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

1) Environmental bulk samples often contain many taxa with biomass differences of several orders of magnitude. This can be problematic in DNA metabarcoding and metagenomic high throughput sequencing approaches, as large specimens contribute disproportionate 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 and balancing the amounts of tissue used per size fraction should improve detection rates, but this approach has not been systematically tested.

2) Here we tested the effects of size sorting on taxa detection using freshwater macroinvertebrates. Kick sampling was performed at two locations of a low-mountain stream in West Germany, specimens were morphologically identified and sorted into small, medium and large size classes (< 2.5x5, 5x10 and up to 10x20 mm). Tissue from the 3 size categories was extracted individually, and pooled to simulate samples that were not sorted by biomass and samples that were sorted and then pooled so that each specimen contributed approximately equal amounts of biomass. DNA from all five extractions of samples from both sites was amplified using four different DNA metabarcoding primer sets targeting the Cytochrome c oxidase I (COI) gene. The library was sequenced on a HiSeq Illumina sequencer.

3) Sorting taxa by size and pooling them proportionately according to their abundance lead to a more equal amplification 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.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 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.

4) Our results demonstrate that even a coarse size sorting can substantially improve detection of taxa 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.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 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), thousand specimens can be easily collected. However, manually identifying hundreds or thousands of specimens in a single sample is often not feasible. Bulk samples, which previously took weeks or month 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 DNA 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)) 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.* 2015)). 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 by the lack of adequate reference data (e.g. mitochondrial genomes) and a high sequencing depth is required (Crampton-Platt *et al.* 2016). While these problems might be solved at least partially by optimised degenerate primers (Elbrecht & Leese 2016), reduced sequencing costs and mitogenome capture (Tang *et al.* 2014), both metabarcoding and metagenomics are biased 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 unsorted samples in bulk, 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 taxon specific primer bias, unrelated to specimen size). Thus just a few big specimens in a sample will likely make up a majority of the reads, requiring higher sequencing depth to also detect small or rare specimens and taxa. Some studies have already sorted their samples into different size fractions, for DNA metabarcoding because of this biomass introduced bias (Leray & Knowlton 2015; Wangensteen & Turon 2016). However, to our knowledge the effect of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

fractioning samples by specimen biomass against the complete sample without pre-sorting of specimens has never been systematically tested and quantified. (Morinière *et al.* 2016) have found effects of sorting malaise trap samples by insect orders (potentially caused by unequal sequencing depth), but also encouraged the testing of size based sample fractioning. In this study we systematically quantified the effects of biomass-sorting on taxon recovery using two complete stream macroinvertebrate kick samples, 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 1 & S1. Specimens of each size class were then homogenized and DNA extracted. DNA from each size fraction was pooled proportionately to the dry weight of each size category, to generate an **un**sorted sample (Un), and additionally pooled by specimen abundance in each size category, to generate a **so**rted sample (So) in which ideally all specimens are equally well represented. By metabarcoding these unsorted and sorted samples, we can precisely investigate the effects of sample size sorting on taxa recovery.

2) Material and Methods

Sample collection and processing

Figure S2 gives an overview of sample sorting and laboratory processing steps. Macroinvertebrates were collected at two sampling points of the small low mountain range stream Kleine Schmalenau in West Germany (Arnsberger Wald). The main stream (P8, N51.43623 E8.13721) and a small tributary (P10, N51.43295 E8.14350) were each sampled with 5 kick samples per spot (0.45 m^2 area) following the general principle of the multi habitat sampling protocol also used in 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 based on morphology to the lowest taxonomic level that could be determined accurately and consistently.

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. Specimens below 2.5x5 mm 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). For thin but long specimens like e.g. chironomids (non-biting midges), the total surface was considered and evaluated if it fitted into the respective rectangle. Terrestrial taxa and Trichoptera (caddisfly) cases were included in the samples.

DNA extraction and tissue pooling

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

Specimens of each size category were dried overnight 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 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.

We wanted to extract and sequence DNA from each size category (S, M & L), but also simulate extracting DNA from 1) proportionally pooled specimens based on number of specimens in each size category (So = "sorted") and 2) also extract DNA from the whole sample as if it was never sorted into S, M and L (Un = "unsorted") by pooling tissue based on total dry tissue weight in each size category. While we could have pooled two subsets of homogenised tissue from each size category, we wanted to avoid this as tissue subsets might differ slightly in taxa composition. Thus, tissue from each category was digested following a modified salt DNA extraction protocol and then the lysate pooled (Sunnucks & Hales 1996), Figure S3). Seven replicates were in each size category, which were united after tissue lysis. This way a higher amount of tissue could be digested for the S, M and L samples. DNA extraction was then carried out in triplicate using 3 times 530 µl TNES extraction buffer for S, M and L as well as the sorted (So) and unsorted (Un) samples. Figure 1 gives a schematic overview of how S, M and L were pooled to generate sorted and unsorted samples (see Figure S1 for exact proportions). To generate the unsorted samples (Un), buffer from S, M and L was pooled based on dry specimen weight in each category. The same S, M and L solutions were additionally pooled proportionally to the total number of specimens in each size category (So), so all specimens contribute approximately equal amounts of DNA in the bulk extraction. 15 µl DNA were pooled from each of the 3 extraction replicates and digested with 1 µl RNAse A and cleaned up using a MinElute Reaction Cleanup Kit (Qiagen, NL) with resuspension in ddH_2O. DNA concentrations were quantified fluorometric using a Qubit with (HS Kit) and concentrations adjusted to 25 ng/µl.

DNA metabarcoding and bioinformatics

All 10 samples (sample site 8 and 10 with S, M, L, Un, So each) were amplified with the four freshwater macroinvertebrate fusion primer sets BF / BR as described in (Elbrecht & Leese 2016) (see Figure S4 for sample tagging). Each PCR reaction was composed of 1× PCR buffer (including 2.5 mM Mg^{2+}), 0.2 mM dNTPs, 0.5 µM of each primer, 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 5 wells before PCR. PCR reactions were run in a Biometra TAdvanced Thermocycler using the flowing 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

inhibitors present in the samples. PCR products were purified and left size selected 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. Sequencing was done on one lane of an Illumina HiSeq 2500 system with a rapid Run 250 bp PE v2 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.

Figure S5 gives an overview of our bioinformatic pipeline. We used the UPARSE pipeline in combination with custom R scripts (Dryad DOI - NA) for data processing (Edgar 2013). 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 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).

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. Then for each sample, OTUs with less or equal than 0.01% were set to 0% sequence abundance. Taxonomy was assigned to remaining OTUs using an R script searching the BOLD and NCBI database.

3) Results

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 (SRR3399056 and SRR3399057). On average 1.71 (SD = 0.29, lane 1) and 2.17 (SD = 0.38, lane 2) million read pairs were

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

obtained for each sample after demultiplexing (Figure S4). Read quality 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).

The OTU raw data is available in Table S1 as well as morphology based identifications and taxa abundances in Table S2. After clustering and discarding low abundance OTUs, a total of 314 OTUs remained in the data set (Figure S7, Table S1). 71% of these OTUs could be reliably identified with available reference databases, with 58% of the OTUs belonging to target invertebrate taxa (Figure S8). High abundance OTUs (at least 0.1% of reads) belonged all to the invertebrate target groups, with 45 of 52 these taxa 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 much target taxa were detected than with morphology based identification alone, with 5 times more on species level alone (Figure S9). The main stream (P8) and tributary (P10) could be clearly distinguished, with only 14.3% of OTUs shared between both sites (Figure S8, 36.4% similarity based on morphological identification, Table S1).

Sorting the sample into 3 size categories and proportional pooling of DNA extracts by amount of specimens in each category reduced the bias introduced by large specimens substantially (Figure 2). The sorted samples (So) resembled the composition of the original sample much better than the unsorted samples (Un). 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 3). By using the S M and L samples as controls, we could estimate the **e**xpected (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 could be 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 show 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 less than with 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). The reduced amount of taxa detected with the unsorted samples, persists when the sequencing depth is reduced (Figure 3). Sample sorting does reduce the required sequencing depth to detect the same amount of taxa by ~5 times, compared to the unsorted samples.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

180

181 # 4) Discussion

182 ## 4.1) Data basis and reliability of results

183 Freshwater macroinvertebrate COI reference databases were quite reliable for the studied ecosystem. Common taxa in our
184 samples often had 100% matches to BOLD and NCBI databases. However, the reliability of DNA barcoding database
185 entries depends highly on the expert knowledge and accuracy when morphologically identifying specimens. Not all
186 database entries have resolved taxonomy or possibly misidentified specimens, potentially inflating the number of taxa
187 detected. While most metabarcoding studies will likely be affected by this issue and also the number of morphotaxa
188 detected with OTUs in this study might be inflated, the over all investigation of size sorting is likely unaffected as the bias
189 will be the same for all samples and size categories. Nevertheless, DNA based identifications can be more accurate than
190 classical morphology based identification (Stein et al. 2013; Sweeney et al. 2011) as we also show with our two kick
191 samples in this project. We did make sure to only morphologically identify specimens to a level we were confident the
192 identifications were correct, as we relate the size classes of each taxon back to the metabarcoding data (Figure 2).
193 Furthermore, we show with our dataset that stochastic effects during Illumina sequencing affect mainly low abundant OTUs.
194 While this effect adds variability to low abundant OTUs, it does not affect the detection of OTUs with an abundance of >
195 0.01%. However, when analysing OTUs with lower abundance stochastic effects should be taken into consideration.

196

197 ## 4.2) Effects of sorting metabarcoding samples by specimen size

198 We sorted two samples by specimen size (resembling biomass) into small, medium and large specimens and pooled them
199 proportionately by specimen abundance per size class to compare these results against unsorted samples. Our results
200 demonstrate unambiguously that, as expected, read abundances of the unsorted samples were dominated by few taxa with
201 large specimens that contribute the majority of DNA when bulk extracting samples. This not only does skew the read
202 abundances in favor of biomass rich specimens, but also some smaller and less abundant taxa remained undetected (on
203 average 30% fewer taxa detected in the unsorted samples). The sorted samples only need 1/5 of the sequencing depth, to
204 detect the same amount of taxa as in the unsorted samples. This means that sorting metabarcoding bulk samples by
205 specimen biomass can substantially reduce sequencing costs. While we only manually sorted our samples into 3 size
206 categories, further cost reductions might be possible by sorting samples into more size categories. It is likely that larger
207 specimens will have similar effects on metagenomic bulk samples, thus sorting by specimen size might also likely be viable

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

for these samples.

4.3) 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 the 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 bulk samples are variable in biomass and sequencing depth should be maximised, thus sorting your samples will increase the number of taxa detected. The method of size sorting depends on sample composition and characteristics. If samples just contain a few large specimens, 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. In this study, we measured the size of each individual specimen under a stereo microscope to get very accurate size classes needed to test this method. With approximately 2-3 hours for each sample and 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 10 mm to 63 µm (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 of 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

4.4) Conclusions

We demonstrated that sorting metabarcoding samples into 3 specimen size categories and then pooling the tissue proportionally to the number of specimens in each size class, can reduce the amount of required sequencing depth compared to the unsorted complete 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 your metabarcoding project. Also some metagenomic projects will likely profit from presorting samples by biomass, but we did not explicitly test this here so we can only hypothesise.

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 in the field work, as well as Arne Beermann for taxonomic validation. We further like to thank for Brian Gill, Jan Macher, Edith Vamos, Vera Zizka proofreading this manuscript. The authors report no conflicts of interest.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Author Contributions statement: VE, BP and FL conceived the ideas and designed methodology; BP identified specimens and carried out the laboratory work; VE performed bioinformatic analyses; VE led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

Figures

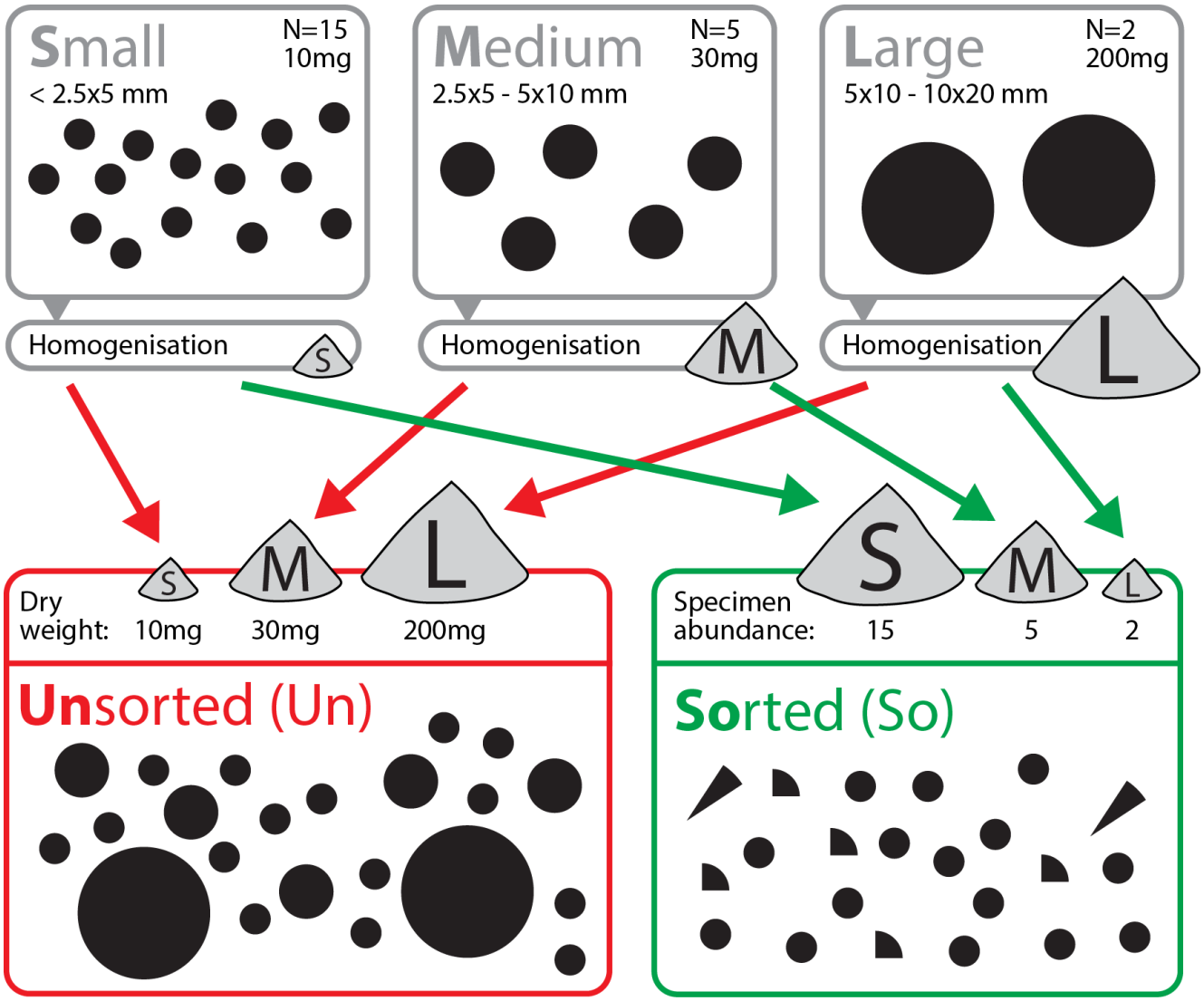


Figure 1: Schematic overview of the laboratory processing of macroinvertebrate samples. Specimens from each sample were sorted into 3 size categories (**S**mall, **M**edium and **L**arge) and then processed individually, pooled to simulate complete samples without sorting (**Un**sorted) and samples in which subsamples are pooled proportionally to taxa abundance (**So**rted). N refers to the number of taxa in each size category in this hypothetical example, with the respective total weight given in mg. See Figure S1 for actual numbers of the used samples.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

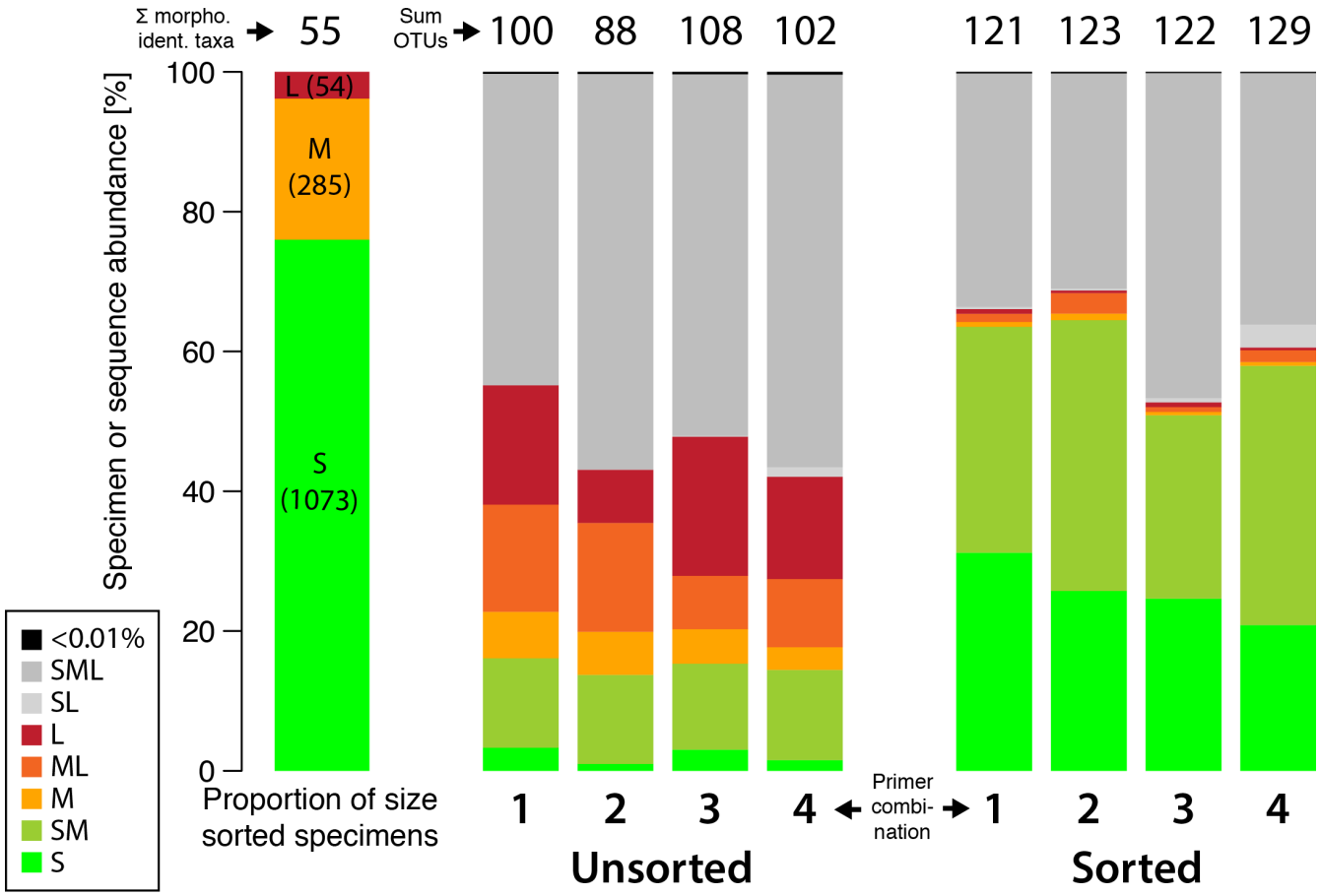
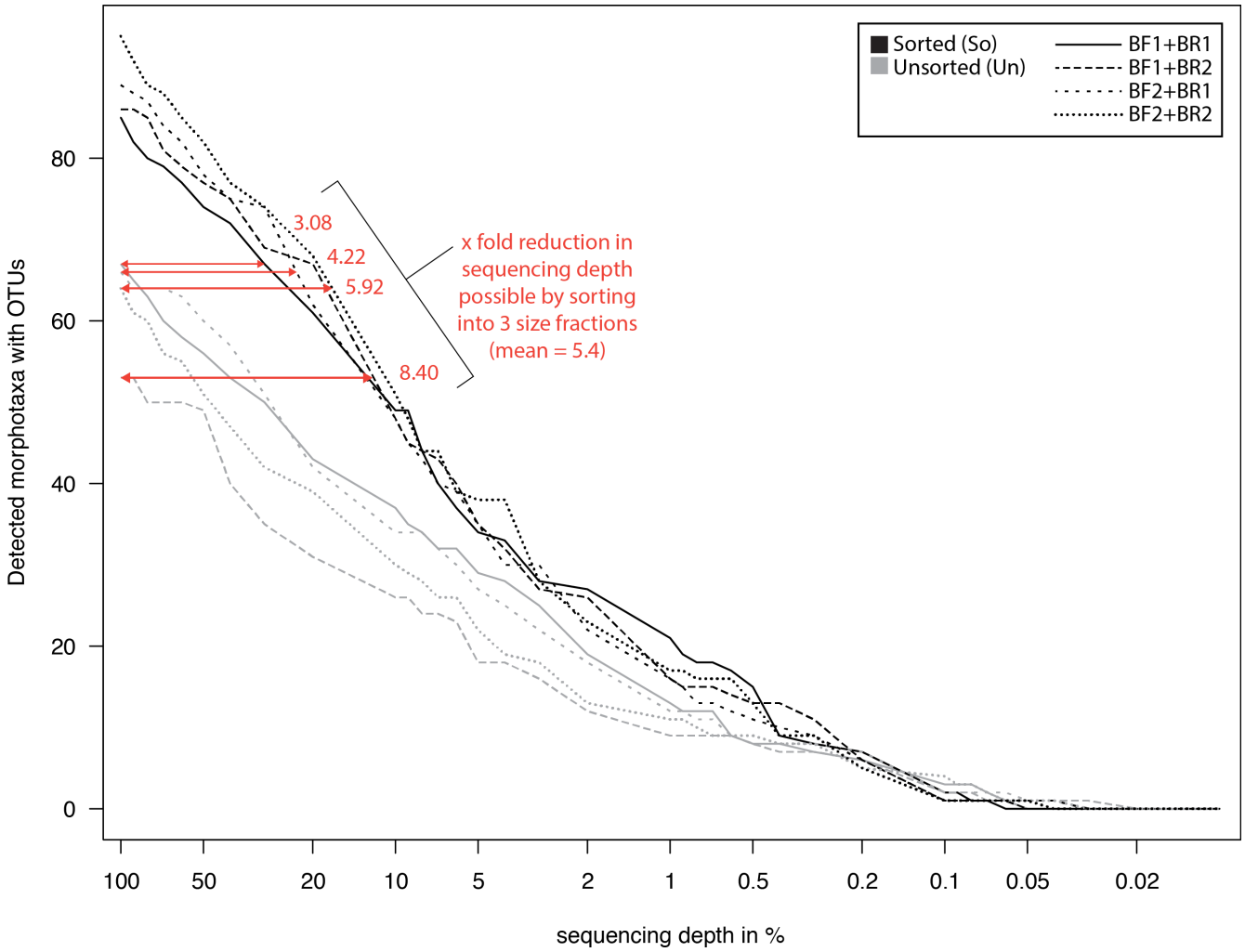


Figure 2: Comparison of specimen number in each size category against the respective OTU read abundance of unsorted (Un) and sorted samples (So) with 4 different primer sets. Each size category (S, M and L) was also sequenced individually, and thus the data could be used to assign size classes to the OTUs in the sorted and not sorted samples. Sometimes, one OTU included reads of specimens from more than one size class, leading to assignment of several size categories (Gray = OTU containing specimens in S, M and L). Numbers above the plot give toe total amount 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.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017



280

281 **Figure 3**: Amount of detected taxa based on OTUs with unsorted (Un) and sorted samples (So), considering different

282 sequencing depth. The sequencing depth is plotted on a logarithmic scale.

283

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

References

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., 2016. Validation and development of freshwater invertebrate metabarcoding COI primers for Environmental Impact Assessment. *PeerJ PrePrints*, pp.1–16.

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.

Hajibabaei, M. et al., 2011. Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. *PloS one*.

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

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*, pp.201424997–6.

Liu, S. et al., 2015. Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular ecology resources*.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17, pp.10–12.

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

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

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

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017

353

354 **Figure S7.** Plot of OTU table

355 Visualisation of taxa detected within S, M, L, Un, So DNA extractions, with 4 different primer combinations. Data is also

356 compared to morphological identifications and number of specimens of each morphologically identified taxon.

357

358 **Figure S8.** Database completeness

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

360

361 **Figure S9.** Taxa identification with metabarcoding and morphology

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

363

364 **Figure S10.** Taxa detection in sorted and unsorted samples

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

366

367 **Table S1.** OTU table

368 Detailed OTU table giving the number of reads for each sample, including assigned taxonomy and OTU sequence. OTUs

369 with below 0.01% sequence abundance in each sample (highlighted in Orange), were set to 0 for statistical analysis.

370

371 **Table S2.** Morphologically identified taxa

372 Table giving an overview of morphologically identified taxa and abundance of specimens in S, M and L for both sample

373 locations.

374

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2561v2 | CC BY 4.0 Open Access | rec: 22 Jan 2017, publ: 22 Jan 2017