples recovered 30% more taxa than the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

unsorted samples, at the same sequencing depth. Our results imply that sequencing depth can be decreased approximately five-fold when sorting the samples into three size classes and pooling by specimen abundance.

Our study demonstrates that even a coarse size sorting can substantially improve taxa detection using DNA metabarcoding. While high throughput sequencing will become more accessible and cheaper within the next years, sorting bulk samples by specimen biomass is a simple yet efficient method to reduce current sequencing costs.

Keywords: Biomass bias, specimen sorting, metabarcoding, metagenomics, DNA barcoding, ecosystem assessment

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

1) Introduction

Recent advancements in high-throughput sequencing (HTS) and DNA barcoding have improved our ability to rapidly assess biodiversity. By using traps or manual collection methods (e.g. nets), thousands of specimens can be easily collected. However, manually identifying hundreds or thousands of specimens in a single sample is often not feasible, especially if species level identification is needed (Haase et al. 2004). Bulk samples, which previously took weeks or months to determine morphologically, can now be homogenised and their DNA extracted for sequencing based identification within days. The power, accuracy and cost effectiveness of these HTS based assessments have already been demonstrated (e.g. Ji et al. 2013; Tang et al. 2015; Gómez-Rodríguez et al. 2015; Leray & Knowlton 2015; Gibson et al. 2014; Hajibabaei et al. 2011; Zimmermann et al. 2014; Dowle et al. 2015; Elbrecht et al. 2017) and sequencing costs are expected to further decline in the future.

In DNA based ecosystem assessment we can distinguish between two approaches: 1) a target gene fragment is amplified and compared to a DNA barcoding database (metabarcoding, see Taberlet et al. 2012) or 2) the extracted DNA from the bulk sample is shotgun sequenced directly without PCR and can be optionally enriched for target genes (metagenomics, see Liu et al. 2016; Crampton-Platt et al. 2016). Both approaches have specific advantages and drawbacks: metabarcoding is severely limited by PCR bias, preventing estimates of taxa biomass and potentially not detecting all taxa present in the sample (Elbrecht & Leese 2015; Piñol et al. 2014; Leray & Knowlton 2015). While metagenomics might overcome these PCR based problems, this approach is currently limited because only little reference data is available (e.g. mitochondrial genomes) and a high sequencing depth is required (Crampton-Platt *et al.* 2016). Additionally, both approaches are likely affected by variable cell densities and types, as well as variable mitochondrial genome copy numbers between taxa and specimen life stages (Ballard & Whitlock 2004; Moraes 2001), which is potentially affecting taxa detection. While these problems might be solved at least partially by optimised degenerate primers (Elbrecht & Leese 2017), reduced sequencing costs and mitogenome capture (Tang *et al.* 2014), both metabarcoding and metagenomics are potentially affected by an additional factor: variable taxa biomass.

Environmental samples usually contain a diverse set of taxa spanning often several orders of magnitude in specimen sizes and biomass. When extracting complete bulk samples, large biomass rich specimens will contribute significantly more DNA to the final bulk DNA isolate than small organisms with little biomass. We demonstrated this previously, by bulk extracting DNA from 31 specimens of the same stonefly (Plecoptera) species with varying specimen biomass, and found a clear significant linear correlation between obtained reads and dry specimen weight ($p<0.001$, $R^2 = 0.65$, (Elbrecht & Leese 2015). We hypothesise that also in more species rich samples, taxa biomass translates directly into read abundance (assuming no

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

primer bias within species Elbrecht & Leese 2015). Thus just a few big specimens in a sample will likely make up the majority of the reads, requiring higher sequencing depth to also detect small specimens and rare taxa. The effects of large specimens might be also further amplified by primer bias increasing or decreasing the number of reads obtained for a taxon (Elbrecht & Leese 2015; Piñol et al. 2014). Some studies have already applied samples sorting into different size fractions, for DNA metabarcoding because of this biomass introduced bias (Leray & Knowlton 2015; Wangensteen & Turon 2016). However, the effect of fractioning samples by specimen biomass against the complete sample without pre-sorting of specimens has not jet been systematically tested and quantified. Morinière *et al.* 2016 detected additional taxa when sorting malaise trap samples by insect orders (which however could have been caused by unequal sequencing depth between the samples). The authors further encourage to also test the effects of fractioning samples by specimen biomass.

In this study we systematically quantified the effects of biomass sorting on taxon recovery using two complete stream macroinvertebrate kick samples (mostly larval specimens), morphologically identifying and sorting them into three biomass categories based on specimen sizes: **s**mall (S), **m**edium (M) and **l**arge (L), see Figure 1A & S1. These size fractions were used to generate mock samples to compare the effect of extracting all specimens together without sorting ("**Un**sorted") against pooling the sorted samples according to number specimens in each sample ("**So**rted"), to archive a more equal representation of small specimens in the extracted sample (Figure 1). While it is difficult to accurately pool the ground tissue of each size category (Figure 1B), pooling extracted DNA might be potentially biased by variable cell sizes and mitochondrial copy numbers in different taxa (Figure 1D, Bendich 1987; Lemire 2005). Thus we decided to pool the DNA extraction buffer after tissue digestion of S, M and L for mock sample generation, as the lysis buffer has the same DNA proportions as the ground tissue but can be more precisely pooled (by pipetting, see Figure 1C). Additionally, DNA from each size category was extracted and sequenced individually, to estimate which taxa are present in each and are thus expected to be also detected in the mock samples. By metabarcoding these individual size fractions as well unsorted and sorted samples mock samples, we can precisely investigate the effects of sample sorting by specimen size on taxa recovery.

2) Material and Methods

Figure S2 gives an overview of how samples were collected, extracted, pooled into mock communities and metabarcoded as will be discussed in the following.

Sample collection and processing

Macroinvertebrates were collected at two sampling points of the small low-mountain range stream Kleine Schmalenau in Germany (Arnsberger Wald). The main stream (site P8, N51.43623 E8.13721) and a small tributary (site P10, N51.43295

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

E8.14350) were sampled with five kick samples per sampling site (0.45 m^2 area) following the general principle of the multi-habitat sampling protocol also used in the German implementation of the EU Water Framework Directive (Meier *et al.* 2006). Collected specimens were stored in 96% ethanol at -20°C for later molecular analysis. All invertebrates were counted and identified morphologically to the lowest taxonomic level that could be accurately and consistently determined given the available literature, larval life stage and specimen condition (Table S1).

Specimens from the two samples were each sorted into three size categories under a Zeiss Stemi 2000 stereo microscope by placing them onto millimetre paper (Figure S1, C). Specimens below 2.5x5 mm body size (length x height, excluding thin extremities and appendices) were sorted into small (S) specimens up to 5x10 mm into medium (M) and everything bigger than that into large specimens (L, max 10x20 mm, see Figure S1, C). For thin but long specimens like e.g. chironomids (non-biting midges), the total surface was considered and evaluated if it would fit into the surface of the respective rectangle (e.g. all chironomids were sorted into the small size category despite being some times longer than 5 mm). Antennae and cerci were not counted in the measurement of body length. Goal of the sorting by specimen size, was to visually separate the specimens into size categories as a proxy for biomass that is difficult to measure on ethanol wet specimens (see Figure S1). Terrestrial taxa and Trichoptera (caddisfly) quivers were included in the samples, as it is not realistic to remove non-target organisms or empty shells and quivers in routine monitoring samples.

DNA extraction and tissue pooling

Specimens of each size category were dried overnight at room temperature in sterile Petri dishes to remove the ethanol. Total dry specimen weight in each size category was measured (in duplicates) on a Sartorius RC 210D scale. Specimens from each category were homogenised (Figure 2D) using an IKA ULTRA-TURRAX Tube Drive control system with sterile 20 ml tubes and 10 steel beads (5 mm Ø) by grinding at 4000 rpm for 30 minutes (IKA, Staufen im Breisgau, Gemrany). In this study we wanted to compare the taxa recovery between samples sorted by specimen size and then proportionally pooled by specimen abundance (So) against unsorted, i.e. complete samples (Un). Thus five different DNA extractions were prepared for each of the two sampling sites (Figure 2). First of all, DNA from each size category (S, M and L) was separately digested using a standard salt extraction protocol (Sunnucks & Hales 1996) (see Figure S3). Seven tissue aliquots were digested and united per size category (Figure 2F), to obtain sufficient amounts of digested tissue for pooling (Figure 2G). Then three aliquots of digested tissue were then used to generate the sorted and unsorted mock samples. Tissue digested in DNA extraction buffer was used, as it can be precisely pooled in specific proportions (unlike ground tissue), while not introducing biases based on variation in cell density and mitochondrial copy numbers which possibly affect

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

extracted DNA (Figure 1). However, the amount of tissue used in digestion of S, M and L samples was not always similar (Figure 2E), which has to be accounted for when pooling the digested tissue for mock community generation (Figure 2I). This however was mistakenly not done for the sorted samples (Figure 2H), where digested tissue was pooled based on number of specimens in each size category to reduce the influence of large specimens in the extraction. This mock sample was compared with an unsorted sample pooled based on specimen weight (Figure 2J) that retains the original tissue proportions in the sample, i.e. bulk extraction of the complete sample. Additionally, all S, M and L aliquots were extracted separately and used as individual metabarcoding samples, to be included as positive controls (Figure 2). All extractions from the digested tissue were done in triplicates and united into one single aliquot, to increase the amount of DNA available for each sample.

45 µl DNA from each sample (S, M, L, Un, So for sampling site P8 and P10) was digested with 1 µl RNAse A (10 mg/mL, Thermo Fisher Scientific, MA, USA) and cleaned up using a MinElute Reaction Cleanup Kit (Qiagen, Venlo, Netherlands) with resuspension in ddH_2O. DNA concentrations were quantified fluorometrically using a Qubit (HS Kit, Thermo Fisher Scientific, Waltham, MA, USA) and concentrations adjusted to 25 ng/µl.

DNA metabarcoding and bioinformatics

All 10 samples (S, M, L, Un, So for sampling site P8 and P10) were amplified with the four freshwater macroinvertebrate fusion primer sets BF/BR (Elbrecht & Leese 2017). The four primer combinations are targeting a 217 to 421 bp long fragment of the Cytochrome c oxidase I (COI) gene. Figure S4 gives an overview of sample tagging using fusion primers with inline barcodes. Each PCR reaction was composed of 1× PCR buffer (including 2.5 mM Mg^{2+}), 0.2 mM dNTPs (Thermo Fisher Scientific, MA, USA), 0.5 µM of each primer (Biomers, Ulm, Germany), 0.025 U/µL of HotMaster Taq (5Prime, Gaithersburg, MD, USA), 0.5 mg/µl molecular grade BSA (NEB, MA, USA), 12.5 ng DNA, filled up with HPLC H_2O to a total volume of 250 µL. Each 250 µL PCR reaction mix was divided into five wells before PCR. PCR reactions were run in a Biometra TAdvanced Thermocycler using the following program 94°C for 3 min, 40 cycles of 94°C for 30 s, 50°C for 30 s, and 65°C for 2 min, and 65°C for 5 min. High reaction volume and BSA was necessary due to PCR inhibitors present in the samples. PCR products were purified and size selected (left sided) using SPRIselect with a ratio of 0.8x (Beckman Coulter, CA, USA) and quantified with a Qubit fluorometer (HS Kit, Thermo Fisher Scientific, MA, USA). Samples were pooled to equal molarity, and the final library purified with the MinElute Reaction Cleanup Kit (Qiagen, NL), as a precaution because the BSA used in the PCR caused adhesion of magnets to the tube walls in the PCR clean-up with SPRIselect. Paired-end sequencing was done on one lane of an Illumina HiSeq 2500 system with a rapid run 250 bp PE v2

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

sequencing kit and 5% PhiX spike-in. However, sequences contained ambiguous bases at two positions, due to air bubbles in the flow cell (SRR3399055). Thus the run was repeated, this time loading two lanes with the same library in slightly different cluster density, again with a 5% PhiX spike-in.

We used the UPARSE pipeline in combination with custom R scripts (Dryad COI - Not jet available!) for data processing (Edgar 2013, Figure S5). Reads from both lanes were demultiplexed with a R script and paired end reads merged using Usearch v8.1.1861 `-fastq_mergepairs` with `-fastq_maxdiffs` and `-fastq_maxdiffpct 99` (Edgar & Flyvbjerg 2015). Primers were removed with Cutadapt version 1.9 on default settings (Martin 2011). Sequences were trimmed to the same 217 bp region amplified by the BF1+BR1 primer set and the reverse complement build if necessary using `fastx_truncate / fastx_revcomp`. Only sequences with 207 - 227 bp were length used in further analysis (filtered with Cutadapt). Low quality sequences were then filtered from all samples using `fastq_filter` with `maxee = 1`. Sequences from all samples were then pooled, dereplicated (`minuniquesize = 3`) and then clustered into operational taxonomic units (OTUs) using `cluster_otus` with 97% identity (Edgar 2013) (includes chimera removal). A threshold of 97% was used to reduce the effect of sequencing errors, which might lead to the generation of additional "false" OTUs.

Pre-processed reads (Figure S5, step B) of all samples were dereplicated again using `derep_fulllength`, but singletons were included. Sequences of each sample were matched against the OTUs with a minimum match of 97% using `usearch_global`. As the sample library was loaded on both lanes, hit tables from both HiSeq lanes were combined, because they only represent sequencing replicates. Only OTUs with a read abundance above 0.01% in at least one sample were considered in downstream analysis. Within each sample, OTUs with less or equal than 0.01% were set to 0% sequence abundance to reduce the number of false positive OTUs. Taxonomy was assigned to the remaining OTUs using an R script searching the BOLD and NCBI database independently. Conflicting taxonomy was resolved on a case-by-case basis (with falling back to a coarser taxonomic level if the correct assignment was no evident). Only OTUs reliably identified as freshwater macroinvertebrates were included in the main analysis.

3) Results

Weight measurements of the tissue was done twice independently, with consistent results between replicates (SD = 0.083 mg). The library was sequenced on a HiSeq rapid run with a cluster density of 438 k/mm^2 and 542 k/mm^2 for lane 1 and 2 (raw data available on the NCBI RSA archive: SRR3399056 and SRR3399057). On average 1.71 (SD = 0.29, lane 1) and 2.17 (SD = 0.38, lane 2) million read pairs were obtained for each sample after demultiplexing (Figure S4). Read quality

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

varied with amplicon length and cluster density (Figure S4), but did not affect results strongly as OTU abundance was very similar between both lanes (= sequencing replicates of identical library). However, stochastic effects between both lanes increased for OTUs with low read abundance (Figure S6, variability between replicates for abundant OTUs >10%, SD = 0.007, OTUs with 0.1-0.01>% abundance, SD = 0.077).

The OTU raw data are available in Table S2 and morphology based identifications and taxa abundances in Table S3. After clustering and discarding low abundance OTUs, a total of 314 OTUs remained in the data set (Figure S7, Table S2). 71% of these OTUs could be reliably identified with available reference databases, with 58% of the OTUs belonging to freshwater macroinvertebrate taxa (Figure S8). All high abundance OTUs (at least 0.1% of reads) were identified as macroinvertebrate. Of these taxa 45 of 52 were reliably identified at species level, of which about 3/4 had 100% similarity matches to reference sequences. Low abundance OTUs (<0.1%) often showed poor matches to data bases or could not be identified at all (see Figure S8). With DNA metabarcoding over twice as many macroinvertebrate taxa and five times more species were detected than with morphology based identification (Figure S9). The main stream (P8) and tributary (P10) could be clearly distinguished, with only 14.3% of OTUs shared between both sites (Figure S7, 36.4% similarity based on morphological identification, Table S1).

Sorting the sample into three size categories and proportional pooling of DNA extracts by amount of specimens in each category reduced the dominance of large specimens substantially (Figure 3). The sorted samples (So) resembled the composition of the original sample much better (average difference to original composition = 2.3-fold, SD = 2.49) than the unsorted samples (Un, average difference = 9.0-fold, SD = 7.88, Figure 3). Using the four primer sets an average of 88.75 (SD = 6.46) invertebrate taxa were detected in the sorted samples, compared against 62.5 (SD = 4.5) in the unsorted samples (30% less, paired t-test, $p = 0.005$, Figure 4, no rarefaction applied). By using the S, M and L samples as controls, we could estimate the expected (E) amount of taxa we should be detecting with each primer pair (Figure S10). In sorted samples (So) very similar amounts of taxa as in the controls (E) were detected (paired t-test, $p = 0.17$). However, on average only 80% (SD = 8%) of the expected number of taxa were detected when the complete sample was extracted without sorting (Figure S10 A, paired t-test, $p < 0.001$). The same trend was observed when looking at Shannon Diversity (Figure S10 B, paired t-test, E vs So; $p = 0.9153$, E vs Un; $p < 0.001$). When comparing the taxa detected with metabarcoding against the taxa list based on morphological identification, again the unsorted samples showed decreased detection rates (67%, SD = 3%, paired t-test, $p = 0.006$). However, also with sorting, only 74% (SD = 3%) of the morphologically identified taxa were detected with each primer set, which however was not significantly different than the detection rates in the controls E (paired t-test, $p = 0.23$, Figure S10 C). Six morphologically identified taxa were not detected in our metabarcoding dataset (Figure S7, Table S3, morphologically determined specimens "Plecoptera" and "Insecta" are counted as detected here, as several insect and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

Plecoptera OTUs were detected in the dataset). The reduced amount of taxa detected with the unsorted samples, persists when the sequencing depth is reduced (Figure 4). Sample sorting does reduce the required sequencing depth to detect the same amount of taxa by ~5 times, compared to the unsorted samples.

4) Discussion

4.1) Effects of sorting metabarcoding samples by specimen size

We sorted two samples by specimen size (resembling biomass) into small, medium and large specimens and pooled them proportionately by specimen abundance per size class to compare these results against unsorted samples. Our results demonstrate that read abundances of the unsorted samples were dominated by few biomass rich taxa that contribute the majority of DNA in the bulk extraction. This does not only skew the read abundances in favor of biomass rich specimens, but also some smaller and less abundant taxa remained undetected (on average ~30% fewer taxa detected in the unsorted samples). The sorted samples only need 1/5 of the sequencing depth, to detect the same amount of taxa as in the unsorted samples. This means that sorting metabarcoding bulk samples by specimen size can substantially reduce sequencing costs, if the aim is to detect all taxa present in the sample regardless of biomass. While we only manually sorted our samples into three size categories, further cost reductions might be possible by sorting samples into more size categories. It is likely that larger specimens will have similar effects on metagenomic bulk samples, thus sorting by specimen size and correcting for abundance might also likely be viable for these samples.

Based on basic physical laws it is expected that large specimens are overrepresented in a metabarcoding study when extracted in bulk together with smaller organisms. Thus, it is no real surprise that sorting by specimen size but pooling proportionally by number of specimens in each size category lead to a more equal representation of specimens in the sample and increase the detection of rare and small specimens with DNA metabarcoding. However, also the limitations and shortcomings of this study should be discussed here. While we took great care to reduce factors biasing ours result, e.g. by extracting all samples from the same digested tissue aliquots, we failed to adjust for the amount of tissue digested in these aliquots for the sorted mock samples (Figure 2, H and I). This leads to a slight underrepresentation of small taxa in the mock samples, as for medium and large taxa more tissue was extracted (Figure 2E). While this will not change the overall effects found in this study, it does mean that the positive effects of sample sorting are even slightly underestimated here: With the correct (higher amount) of small specimens used in the mock communities maybe even more taxa could have been detected in the sorted samples. Additionally, this study was only carried out on two sampling sites and with limited morphological identifications. With more time spent and higher taxonomic expertise, probably more taxa could have been identified

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

morphologically. Also, despite the COI reference databases being fairly complete for common macroinvertebrate taxa (Figure 8), there are still gaps and potentially unreliable reference barcodes present, potentially also underestimating diversity. We also show that with our dataset that stochastic effects during Illumina sequencing affect mainly low abundant OTUs, which was recently also confirmed in other studies (Leray & Knowlton 2017). For a full and more detailed discussion of effects limitations of DNA metabarcoding for routine macroinvertebrate monitoring see (Elbrecht et al. 2017). Nevertheless, DNA based identifications can be more accurate than classical morphology based identification (Stein et al. 2013; Sweeney et al. 2011) as we also show with our two kick samples in this project.

4.2) Implications: Not all samples have to be sorted

While we could demonstrate and also quantify the increased resolution and potential cost savings by size sorting metabarcoding bulk samples, we have to acknowledge that these sample sorting steps can be time consuming and potentially also a source of cross contamination between samples. Thus, we do not recommend sorting every sample by specimen biomass right away. First of all, the sample should have specimens varying several magnitudes in biomass. If all specimens have similar sizes, sorting will likely not improve sequencing results. Additionally, the number of samples which can be reliably tagged on a HTS run in combination with the expected sequencing output, might make sorting obsolete if expected sequencing depth per sample is sufficiently high. However, in many cases where bulk samples show substantial variation in biomass, sequencing depth should be sufficiently high to also detect small and rare taxa. Here sorting samples and adjusting for specimen biomass can help to increase the number of taxa detected making it possible to pool more samples on the same sequencing run.

Whether or not the method of size sorting should be used in a study depends on sample composition and characteristics as discussed above, but more importantly if it is necessary to detect small and rare taxa present in the study (e.g. for non-targeted early detection of pests, invasive species or to build barcoding references). It has to be stressed that for most studies, the proportion of the abundant taxa is most relevant, which gets distorted by sample sorting and pooling by abundance of small, medium and large specimens. If samples just contain a few large specimens and abundance data is not that important, one could obtain a small piece of tissue (e.g. a leg of an invertebrate) and remove the rest of the specimen from the sample (as done by Ji *et al.* 2013 for example). Especially, if only presence-absence data is desired, this is a good trade off to reduce the negative influence of a few large specimens on the dataset, without sorting the complete sample. However, treating samples to reduce the influence of biomass rich specimens should be done systematically across samples to not introduce processing biases. In this study, sorted individual specimens into three size categories under a stereo microscope to get very

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

accurate size classes needed to test this method. With approximately 2-3 hours for each sample and the additional workload for DNA extraction, this is a highly time consuming step, making the technique of size sorting samples impractical for large sample quantities. Studies on marine invertebrate did size sort samples by sieving the samples with different sieve sizes from 63 µm to 10 mm (Leray & Knowlton 2015; Wangensteen & Turon 2016). Sieving is probably the only feasible method for processing large numbers of samples, but good care has to be taken when cleaning the sieves between samples, to prevent cross contamination. Sieving might also change the community composition as very small bacteria on surfaces and small organism might get lost, and broken off body parts (e.g. legs, antennas) or tissue parts from prey animals might end up in the lowest size fraction (Leray & Knowlton 2015; Wangensteen & Turon 2016). These effects have to be taken into consideration when looking at each size fraction individually. However, if the goal is to obtain a presence-absence taxa list for a complete sample, sieving and proportional pooling might be an ideal solution to minimize bias introduced by large specimens in the samples. Using dry specimen weight for each size fraction can be used to roughly estimate the number of taxa in each size fraction, which can then be used to pool the DNA proportionately, instead of sequencing each size fraction individually.

4.3) Conclusions

We demonstrated that sorting metabarcoding samples into three specimen size categories and then pooling the tissue fractions proportionally to the number of specimens in each size class, can reduce the amount of required sequencing depth compared to the unsorted sample by 80%. Sample sorting leads to a more balanced taxa assessment, dramatically reducing the overrepresentation of large specimens on the dataset. While size sorting of bulk samples might not be necessary or suitable for all samples, ecosystems or research questions, we encourage to evaluate if sample fractioning could be beneficial and feasible in metabarcoding projects. Also, some metagenomic projects will likely profit from presorting samples by biomass, but we did not explicitly test this here.

Acknowledgements: FL and VE are supported by a grant of the Kurt Eberhard Bode foundation to FL. We thank Volodymyr Pushkar and Janis Neumann for help with the field work, Arne Beermann for taxonomic validation and Ralph Tollrian for support within the project. We would further like to thank Brian Gill, Jan Macher, Edith Vamos, Vera Zizka as well as the leeselab journal club for proofreading this manuscript as well as two reviewers for helpful comments.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

298 **Declaration of interest:** The authors report no conflicts of interest. The authors alone are responsible for the content and
299 writing of the paper.
300
301 **Author Contributions statement:** VE, BP and FL conceived the ideas and designed methodology; BP collected and
302 identified specimens and carried out the laboratory work; VE performed bioinformatic analyses and wrote the manuscript.
303 All authors contributed critically to the drafts and gave final approval for publication.
304
305

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

Figures

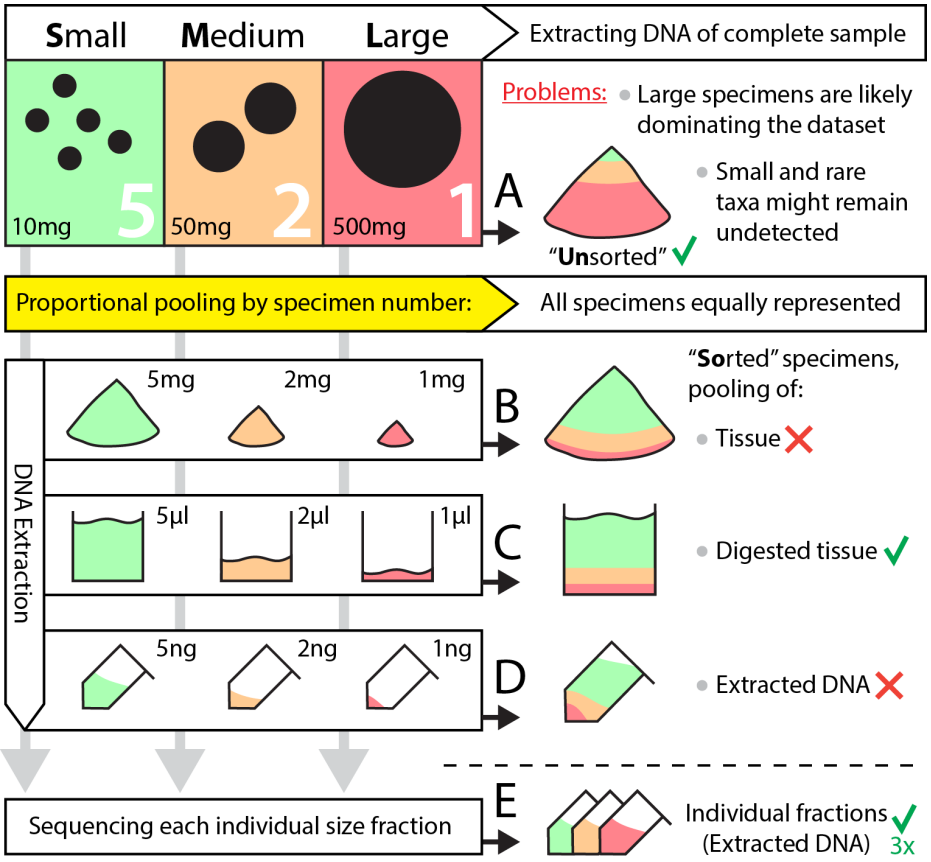


Figure 1: Overview of different strategies to reduce the presence of biomass rich specimens when metabarcoding bulk samples. Aliquots with a green checkmark (✓) were generated and metabarcoded in this study, while those with a red "X" were not tested. Large specimens (L) have substantially more biomass than small specimens (S) and thus contribute more DNA when extracting complete *unsorted* samples (**A**). This likely leads to metabarcoding datasets being dominated by a few biomass rich or abundant taxa, while small and rare ones might remain undetected. If the goal of the study is, to detect all taxa present in the sample, it might make sense to adjust the biomass to have all specimens equally strong represented in the dataset. This can be done by *sorting* specimens into size categories (e.g. small, medium and large specimens), followed by sequencing of individual size fractions (**E**) or pooling them proportionally based on specimen abundance in each fraction (see **B**, **C** and **D**). It is however difficult to precisely pool ground tissue (**B**). Extracted DNA on the other hand has to be quantified and might be affected by copy number variation of mitochondrial genomes between taxa (**D**). Thus, in this study pooled digested tissue from each size category (**C**) was used to investigate the effects of **so**rted and **un**sorted samples.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

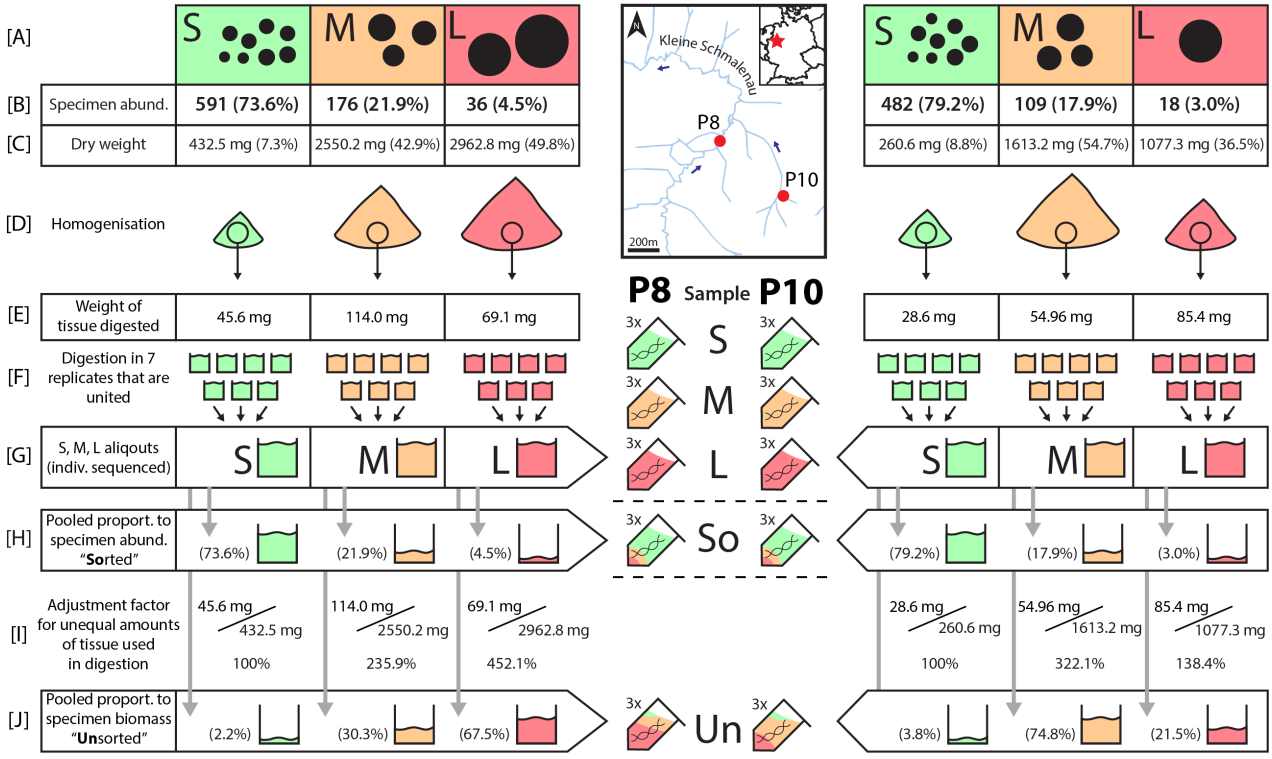


Figure 2: Strategies how digested tissue was pooled, to generate samples which retained the original proportion of small, medium and large specimens, as if the sample has not been sorted ("unsorted" [J]) and a sample where size sorting did take place and specimens of each size category are proportionally pooled by specimen abundance ("sorted" [H]). Specimens of both kick samples were sorted by specimen size into three size categories; small, medium and large [A]. Using the specimen abundance in each category [B], as well as total dry weight [C], "sorted" and "unsorted" samples were generated by pooling digested tissue in specific proportions. To generate a unsorted mock sample digested liquid was pooled based on dry specimen weight in each size category [J] under consideration of how much tissue was used in the digestion [E,I]. To adjust for specimen biomass, the sorted specimens were pooled according to the number of specimens in each size category [H]. For the sorted mock sample [H] we mistakenly did not consider the tissue adjustment factor [I]. After pooling of digested tissue three aliquots were extracted for each category (small, medium, large specimens, sorted and un unsorted), which each were united into a single DNA aliquot used for metabarcoding.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

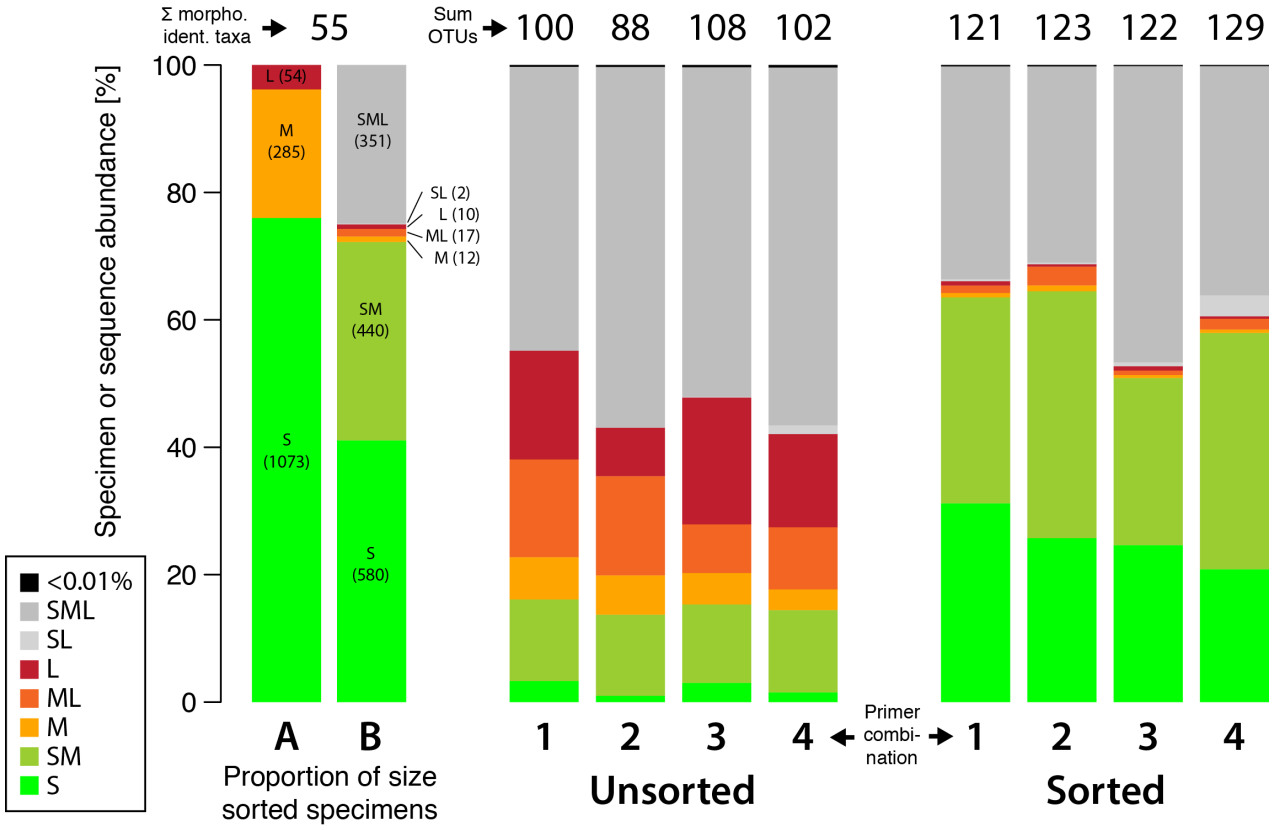
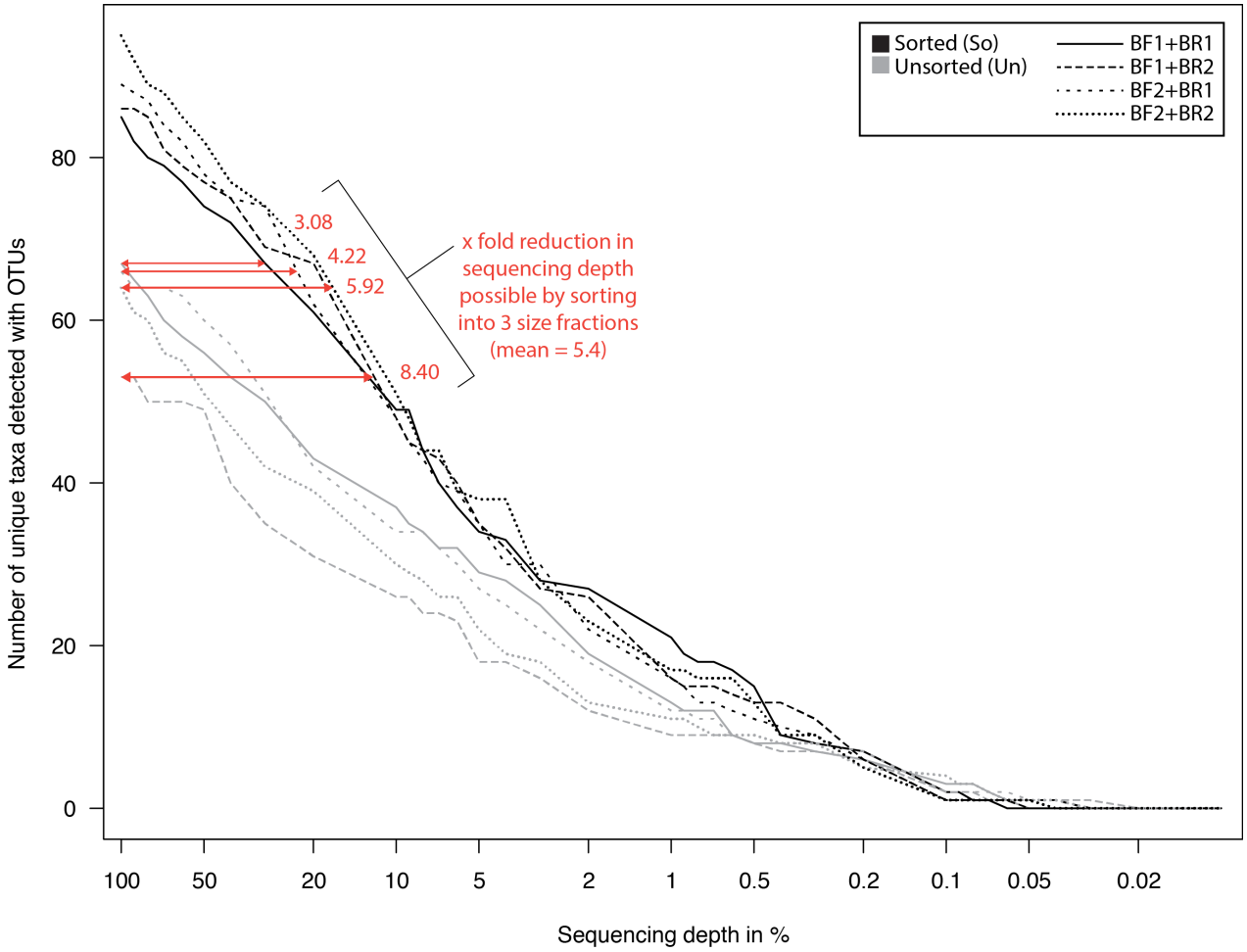


Figure 3: Comparison of specimen number in each size category against the respective OTU read abundance of unsorted and sorted samples with 4 different primer sets. The proportion of sorted specimens is shown in barplot A, while plot B is showing how many morphologically identified taxa are sharing the same size categories. For example, if a taxon is represented by small, medium and large specimens it gets assigned to 'SML' (grey) in the metabarcoding dataset as all specimens contribute DNA which clusters into the same OTU regardless of specimen size. Thus, reads can not always be reassigned to small, medium or large specimens, but a combination of those (see also figure S7). Numbers above the plot give the total number of taxa identified with morphology and the number of OTUs detected with each primer set for unsorted and sorted samples. The numbers 1 - 4 below the plots indicate the different primer combinations used; 1 = BF1+BR1, 2 = BF1+BR2, 3 = BF2+BR1, 4 = BF2+BR2.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017



344

345 **Figure 4**: Amount of detected taxa based on OTUs with unsorted (Un) and sorted samples (So) for the four tested primer

346 combinations, considering different sequencing depths. The sequencing depth is plotted on a logarithmic scale.

347

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

References

Ballard, J.W.O. & Whitlock, M.C., 2004. The incomplete natural history of mitochondria. 13(4), pp.729–744.

Bendich, A.J., 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays: news and reviews in molecular, cellular and developmental biology*, 6(6), pp.279–282.

Crampton-Platt, A. et al., 2016. Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, pp.1–11.

Dowle, E.J., Pochon, X. & Banks, J.C., 2015. Targeted gene enrichment and high‐throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Molecular Ecology*.

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

Edgar, R.C. & Flyvbjerg, H., 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, 31(21), pp.3476–3482.

Elbrecht, V. & Leese, F., 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol M. Hajibabaei, ed. *PloS one*, 10(7), pp.e0130324–16.

Elbrecht, V. & Leese, F., 2017. Validation and development of freshwater invertebrate metabarcoding COI primers for Environmental Impact Assessment. *Frontiers in Freshwater Science*. Available at: http://journal.frontiersin.org/article/10.3389/fenvs.2017.00011/abstract#.

Elbrecht, V. et al., 2017. Assessing strengths and weaknesses of DNA metabarcoding based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, pp.1–21.

Gibson, J. et al., 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences*, 111(22), pp.8007–8012.

Gómez-Rodríguez, C. et al., 2015. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages M. Gilbert, ed. *Methods in Ecology and Evolution*, 6(8), pp.883–894.

Haase, P. et al., 2004. Assessing streams in Germany with benthic invertebrates: development of a practical standardised protocol for macroinvertebrate sampling and sorting. *Limnologica*, 34(4), pp.349–365.

Hajibabaei, M. et al., 2011. Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. *PloS one*, 6(4), pp.1–7.

Ji, Y. et al., 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. M. Holyoak, ed. *Ecology letters*, 16(10), pp.1245–1257.

Lemire, B., 2005. Mitochondrial genetics. *WormBook : the online review of C. elegans biology*, pp.1–10.

Leray, M. & Knowlton, N., 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 112(5), pp.2076–2081.

Leray, M. & Knowlton, N., 2017. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, 5, pp.e3006–27.

Liu, S. et al., 2016. Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular ecology resources*, 16(2), pp.470–479.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

pp.10–12.

Meier, C. et al., 2006. Methodisches Handbuch Fließgewässerbewertung. pp.1–110.

Moraes, C.T., 2001. What regulates mitochondrial DNA copy number in animal cells? *Trends in genetics : TIG*, 17(4), pp.199–205.

Morinière, J. et al., 2016. Species Identification in Malaise Trap Samples by DNA Barcoding Based on NGS Technologies and a Scoring Matrix D. Fontaneto, ed. *PloS one*, 11(5), pp.e0155497–14.

Piñol, J. et al., 2014. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular ecology resources*, 15(4), pp.1–12.

Stein, E.D. et al., 2013. Does DNA barcoding improve performance of traditional stream bioassessment metrics? *Freshwater Science*, 33(1), pp.302–311.

Sunnucks, P. & Hales, D.F., 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). *Molecular biology and evolution*, 13(3), pp.510–524.

Sweeney, B.W. et al., 2011. Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, 30(1), pp.195–216.

Taberlet, P. et al., 2012. Environmental DNA. *Molecular Ecology*, 21(8), pp.1789–1793.

Tang, M. et al., 2015. High-throughput monitoring of wild bee diversity and abundance via mitogenomics M. Gilbert, ed. *Methods in Ecology and Evolution*, 6(9), pp.1034–1043.

Tang, M. et al., 2014. Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic acids research*, 42(22), pp.gku917–e166.

Wangensteen, O.S. & Turon, X., 2016. Metabarcoding techniques for assessing biodiversity of marine animal forests. *Marine Animal Forests. The Ecology of Benthic Biodiversity Hotspots.*, pp.1–34.

Zimmermann, J. et al., 2014. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular ecology resources*, pp.1–17.

Supporting information

Figure S1. Pictures of sorted specimens

Pictures of the specimens sorted into small, medium and large individuals. Also provides information on how S, M and L tissue was pooled to generate the proportionally sorted (So) and unsorted (Un) samples.

Figure S2. Flowchart detailing laboratory processing

Overview of the steps carried out for sample sorting and processing in the laboratory.

Figure S3. DNA extraction protocol

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

Shows the step where the digested buffers of S, M and L were pooled to generate unsorted (Un) and sorted (So) samples.

Figure S4. Sequencing depth and sequences discarded in bioinformatic processing

Barplot showing the number of total reads and proportion of sequences discarded in subsequent bioinformatic processing steps for all samples.

Figure S5. Flowchart detailing the bioinformatic pipeline

Figure giving an overview of the metabarcoding pipeline applied to this dataset.

Figure S6. Reproducibility between HiSeq lanes

Comparison of relative OTUs abundances between both HiSeq lanes.

Figure S7. Plot of OTU table

Visualisation of taxa detected within S, M, L, Un, So DNA extractions, with 4 different primer combinations. Data is also compared to morphological identifications and number of specimens of each morphologically identified taxon.

Figure S8. Database completeness

Plot showing the percent match of each OTU to the reference database, under consideration of read abundance.

Figure S9. Taxa identification with metabarcoding and morphology

Comparison of number of taxa identified with morphology and DNA metabarcoding on different taxonomic resolutions.

Figure S10. Taxa detection in sorted and unsorted samples

Comparison of the amount of diversity and taxa detected in sorted samples (So) and unsorted samples (Un).

Table S1. Identification literature

Overview of identification keys used for the different macroinvertebrate groups.

Table S2. OTU table

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

449 Detailed OTU table giving the number of reads for each sample, including assigned taxonomy and OTU sequence. OTUs
450 with below 0.01% sequence abundance in each sample (highlighted in Orange), were set to 0 for statistical analysis.
451
452 **Table S3.** Morphologically identified taxa
453 Table giving an overview of morphologically identified taxa and abundance of specimens in S, M and L for both sample
454 locations.
455

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v3 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017