DR. CARMELO ANDUJAR (Orcid ID : 0000-0001-9759-7402)

PROF. ALFRIED VOGLER (Orcid ID : 0000-0002-2462-3718)

Article type : Original Article

Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill

Carmelo Andújar[1,2,3], Paula Arribas[1,2,3], Clare Gray[2], Katherine Bruce[4], Guy Woodward[2], Douglas W. Yu[5,6], Alfried P. Vogler[1,2]

[1] Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK

[2] Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, SL5 7PY, UK

[3] Grupo de Ecología y Evolución en Islas, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), San Cristóbal de la Laguna 38206, Spain

[4] NatureMetrics Ltd, CABI Site, Bakeham Lane, Egham, Surrey, TW20 9TY, UK

[5] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

[6] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ, UK

Correspondence to:

Dr Carmelo Andujar, Grupo de Ecología y Evolución en Islas, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), San Cristóbal de la Laguna 38206, Spain

Email candujar@um.es

Running title: Metabarcoding of freshwater invertebrates

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.14410

Abstract

Biomonitoring underpins the environmental assessment of freshwater ecosystems and guides management and conservation. Current methodology for surveys of (macro)invertebrates uses coarse taxonomic identification where species-level resolution is difficult to obtain. Next-generation sequencing of entire assemblages (metabarcoding) provides a new approach for species detection, but requires further validation. We used metabarcoding of invertebrate assemblages with two fragments of the *cox1* "barcode" and partial nuclear ribosomal (*SSU*) genes, to assess the effects of a pesticide spill in the River Kennet (Southern England). Operational Taxonomic Unit (OTU) recovery was tested under 72 parameters (read denoising, filtering, pair merging and clustering). Similar taxonomic profiles were obtained under a broad range of parameters. The SSU marker recovered Platyhelminthes and Nematoda, missed by *cox1*, while Rotifera were only amplified with *cox1*. A reference set was created from all available barcode entries for Arthropoda in the BOLD database and clustered into OTUs. The River Kennet metabarcoding produced matches to 207 of these reference OTUs, five times the number of species recognised with morphological monitoring. The increase was due to: greater taxonomic resolution (e.g. splitting a single morphotaxon 'Chironomidae' into 55 named OTUs); splitting of Linnaean binomials into multiple molecular OTUs; and the use of a filtration-flotation protocol for extraction of minute specimens (meiofauna). Community analyses revealed strong differences between "impacted" vs. "control" samples, detectable with each gene marker, for each major taxonomic group, and for meio- and macro-faunal samples separately. Thus, highly resolved taxonomic data can be extracted at a fraction of the time and cost of traditional non-molecular methods, opening new avenues for freshwater invertebrate biodiversity monitoring and molecular ecology.

INTRODUCTION

The freshwater biota is affected by a host of natural environmental drivers and, increasingly, anthropogenic disturbances that alter local species assemblages. Biomonitoring therefore is required to assess the ecological status of freshwaters and to enforce their protection through legislation, such as the Water Framework Directive (WFD) of the European Union and the US Clean Water Act (United States 1972; European Commission 2000). However, this field of applied ecology is still largely reliant on techniques that were developed over a century ago, albeit with some statistical advances, tweaks, and adjustments in the intervening years, and has been roundly criticised for failing to adapt to a rapidly changing world (Friberg *et al.* 2011). The vast majority of biomonitoring schemes still relies on identifying macroinvertebrates by eye, or at best via microscopy, to a coarse level of taxonomic resolution, and the molecular revolution that is overtaking mainstream ecology has yet to be embraced (Pauls *et al.* 2014; Bohan *et al.* 2017). Because of the need for rapid and cost-

effective approaches, it is routine practice that many taxa are not identified to individual species but instead are lumped taxonomically, e.g. by family, as used in RIVPACS and AUSIVAS systems (Wright *et al.* 2000), or, less frequently, into trait-based groupings, such as "riverflies". The taxonomically difficult groups in which most of the aquatic biodiversity resides (e.g. chironomid midges) are typically either ignored or treated as a single entity (Schmidt-Kloiber & Nijboer 2004; Jones 2008).

These labour-saving shortcuts can nonetheless provide a broad assessment of the ecological state of a water body, despite the huge amounts of environmental-status information that are inevitably jettisoned in the process, and has been successfully used for the assessment of habitat and water quality for many decades (Camargo 1993; Marshall *et al.* 2006; Sánchez-Montoya *et al.* 2007). However, population responses to changes in water quality can differ between even closely related species, and so taxonomically coarse inventories may miss the full impact of important environmental stressors (Stubauer & Moog 2000; Chessman *et al.* 2002; Gutiérrez-Cánovas *et al.* 2008). Species-level identification can establish the link to known ecological, physiological and behavioural traits, which may reflect differential responses to environmental conditions, and also may reveal the membership in feeding groups and position in trophic networks (Bohan *et al.* 2017). These distinctions are lost if the assessment is at the level of genera or families, or other such coarse groupings (Schmidt-Kloiber & Hering 2015; Leese *et al.* 2016).

Recent protocols for metabarcoding, i.e. the sequencing of PCR amplicons from environmental specimen mixtures, could provide faster and more highly-resolved taxonomic identification of complex assemblages (Taberlet *et al.* 2012). This methodology applies Hebert et al.'s (2003) idea of species identification through short diagnostic DNA barcodes (a fragment of the *cox1* gene) to the community level, and thanks to new high-throughput sequencing (HTS) technology, the effort required for DNA barcoding of an entire assemblage now is not much greater than required for a single specimen with Sanger-sequencing (Taylor & Harris 2012; Brandon-Mong *et al.* 2015). Metabarcoding permits the simultaneous analysis of large numbers of minute specimens obtained from environmental samples, such as soil and leaf litter (Yang *et al.* 2014; Arribas *et al.* 2016; Zinger *et al.* 2016), the deep sea (Esling *et al.* 2015; Guardiola *et al.* 2015; Leray & Knowlton 2015; Lanzén *et al.* 2016), and freshwater sediments and the water column (Elbrecht & Leese 2015; Bista *et al.* 2017). Metabarcoding can thus provide the elusive species-resolution desired for biomonitoring of entire ecosystems, and also for capturing the large proportion of organisms that are either too small to see or identify using traditional sorting and microscopy techniques (Creer *et al.* 2010; Ji *et al.* 2013; Hajibabaei *et al.* 2016; Bohan *et al.* 2017).

Here, we applied metabarcoding to study the consequences of an insecticide spill on invertebrate freshwater communities in a large lowland river as a test case. On July 1, 2013, a pulse of the

organophosphate chlorpyrifos in the River Kennet, the largest tributary of the River Thames in southern England, led to population crashes and localised extinctions of many invertebrate taxa (see Thompson *et al.* 2016 for details). We used samples collected upstream and downstream from the spill site over several km of the river's length to explore the effectiveness of metabarcoding, and to trial new environmental diagnostic protocols for identifying differential responses of invertebrate communities to a profound environmental perturbation. For example, some components of the local community such as the dominant detritivore, the amphipod *Gammarus pulex*, were greatly reduced in number downstream from the spill, whereas other taxa, especially those with an aerial adult life stage, were far less affected, and at a later sampling time returned to post-spill levels, possibly due to their ability to recolonize rapidly. The Chironomidae (non-biting midges) as a group greatly increased in abundance after the spill. However, because community composition was only measured at higher taxonomic levels, rather than with species-level resolution, it is not possible to gain further insight into the mechanisms of ecological resilience and recovery after the spill. Specifically, the species composition of the post-impact chironomid community might be largely unchanged from the pre-impact community, or, despite the increased abundance, it might be composed of a subset of that community (nestedness) or of a new set of species dispersed from elsewhere (turnover).

The *cox1* gene is the obvious choice of a marker for metabarcoding of aquatic invertebrates, but due to the constraints on read length, the widely used Illumina platform is not suited for sequencing the full-length amplicon of the barcode region (658 bp). We have metabarcoded two gene fragments covering the entire *cox1* barcode region using two primer pairs shown to have broad target ranges (Arribas *et al.* 2016), which here were applied for the first time to freshwater invertebrate communities. The parallel use of two barcode fragments provides a test of amplification breadth and potential biases due to primer choice, which could affect the success of species detection and delimitation. A major concern is that PCR amplification of the *cox1* region in several aquatic phyla is generally low and thus this region may produce bias in the detectable species assemblages (Deagle *et al.* 2014; Lobo *et al.* 2015; Creer *et al.* 2016). We therefore also conducted metabarcoding with the nuclear 18S rRNA (*SSU*) gene, frequently used for sequencing marine meiofaunal communities but never tested in freshwater ecosystems to our knowledge. This gene contains highly conserved regions bracketing more variable segments and thus is less affected by primer bias across a larger phylogenetic range of taxa. However, lower sequence variation in *SSU* generally underestimates the true species diversity (Tang *et al.* 2012). The resulting metabarcode sequences are typically first clustered into *de novo* generated species proxies, i.e. Operational Taxonomic Units (OTUs) (Blaxter *et al.* 2005), that can be directly used for downstream ecological analyses. Species identification is critical for many uses of these data, and can be obtained against existing databases of DNA sequences from fully identified specimens available at public databases (NCBI or BOLD). These reference sets can be used in two ways, either by matching the *de novo* generated OTUs against the external

reference sequences, in a 'taxonomy independent' approach, or by matching the raw sequence reads directly to the reference set without prior OTU clustering in a 'taxonomy dependent' approach (Schloss & Westcott 2011), which has been used on various occasions to test species presence or absence (e.g. Shokralla *et al.* 2014; Arribas *et al.* 2016).

The River Kennet pesticide spill, characterised previously with conventional approaches (Thompson *et al.* 2016), was used to trial the metabarcoding methodology for freshwater invertebrates. This included the development of protocols for extraction of meio- and macro-fauna from bulk sediment samples, the evaluation of existing universal primers for amplification of the *cox1* and nuclear 18S ribosomal RNA (*SSU*) genes, and the calibration of bioinformatics tools and parameter settings for accurate estimates of species numbers and species identification. For identification of the local community we made use of the rapidly growing publicly available taxonomic sequence databases, whose species representation is increasingly complete at least for this ecosystem in Western Europe. Given the high quality of sequence data achievable with recent Illumina technology, future biomonitoring schemes may shift to the use of metabarcoding.

MATERIALS AND METHODS

Study site and sampling protocol

The River Kennet is a lowland chalk river that was affected by widespread macroinvertebrate mortality along a 15-km stretch downstream from an insecticide spill site (Thompson *et al.* 2016). Invertebrates were collected using a Surber sampler (0.0625 m^2, 335 µm mesh) at three upstream control and three downstream impacted reaches, each 50 m long, along a ca. 6 km river stretch (including the four sites sampled in Thompson *et al.* 2016). Sites were ca. 1 km apart, with similar channel forms and riparian surroundings, and were sampled at two times: time 1 (12th July 2013), 11 days after the spill; time 2 (17th September 2013) (Suppl. Fig. S1). The latter was the same time as samples used in Thompson *et al.* (2016). The sampling regime permitted to explore the immediate effect downstream of the spill point relative to the unaffected upstream sites, and the short-term recovery of the arthropod community 2.5 months after the spill. One Surber sample per site and time was preserved immediately after collection in absolute ethanol and transferred to the laboratory, where we removed debris by hand and subsequently filtered the remainder through a 1 mm wire mesh sieve to retain macrofauna. The smaller material not retained by this sieve was then passed through a 45 µm wire mesh sieve to capture the meiofaunal fraction (size < 1mm) while flushing out microorganisms and silt with copious amounts of water (Fonseca *et al.* 2010; Arribas *et al.* 2016). Note that the original Surber used a 335 µm mesh but many organisms below this size were retained in the Surber sample by debris and no effort was made to remove small organisms at this stage. The

filtrate from this second step was cleaned by flotation using LUDOX 40^{TM} (Burgess 2001) to separate organisms, which tend to float, from inorganic particles, which tend to sink. The floating layer was extracted for DNA to represent the sampled meiofauna. Each sample was processed separately for the macro- and meiofauna, for a final number of 24 samples used for DNA extraction and sequencing (Fig. 1).

DNA extraction and Illumina sequencing

Each sample was dried and homogenised in a Falcon tube, and DNA was extracted from 200 µl of sample lysate using a DNeasy Blood and Tissue Spin-Column Kit (Qiagen). Three DNA markers were individually amplified: a fragment of the *SSU* gene, and two fragments (*bc5'* and *bc3'*) within the mitochondrial *cox1* barcode region. The two fragments were bracketed by the "Folmer" primers used for amplification of the standard animal barcode (Hebert et al., 2003), but with a higher degree of degeneracy. The *bc5'* fragment corresponds to ≈350 bp of the 5' end of the *cox1* barcode fragment, and was amplified using primers already validated in a wide variety of arthropods (Fol-degen-for: 5' TCNACNAAYCAYAARRAYATYGG (Yu *et al.* 2012) and Ill_C_R: 5' GGIGGRTAIACIGTTCAICC (Shokralla *et al.* 2015). Similarly, the *bc3'* fragment corresponding to ≈420 bp of the 3' end of the *cox1* barcode was amplified with primers Ill_B_F (5'-CCIGAYATRGCITTYCCICG) (Shokralla *et al.* 2015) and Fol-degen-rev (5'-TANACYTCNGGRTGNCCRAARAAYCA) (Yu *et al.* 2012). The *SSU* marker was amplified using primers SSU-FO4 (5'-GCTTGTCTCAAAGATTAAGCC) and SSU-R22 (5'-GCCTGCTGCCTTCCTTGGA), producing a fragment of varying length of 300 to 400 bp (Blaxter *et al.* 1998).

Primers were modified to include an overhang adapter sequence for subsequent nested PCR, in analogy to the Illumina protocol for sequencing the 16S rRNA gene in microbial samples (16S Library Preparation Protocol at http://support.illumina.com) (see Arribas *et al.* 2016). For each sample, three independent reactions for each pair of primers were performed, and the PCR amplicons were pooled. All information regarding PCR reagents and conditions was included in Data S1. Amplicon pools were cleaned using Ampure XP magnetic beads, after which these primary amplicons were used as template for a limited-cycle secondary PCR amplification to add dual-index barcodes and the P5 and P7 Illumina sequencing adapters (Nextera XT Index Kit; Illumina, San Diego, CA, USA). For each sample, the three gene fragments were processed individually but using the same indexes for sample tagging, thus combined in a single library and reducing the costs. The 24 resulting metabarcoding libraries were sequenced on an Illumina MiSeq sequencer (2x300 bp paired end reads) on 1.5% of the flow cell each, to produce paired reads (R1 and R2) with a given dual tag unique combination for each sample.

Creating a reference sequence database

A custom reference set of OTUs for Arthropoda was created from sequences obtained from the BOLD Public Data Portal (Ratnasingham & Hebert 2007; accessed on 8th January 2017). All available full-length (658 bp) *cox1* sequences for Arthropoda were retrieved from BOLD and subsequently clustered with Usearch v7 (Edgar 2013) under a 3% similarity threshold. This resulted in thousands of OTUs (referred to as *BOLD-OTUs* from hereon), each of which was based on variable numbers of primary entries in the BOLD database, ranging from just a single sequence to several hundred sequences in some cases. For simplicity, the *centroid* sequence of each *BOLD-OTU* was used as the "representative sequence" in subsequent analyses for the taxonomic identification of metabarcoding sequences. The BOLD database already provides clustering of barcode data based on a graph theory method that produces the so-called BIN (Barcode Identification Number) groups (Ratnasingham & Hebert 2013). We established the correspondence of our *BOLD-OTUs* with the BINs based on the representative sequences, which permitted to attach a BIN number and, where available, the associated species name to the *BOLD-OTUs.* We obtained species names for most of the OTUs that were matched by the metabarcoding study, but there were 18 cases where metabarcoding reads matched a single sequence on BOLD that was not attached to any named BIN group and which was identified to order level only.

In general, each *BOLD-OTU* corresponded to a unique Linnaean species name, but in several cases the same species name was attached to multiple *BOLD-OTUs*, indicating high intraspecific genetic diversity, identification problems in the reference database, or the existence of cryptic species (Table 1, Suppl. Table 4). The BINs on the BOLD database equally are affected by splitting of Linnaean species. For example, sequences associated with the isopod *Asellus aquaticus* were assigned to eight separate *BOLD-OTUs* matching 8 different BINs. High intraspecific variation (>3% divergence in *cox1*) is a well-established observation in the case of *Asellus* (Sworobowicz *et al.* 2015). Three *BOLD-OTUs* were assigned to *Baetis rhodani* (Ephemeroptera), which is also reflected in the incomplete taxonomy of this species complex (Williams *et al.* 2006; Bisconti *et al.* 2016).

Bioinformatic read processing

Various bioinformatics steps were applied to reduce the proportion of low-quality data (Schirmer *et al.* 2015). These steps included the trimming of 3' ends, merging R1 and R2 reads, and the detection and removal of hybrid molecules formed during PCR from mixed templates. Raw reads were quality checked in Fastqc (Babraham Institute 2013) and subsequently de-multiplexed to get independent datasets for each of the three DNA fragments, using the *fastx_barcode_splitter.pl* option of the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/; accessed 18/07/2016). Primers were

trimmed using *fastx_trimmer* and reads were processed in Trimmomatic (Bolger *et al.* 2014) using TRAILING:20 MINLEN:250 CROP:270 (CROP 250 for R2). R1 and R2 reads were paired using Pairfq 0.16 (Staton 2013) and the *makepairs* option.

Alternative procedures for read denoising, pair merging, quality filtering and clustering method were tested for each DNA marker (Suppl. Fig. S2). These analyses included: (i) Four different denoising parameters using BFC (Li 2015) (s=0.35 K=33; s=2 K=33; s=20 K=33; without de-noising); (ii) two procedures for pair merging using either Pear v0.9.6 (Zhang *et al.* 2014) with *–q 26* and other default parameters *or* Usearch v7 (Edgar 2013) as above; and (iii) three alternative quality filtering parameters in Usearch (Maxee=0.5; Maxee=1; without Maxee). Processed sequences were OTU clustered with three different algorithms: Usearch (greedy heuristic approach; Edgar 2013) (-*cluster_otus* option); CROP (Bayesian approach; Hao *et al.* 2011); and Swarm (agglomerative approach; (Mahé *et al.* 2014) (Suppl. Fig. S2).

The combination of 24 read processing settings and 3 OTU clustering methods yielded a total of 72 OTU sets for each of the three metabarcode fragments. For each DNA fragment (bc5', bc3', *SSU*) we estimated the number of exclusive and shared OTUs between each pair of the 72 OTU sets obtained. This required that we possess a list of all OTUs present in the various sets. We assembled the 72 OTU sets with a minimum similarity threshold of 3% in Geneious v7.1.9 (Kearse *et al.* 2012). The resulting assemblies were exported as a 50% majority rule consensus sequence to represent each OTU, and used as references to subsequently map the OTUs obtained for each of the 72 OTU sets under a 3% similarity threshold, using the command *–usearch_global*. The proportion of shared and exclusive OTUs in each pair of the 72 OTU sets were estimated from the OTU table using R.

Based on the results from these tests (below), we used the following parameters for all further analyses: reads were denoised using BFC (Li 2015), and processed following several steps of the Usearch (Edgar 2013) pipeline: reads were merged (option *mergepairs –fastq_minovlen 150 (130 for bc3')*, *-fastq_maxdiffs 30*), quality filtered (Maxee=1), dereplicated (*-derep_fulllength*) and sorted (-*sortbysize* options). Sequences with only one read (*-minsize 2*) were excluded, and a *de novo* chimera checking was conducted (*-uchime_denovo* option).

De novo OTU generation from metabarcoding

Quality-filtered metabarcode sequences were subjected to clustering with each of the three clustering algorithms using 3% and 10% dissimilarity thresholds for the *cox1* bc5' and bc3' fragments, and 3% threshold for the *SSU* gene fragment. Each OTU set was filtered to retain only OTUs, which corresponded to the targeted invertebrates and to remove bacterial and other sequences. For this

purpose, Geneious was used in batch mode to align each OTU set (the representative sequence, i.e. the centroid according to Usearch) with MAFFT options FFT-NS-2 (Katoh & Standley 2013) and to generate an UPGMA tree based on Tamura-Nei distances. Only the OTUs included in the largest clade of the UPGMA tree within a 30% dissimilarity threshold were retained, and all other sequences presumably not representing the targeted gene fragment were removed with a custom R script using the libraries *ape* (Paradis *et al.* 2004), *rncl* and *stringr* (Wickham 2013). Finally, the retained *cox1* sequences were aligned in Geneious using MAFFT and the Translation Align option, and subsequently sequences with insertions, deletions or stop codons disturbing the reading frame were excluded.

Identification of OTUs against NCBI and the *BOLD-OTUs* reference set

Assignments of OTUs to high level taxonomic categories were conducted with the lowest common ancestor (LCA) algorithm implemented in Megan v5 (Huson *et al.* 2007). Each OTU representative sequence was subjected to BLAST searches against the NCBI *nt* database (December 2016; *blastn -outfmt 5 -evalue 0.001*). BLAST matches were fed into Megan to compute the taxonomic affinity of each OTU. We accepted the taxonomic ranks in the NCBI Taxonomy database (December 2016) and Megan was used to estimate richness for phyla within Metazoa, classes within Arthropoda, and orders within Insecta and Crustacea.

Secondly, the identifications were conducted against *BOLD-OTUs.* Analyses were based either on metabarcoding reads only, after filtering, or after the *de novo* OTU clustering step described earlier. Processed reads of *bc5'* and *bc3'* were matched against *BOLD-OTUs* using the *-usearch_global* option with the same 3% threshold. The python script *uc2otutab.py* was used to generate a list of matched OTUs. For the identification of *de novo* OTUs from clustering with Usearch, Swarm and CROP the same protocol of matching the *BOLD-OTUs* in the reference database was applied to the representative sequences from each OTU (at 3% threshold).

Community composition and indicative species analyses

Ecological statistical analyses for total beta diversity and the associated turnover and nestedness components (Baselga 2010) across sampling sites were conducted for the Metazoa, Arthropoda, Insecta and Crustacea sub-datasets. Community composition tables (OTU x site tables) were obtained by matching Illumina processed reads (*blastn -outfmt 5 -evalue 0.001*) against (a) the selected sets of *de novo* OTUs obtained with Usearch (3% and 10% similarity thresholds for *bc5'* and *bc3'*, and 3% for *SSU);* and (b) the *BOLD-OTUs* reference dataset (for *bc5'* and *bc3'* gene fragments, only for

Arthropoda, Insecta and Crustacea sub-datasets). When using *de novo* generated OTUs, taxonomic assignments obtained in Megan (see above) were used to extract sub-datasets of OTUs identified as (a) Metazoa, (b) Arthropoda, (c) Insecta and (d) Crustacea with an LCA parameter value of 90. Likewise, the *BOLD-OTUs* reference sequences were taxonomically assigned to Arthropoda, Insecta and Crustacea sub-datasets.

Taxonomic subsets of data were used to conduct analyses of community composition, either for the meio- and macro-fauna samples separately or after combining both samples from the same site and time as a single sample. Distance matrices by pairs of sites were generated for total beta diversity (Sorensen index), turnover (Simpson index; species replacement, without the effect of variation in richness) and nestedness (Sorensen - Simpson index; pure richness effect) using the R library *betapart* (Baselga & Orme 2012), and the R library *vegan* (Oksanen *et al.* 2013) was used to perform Nonparametric Multidimensional Scaling analyses (NMDS). Analyses were plotted with the *ordisurf* option to generate clines based on richness values and *ordispider* to connect the samples from upstream and downstream of the spill. Finally, a Permutational Anova (Permanova) was conducted using the function *adonis*, and the significance of differences was assessed using a stress test and the *envfit* test of correlation. Jaccard distances were used to verify the existence of significant differences in the communities upstream and downstream of the spill. We additionally used the *mvabund* package for the analysis of multivariate abundance data (Wang *et al.* 2012) to test upstream and downstream effects (factor *spill*) on community composition (*manyglm* function) and to identify which individual species show significant differences in their distribution between control and impacted sites (*anova.manyglm* function; test=LR, nBoot=999) using the results from the reference database (taxonomy-dependent) approach. Finally, we used Dufrene-Legendre indicator species analysis (*indval*) (Dufrêne & Legendre 1997) using the R package *labdsv* (Roberts 2007) to compare control versus impacted sites.

RESULTS

Sequencing, read processing and OTU clustering

OTU delimitation was conducted on tens of thousands of Illumina (2x300 bp) sequence reads per sample from both the meio- and macro-fauna extracts (ranging 41,306 - 177,653 for *bc3'*; 5,487 - 100,344 for *bc5'*; and 34,555 - 107,016 for *SSU*) (Suppl. Table S1). A single sample (Impacted site 1-Time 1: IS1-T1) failed to produce a PCR product and was removed from the analysis. For the remaining 11 samples, Spearman's *rho* correlation tests revealed no correlation between raw read numbers and the number of OTUs classified as Metazoa for the three DNA fragments (Suppl. Fig. S3), i.e. different read abundances between samples did not affect the number of OTUs recovered.

The raw reads were subjected to basic quality filtering, which reduced the usable reads to about 90% of the initial reads for the *SSU* marker, while for both *cox1* markers in many samples a much greater proportion was lost, sometimes retaining only 10-20% of reads. There were no systematic differences in total read numbers after quality filtering between upstream and downstream or between meio- and macro-fauna samples (Suppl. Table S1). The paired-end merging of R1 and R2 (forward-reverse) reads resulted in a further reduction, but was broadly uniform across samples (Suppl. Table S1). Various modifications of the basic protocol for quality filtering and read merging under 24 different parameter settings were applied. The resulting merged reads were then subjected to OTU clustering using three different clustering methods (Suppl. Fig. S2). After excluding OTUs not matching the expected gene fragments or taxonomic groups using a tree-based method (see Materials and Methods), the number of OTUs under these various parameter settings ranged from 479 to 810 for the *bc5'* fragment, from 543 to 1150 for *bc3'*, and from 193 to 590 for *SSU* (Suppl. Table S2).

The proportion of shared and exclusive OTUs obtained (for each gene) showed broadly similar results across the parameter settings (Fig. 2). The OTUs obtained with different clustering methods (Usearch, CROP and Swarm) were shared in 91.1% and 92.7% of the cases for the *bc5'* and *bc3'* fragments, respectively, with the major difference attributed to the CROP clustering method that lacked 4.6% of OTUs obtained with the other methods. The same trend was observed for *SSU*, where a 65.6% of the OTUs were generated with the three methods and an additional 20.3% were shared between Usearch and Swarm, but not with CROP. The implementation of a denoising tool (BFC software) resulted in a moderate effect on the number of obtained OTUs (~90% of OTUs shared), except for *s*=20 which resulted in the absence of 30%-50% of the OTUs. The use of the Maxee filtering option also produced a moderate effect on OTU recovery (~90% of OTUs shared among parameters settings). Finally, using PEAR or Usearch for the merging step resulted in ~90% shared OTUs (Fig. 2).

Taxonomic profiles based on *de novo* OTUs using Megan

Next, *de novo* OTUs were identified against the Genbank database via LCA assignment to major taxon (see Materials and Methods). We only assessed the OTU set obtained under a single representative parameter setting (BFC with *s*=0.5 for read denoising, *Maxee=1* for read filtering, and *mergepairs* in Usearch for read merging), to which we applied the three clustering methods (Usearch, Swarm, CROP) under similarity thresholds of 3% and 10% for the *cox1* fragments, and 3% for the *SSU* fragment (Fig. 3). In the *bc5'* and *bc3'* fragments, the LCA assignment showed that samples were dominated by OTUs of Arthropoda, followed by Rotifera and Annelida. Within Arthropoda, Insecta was the most abundant group, followed by Arachnida and various classes within the subphylum Crustacea (Fig. 3). The Insecta were dominated by Diptera, followed by Coleoptera, Ephemeroptera and Trichoptera. In the *SSU* data, the OTU composition was also dominated by Arthropoda but the

total number of OTUs was substantially lower compared to those from *cox1*, in particular in the Insecta and specifically the Diptera, which was reduced to approximately 1/4th of the *bc3'* and *bc5'* OTUs (Fig. 4; Suppl. Table S3). In Crustacea, both genes detected the major orders such as Podocopida, Isopoda, Amphipoda, Cyclopoida and Diplostraca. OTU numbers were reduced in *SSU* compared to the *cox1* marker in Annelida, and Rotifera were not detected with *SSU* at all, despite the recovery of several dozen OTUs with *cox1*. In contrast, the *SSU* dataset retrieved a similar number of OTUs of Mollusca, while for Platyhelminthes and Nematoda most OTUs were only detectable with the *SSU* marker (Suppl. Table S3). Taxonomic profiles obtained at a 10% similarity threshold showed the same trend as described above, but with a substantial reduction in the total number of OTUs (Fig. 4; Suppl. Table S3).

Taxonomic assignment against the *BOLD-OTU* reference database

Direct mapping of sequence reads against the reference set resulted in matches to 207 *BOLD-OTUs.* The majority of these *BOLD-OTUs* matched a unique BIN in the BOLD database (Table 1 for Diptera and Suppl. Table S4 for the complete dataset). Eighteen of these 207 *BOLD-OTUs* included a single sequence, but were not included by BOLD in their BIN system. In two cases, a pair of *BOLD-OTUs* produced separate clusters of sequences from a single BIN, and data for these were lumped, to maximise consistence with BOLD. Read numbers for each OTU varied by four orders of magnitude, to a maximum of >20000 reads for some species of Baetidae. Out of a total of 207 *BOLD-OTUs* with read matches, 36 OTUs were obtained exclusively with the *bc5'* fragment and 46 OTUs with the *bc3'* fragment, but in most of these cases only a few reads (<10) were obtained.

Next, we mapped the *de novo* generated OTUs against the *BOLD-OTUs*. Species detections obtained in this way closely matched those from the read mapping, although the method was slightly less sensitive in cases of low number of reads that did not produce an OTU (Table 1, Suppl. Table 4). Species detection depending on the three clustering methods (Usearch, Swarm, CROP) or marker (*bc5'* and *bc3'*) was generally consistent, and observed differences tend to occur for species with low read counts (Fig. 4; Suppl. Table S4).

The total numbers of identified OTUs from read matching and *de novo* clustering were generally similar, in particular in Diptera for both *bc5'* and *bc3'*. In other taxonomic groups, the OTU counts from read matching were closer to those at the 10% thresholds of the *de novo* OTU analysis or even lower, as in the case of several classes and orders within Crustacea (Fig. 4; Suppl. Table S3), whose

reference databases were less complete. The reduction at the 10% threshold generally affected cases that were split into multiple OTUs at 3%; for some of these cases, the sequence reads also matched several *BOLD-OTUs* but based on taxonomic assignments were identified as the same species, e.g. some of the eight *BOLD-OTUs* identified as the isopod *Asellus aquaticus* (Fig. 4; Suppl. Table S4).

Comparisons with the taxon list of macroinvertebrates provided by traditional analysis (Thompson *et al.* 2016) were made for Arthropoda based on identifications by read mapping with the *bc5'* and *bc3'* (*cox1*) fragments at the 3% threshold level. The identifications included many exact matches to the Thompson *et al.* (2016) list, although the latter included only 38 arthropod taxa in total, of which 11 taxa were not represented in the OTUs dataset (not even at higher taxonomic level). The missing taxa were from a range of higher taxa, including Ephemeroptera, Plecoptera, Coleoptera, Isopoda, Arachnida and others, while other species in these taxa were easily recovered by the *cox1* metabarcodes. In many cases, Thompson *et al.* (2016) did not separate the entities at the species level. Notably the Chironomidae, which were listed as two taxa (Chironomidae and subfamily Tanypodinae) in Thompson *et al.* (2016) were split into a total of 55 OTUs in the current analysis and most could be identified to species. Similarly, the single entry for Limnephilidae in Thompson *et al.* (2016) was split into 5 OTUs, the entry for Dytiscidae into 3 OTUs, and even genus-level entries were split further, e.g. *Baetis sp.* corresponded to 9 OTUs and *Simulium sp.* corresponded to 4 OTUs in the *cox1* analysis. In other cases, several *BOLD OTUs* were assigned the same binomial (although not unequivocally in all cases; see Suppl. Table 4), indicating the high intraspecific diversity on some species or even the possible existence of cryptic species.

Effect of the spill on community composition and indicative species

The presence or absence of species in various metabarcoding samples was used to establish the turnover and nestedness of assemblages above and below the spill site. NMDS analyses for total beta diversity, nestedness and turnover across sampling sites were conducted for the Metazoa, Arthropoda, Insecta and Crustacea datasets for: (a) OTUs at 3% and 10% for *bc5'* and *bc3';* (b) OTUs at 3% for *SSU;* and (c) for the BOLD reference dataset and *bc5'* and *bc3'* gene fragments (summarized in Table 2). Significant differences in community composition were detected in all cases between impacted and control sites that were mainly driven by a high turnover of species, but not for the nestedness component of beta diversity (Fig. 5, S4-S9). Results were very similar for *de novo* clustering with BLAST+Megan for OTU classification and reads based approaches using *BOLD-OTUs* for identification (Fig. 6, Table 2). The NMDS plots conducted on the meio- and macro-faunal fractions separately also showed very similar patterns (Fig. 7).

Of the 207 OTUs detected by matching the reads to *BOLD-OTUs*, 170 were obtained from the meiofauna sample, against only 126 OTUs from the macrofauna fraction (filtration at >1 mm), and 70 OTUs were unique to the meiofauna. There were clear patterns in the distribution of meio- and macro-faunal samples, for example showing that Brachiopoda, Collembola and some Arachnida were almost entirely found in the former, reflecting their small body size, but most large-bodied insects were also recovered in the meiofauna fraction (in addition to the macrofaunal fraction). Second, combining the total species count of meio- and macro-faunal fraction, the control sites had slightly more species compared to the impacted sites but both had a high proportion of unique OTUs (65 versus 53)

Indval analyses for strong effects of the spill revealed a statistically significant indicative value of *Gammarus pulex* associated to control sites (indval $p<0.05$), while *Asellus aquaticus* and several Chironomidae species (*Tanytarsus eminulus, T. brundini, T. pallidicornis, T. ejuncidus, Cricotopus bicinctus, Paratendipes albimanus*) were associated to impacted sites (indval $p<0.05$) (Table 3). *Mvabund* analyses resulted in significant differences for *Asellus aquaticus* and *Tanytarsus eminulus* ($p<0.05$). Additional tests conducted with the *indval* function to identify indicative species associated to the two sampling periods resulted in the Ephemeroptera *Centroptilum luteolum* as indicative of sampling period 2 (2.5 months after the spill), whereas when the factor spill and the sampling period were considered in combination, two Chironomidae (*Eukiefferiella claripennis* and *Cricotopus bicinctus*) and one Branchiopoda (*Chydorus sphaericus*) were found as indicative of the impacted sites at the sampling period 2.

DISCUSSION

Metabarcoding revealed high species turnover between control and pesticide-impacted sites in the River Kennet, in line with the earlier study based on conventional taxonomic assignments (Thompson *et al.* 2016), but with far more complete and taxonomically resolved data. This was also achieved at a fraction of the cost and time, although absolute abundance data were sacrificed as a result (note that most routine biomonitoring approaches only ever use relative abundance anyway; Friberg *et al.* 2011; Gray *et al.* 2016). The high read depth and accuracy of Illumina sequences, in addition to an increased completeness of reference databases, now can determine the identity of taxa as effectively as conventional biomonitoring, but with the added advantage of capturing data at the species level for virtually all components of the sample – including both the temporary and permanent meiofauna, which are very diverse and abundant yet are routinely ignored due to difficulties in their identification via traditional microscopy (Friberg *et al.* 2011; Gray *et al.* 2016). Our study takes a further step towards this explicit species level approach, by addressing two key issues: the selection of appropriate gene regions and universal primers for metabarcoding, and the choice of the most appropriate bioinformatics tools and parameter settings. If linked with existing DNA databases, metabarcoding

provides the species level resolution of community composition, to identify the indicators of disturbed and undisturbed sites and, ultimately, to bridge biomonitoring of community structure and ecosystem functioning.

Methodological decisions: choice of gene markers and bioinformatics

Researchers are confronted with a wide range of options for data processing and parameter settings at all stages of metabarcoding. The final outcome is affected by the early steps of the wet lab procedures for DNA extraction from various environmental samples and the choice of PCR primers. The two genetic markers used here amplifying portions of the mitochondrial *cox1* and the nuclear 18S rRNA genes illustrate the important effects of marker choice, affecting: amplification of major taxonomic groups, power of separation of species, and the completeness of available reference data. Based on these criteria, the barcode fragment of the *cox1* gene remains the most powerful choice for metabarcoding studies of animal communities. The advantage of using the *cox1* gene is due to (i) the benefits from an already available and growing database for this marker from standard barcoding; (ii) the widely accepted standardization of the marker choice, allowing for a comparison of different studies; and (iii) the demonstration that universal primers recover complex communities composed of a taxonomically wide range of arthropods and other metazoans.

The primers used here for both *cox1* fragments produced rather similar results for total OTU numbers and taxonomic distribution of major groups, indicating that amplification from mixtures is largely reliable. Differences between OTU detection with both primers are mainly at the level of individual species, which requires further investigation, but possibly is due to variation among individual PCR on low-abundance (low biomass) species, as reflected in the fact that disagreements are mostly due to OTUs represented by low numbers of reads. In contrast, the *SSU* (rRNA) marker is far less variable than mitochondrial DNA and thus presumably closely related species are collapsed into single OTUs (Tang *et al.* 2012). This was evident from the lower number of OTUs obtained, in particular for groups containing multiple congeneric species, such as the Chironomidae. However, in addition to the lumping of close relatives, the overall taxonomic profiles also differed strongly, presumably due to the greater conservation of *SSU* primer binding sites across phyla. Specifically, Platyhelminthes and Nematoda were only recovered with *SSU*, whereas the efficiency was lower for Arthropoda. The former groups are rarely if ever considered in traditional biomonitoring, whereas arthropods underpin most schemes around the world. These well-known issues of taxonomic resolution and differences in the spectrum of PCR amplification breadth need to be considered carefully depending on the target taxa. It is unlikely that truly universal primers can be found for amplification of all animal phyla or that a single locus can separate species universally. Even so, poor amplification of some groups for *cox1* possibly can be alleviated with different primer sequences, at least for regions that are partly

overlapping, using multiple PCRs on the DNA extracts, to be included in the Illumina libraries with relatively little added effort. An alternative is the use of slightly more conserved markers, such as the mitochondrial rRNA (12S and 16S) genes suggested in recent metabarcoding studies, which present a compromise between broad, unbiased PCR and species-level differentiation (Taberlet *et al.* 2012), but this approach misses the advantage of comparisons against the available sequence databases for the barcode *cox1* gene fragment.

For the bioinformatics processing, different parameter settings had little effect. Different clustering algorithms employing fundamentally different methodology generated largely uniform OTUs, except perhaps the CROP algorithm, which detected slightly fewer OTUs than the Usearch and Swarm methods. Equally, the first step of denoising was robust over a range of parameters that are broadly within the boundaries of most published applications of the software. The read pairing step and the final filtering based on the error frequency in the reads (the maximal expected error, Maxee) only changed the outcome of the final assembly for <10% of OTUs. Presumably, the denoising procedure mainly eliminates minor variants among the reads that are subsumed into the OTUs at the minimum clustering threshold of 3% that was applied here, and thus their prior elimination has little impact on the final outcome of the OTU delimitation. We also find that many OTUs are generated that are either non-mitochondrial or correspond to non-target taxa, including bacteria. These were removed via a phylogenetic approach using rapid UPGMA trees and retaining only the major target clade for metazoans, thus avoiding later complications with species counts and turnover analysis. As a general rule, stringent parameters settings should be applied, but the appropriate parameter space is fairly wide and so its reassessment is not necessary for each study; for comparability, a particular parameter set should be chosen and applied consistently.

Species definitions in *de novo* clustering and read mapping

A key aspect of the metabarcoding procedure is the recognition and identification of the species in the sample, which requires a valid taxon concept for each species based on solid data for species circumscription (Wheeler 2004). The clustering algorithms define the species limits based on similarity thresholds, akin to the search for a barcoding gap in standard (Sanger) barcoding (Meyer & Paulay 2005), but the actual variation is overlain by sequencing read errors, and true variants from variable mitochondrial copies (heteroplasmy) and nuclear mitochondrial insertions (numts). While the various clustering algorithms broadly agreed on the number and extent of OTU delineations, differences in particular species hypotheses were evident and dependent on the threshold. If set to 3% in the *cox1* marker, OTU delineation largely reflected the species numbers expected from the Linnaean taxonomy, whereas the highly conservative 10% threshold lumped many species. The application of more refined algorithms to the OTU circumscriptions is desirable, although coalescence

based procedures including GMYC and PTP (Pons *et al.* 2006; Zhang *et al.* 2013) currently are not easily applicable to the very large number of reads. In addition, metabarcoding sequences are comparatively short, which limits the resolving power of these sequences in any species delimitation procedure.

The two principal approaches used here for generating community OTU lists, i.e. the 'taxonomy independent' approach based on *de novo* generated OTUs, and the 'taxonomy dependent' approach via read mapping against the reference database (Schloss & Westcott 2011) produced generally similar results (although with slightly fewer entities detected in the latter). When working with *de novo* generated OTUs, special attention should be paid to the method used for classifying OTUs. Megan has been proven as a useful tool based on Blast searches against reference databases, as here conducted. Nevertheless, incompleteness of reference databases can result on inaccurate classifications when relaxed similarity thresholds are considered (e.g. 70%), whereas higher similarity stringency (about 90%) will result in many OTUs with no taxonomic assignation. This is here illustrated by the classification of one OTU retrieved independently with both *cox1* fragments as Echinodermata, a marine group not expected to be present in the target sample. This identification corresponds to a few matches with similarity below 80% and the sequence requires several indels to align to the identified reference. For sequences with such a weak match, Megan may use similarly weak matches with several other reference sequences from distantly related taxa, to return an identification at the deepest taxononomic level corresponding to all of these weak matches. If the focal sequence has no close match in the database, spurious Blast similarities might indicate a best match in a certain area of the reference database, despite representing a lineage quite divergent from these best matches, and thus this sequence is not necessarily related to Echinodermata.

In the read mapping approach, the problem of species delimitation is shifted to the reference databases that define intraspecific variation and link a set of sequences to a Linnaean binomial. In the European freshwater arthropods we could heavily draw on the relatively good species representation in BOLD. For simplicity, we downloaded all available sequences and conducted OTU clustering on these sequences, in analogy to the metabarcode data, to produce a reference database (the *BOLD-OTUs*), after which only those with similarity to the metabarcoding sequences encountered in the target samples were considered further (a total of 207 OTUs). The BIN groups presented in the BOLD database are an alternative compilation of the existing reference data, and we established that they produce similar OTU circumscriptions as those from the Usearch clustering at a 3% similarity threshold. Both compilations show a good general congruence with the Linnaean names, but in several cases multiple BINs and *BOLD-OTUs* were associated to a single binomial, probably reflecting high intraspecific variation in some species whose real species limits need to be evaluated with additional information. The relevance of these closely related OTUs remains unclear but may be related to the presence of multiple divergent mitogenome copies in the focal species. For example,

some isopods, including *A. aquaticus*, exhibit atypical mitogenomes composed of duplicated regions that apparently are maintained constitutively as heteroplasmic copies (Doublet *et al.* 2012). Other cases, such as the ephemeropteran *Baetis rhodani* or the amphipod *Gammarus pulex*, are already well established species complexes consisting of multiple cryptic species (Karaman & Pinkster 1977; Williams *et al.* 2006; Rutschmann *et al.* 2014; Bisconti *et al.* 2016). Taxonomic difficulties associated to these cases may affect the reference sequence databases, such the detection of the oriental *Gammarus nekkensis* with the *bc5'* fragment, whereas *G. pulex* is exclusively found with *bc3'*. It should be noted that *G. pulex* is represented by 5 different BINs at BOLD, whereas *G. nekkensis* forms up to 10 BINs (accessed 13-06-2017). This taxonomic and molecular complexity highlights the need of a careful assessment by expert taxonomists and molecular biologist for consensus species delineation. Yet, beyond these taxonomically complicated cases, the existing barcode database already holds an excellent coverage for European aquatic macroinvertebrates.

Whereas the direct read mapping approach could circumvent the problematic step of OTU clustering, maximising comparability between different studies, OTUs from the *de novo* clustering could still be useful in particular to identify species not yet embedded in reference databases, albeit without link to a Linnaean name. The robustness of the *de novo* approach shown here therefore offers a defensible option for species assemblages whose coverage of reference sets is still patchy, and could be used in biodiversity discovery in hitherto unstudied ecosystems. Ideally, standardized protocols should use a combined approach, where direct mapping for biomonitoring is complemented with *de novo* OTU clustering under defined parameters and a comparison of OTUs with reference databases. This can help to identify gaps in the reference database, through an iterative process, contributing to refine local reference databases. Once a validated reference set is in hand, the straightforward mapping of sequence reads against these DNA-based grouping provides easily repeatable and stable species identification for biomonitoring. This provides a reliable link to taxonomy and resolution to species level even in the most problematic taxa, such as Chironomidae, which can be used to improve conclusions reached by river monitoring programs, and subsequently to improve conservation and management practices (e.g., Chironomidae richness can be stable while high turnover happens after an impact). In future, through the direct link with taxonomic identifications, metabarcoding can then be connected to traits databases available for a large proportion of aquatic invertebrates species in Europe (Tachet *et al.* 2002; Schmidt-Kloiber & Hering 2015), to bridge the critical gap that still exists between structural biodiversity and functional measures related to ecosystem processes (Friberg *et al.* 2011; Woodward *et al.* 2013).

Implications for biomonitoring and ecological studies

The total diversity of identified species in our metabarcoding study far exceeds the number of invertebrate species detected using conventional analyses (Thompson *et al.* 2016) - by at least a factor of 5 - even under the most conservative OTU detection parameters. Three aspects contribute to the high numbers: (i) the greater taxonomic resolution compared to the morphological analysis that was conducted on higher taxonomic levels, especially for the Chironomidae; (ii) the split of binomial species names into multiple molecular OTUs (at the 3% threshold in *cox1*); and (iii) the better detection of minute specimens, especially those in the temporary meiofauna, in part representing early instars of otherwise larger-bodied species, which are often overlooked in visual assessments, even under light microscopy. These factors thus resolve some of the drawbacks of conventional techniques, namely the incomplete species identification, poor separation of cryptic diversity, and incomplete sampling of the freshwater assemblage. The study also highlights some of the remaining challenges of generating complete metabarcoding inventories: the problem of lumping and splitting of Linnaean species, the low primer efficiency for particular taxa, the variation among PCR runs, and the sampling itself which is affected by stochastic error. Our sample consisted of two independent Surber samples each, from three sites above and three below the impact zone. The methodology is now sufficiently well developed to be applied to many more samples, and denser sampling may reveal additional species that show a clear response to environmental conditions. Ultimately these methods should be applicable in a highly consistent manner for regulatory purposes, perhaps after matching these data against existing schemes and indexes for assessing water quality. For instance, the 600+ pre-defined reference stream sites used for the UK-wide RIVPACS biomonitoring scheme would offer an ideal testbed for this 'next-generation biomonitoring' approach (Bohan *et al.* 2017), as would the UK Acid Waters Monitoring Network species-level data that now span several decades and multiple standing and running waters (e.g. Gray *et al.* 2016).

Our analyses confirmed the previously observed shifts in community composition, despite the variation among individual samples, as the ordinations clearly separated samples from the control and impacted sites. This can be discerned at various taxonomic levels, for different arthropod classes, for meio- and macro-fauna, and for the communities established with each of the three markers. The detailed species list now complements this broad-scale information with the traditional approaches, and in particular it shows that the most resilient *r*-selected taxa, such as chironomids, recover most rapidly to occupy vacant niches following the crashes in *K*-selected taxa. Our study confirms that this increase is indeed an increase in species richness, not just in abundance of species present already at lower density. The conventional approach, as is common in freshwater ecology, simply lumped these highly responsive and speciose taxa into a single entity that revealed marked increases in abundance in the absence of potential competitors and predators, but provided no information on chironomid biodiversity. Most of the taxa in the impact samples are orthoclads (grazers on stones and plant

surfaces), with a few others that are detritivores living in soft sediments, in addition to some predatory species. The increase of several species of *Tanytarsus*, a group of dominant sediment-dwelling detritus feeders that have been shown to feed on diatoms in the early larval stages (Ingvason *et al.* 2004), is consistent with the increase of several large diatom species observed in the post-spill sites (Thompson *et al.* 2016). These preliminary data suggest that inferences about the ecological response in terms of both impacts and resilience is masked by limited taxonomic resolution, and that this in turn is likely to reflect marked functional trait shifts that are overlooked in routine biomonitoring schemes. Biomonitoring at present is focused on responses to organic pollution or, to a lesser extent, acidification, whereas responses to other stressors are still poorly characterised in natural systems: the next-generation biomonitoring approach we use here could open the door to improving the sensitivity and power of detection in relation to both response variables and a wider range of drivers than is currently possible..

Conclusions

Our study refines the parameter space of metabarcoding studies generally, and our specific case study highlights its potential for next-generation biomonitoring to advance the current state-of-the-art assessment of water quality and ecological status. At least in terms of detecting relevant changes in community, neither the marker, the read processing or the clustering method or threshold affected our ability to detect the spill's impact on the community. The availability of full species-level inventories for the first time enabled us to exploit the extensive ecological databases that are now available for freshwater species in Europe, and also to begin to elucidate relevant trait differences. In addition, the capacity to use all taxa, rather than a narrow subset for which taxonomic expertise is available, promises to deliver a far more informative and mechanistic understanding of biodiversity in freshwater ecosystems and its responses to environmental stressors.

Acknowledgements

This study was funded by NERC grant NE/M021955 to DYW, APV and KB. We are grateful to Steve Brooks, Benjamin Price and Tjorborn Ekrem for discussions about species responses in freshwater communities.

DATA ACCESSIBILITY

Raw metabarcode data deposited in Dryad: doi:10.5061/dryad.104kg

AUTHORS'S CONTRIBUTIONS

APV and GW conceived the study; CG and GW collected the samples; CA, PA, DWY and APV designed the analyses; CA and PA did the molecular lab work and performed the analyses; CA and APV led the writing and all author contributed to the discussion of the results and the final writing of the manuscript

SUPPLEMENTARY MATERIALS

1. Information on PCR reagents and conditions
2. Supplementary figures
3. Supplementary tables
3. Code for metabarcoding pipelines, read processing and clustering

Table 1. BINs of Diptera from BOLD identified based on *usearch_global* searches under a 3% similarity threshold of (i) processed reads and (ii) OTUs clustered at 3% (details in text).

BIN	Family	Main species id of BIN	bc-5'	bc-3'	r-bc-5'	r-bc-3'	C	I	MC	MS	T1	T2
BOLD:AAJ7051	Agromyzidae	*Agromyza pseudoreptans [19]*	-/-/-/-	R/U/C/S	0	5		b3	b3		b3	
BOLD:ACI4790	Bibionidae	*Dilophus febrilis [21]*	R/U/C/S	R/U/C/S	478	428	b5 b3		b5 b3	b5	b5	b5 b3
BOLD:ACP0608	Cecidomyiidae	Cecidomyiidae *[3]*	R/U/C/S	R/-/-/S	455	31		b5 b3		b5 b3		b5 b3
BOLD:ACS1169	Ceratopogonidae	*Palpomyia flavipes [6]*	R/U/-/-	R/U/C/S	20	41	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ACD1957	Chironomidae	*Apsectrotanypus trifascipennis [9]*	R/U/C/S	R/U/C/S	81	1446	b5 b3	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ADE2432	Chironomidae	*Brillia bifida [6]*	R/-/-/-	-/-/-/-	1	0		b5		b5	b5	
BOLD:ACR1089	Chironomidae	Chironomidae	R/U/C/S	R/U/C/S	550	93	b5 b3	b5 b3	b5	b5 b3	b5 b3	b5 b3
BOLD:ACP8764	Chironomidae	Chironomidae *[13]*	R/U/C/S	-/-/-/-	130	0		b5	b5	b5		b5
BOLD:ACP6740	Chironomidae	Chironomidae *[60]*	R/-/-/-	R/U/C/S	11	9		b5 b3	b3	b5 b3		b5 b3
BOLD:ACP2182	Chironomidae	Chironomidae *[92]*	R/U/C/S	R/U/C/S	33	26		b5 b3		b5 b3		b5 b3
BOLD:AAW5799	Chironomidae	*Conchapelopia hittmairorum [3]*	R/-/-/-	R/-/-/-	1	4		b5 b3		b5 b3		b5 b3
BOLD:AAP5886	Chironomidae	*Conchapelopia melanops [11]*	R/U/C/S	R/U/C/S	1018	331	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ACQ3496	Chironomidae	*Conchapelopia pallidula [1]*	R/U/C/S	-/-/-/-	3	0		b5		b5		b5
BOLD:ACD1670	Chironomidae	*Corynoneura sp.*	R/U/C/S	R/U/C/S	1178	1263	b5 b3	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACT8698	Chironomidae	*Corynoneura sp. [6]*	R/U/C/S	R/U/C/S	21	72	b5 b3	b5 b3		b5 b3		b5 b3
BOLD:AAW5785	Chironomidae	*Cricotopus albiforceps [139]*	R/-/-/-	R/-/-/-	2	4	b5 b3			b5 b3		b5 b3
BOLD:AAF2345	Chironomidae	*Cricotopus annulator [19]*	R/U/C/S	R/U/C/S	30	14		b5 b3	b5 b3	b5		b5 b3
BOLD:AAP5931	Chironomidae	*Cricotopus bicinctus [1]*	R/-/-/-	-/-/-/-	60	0		b5	b5	b5		b5
BOLD:ACU8677	Chironomidae	*Cricotopus bicinctus [1]*	R/-/-/-	-/-/-/-	5	0		b5	b5	b5		b5
BOLD:AAI6018	Chironomidae	*Cricotopus bicinctus [123]*	R/U/C/-	R/-/C/S	5751	4759	b5	b5 b3	b5 b3	*b5 b3	b5	b5 b3
BOLD:AAT9677	Chironomidae	*Cricotopus bicinctus [27]*	R/U/-/S	R/U/C/-	123	1991	b3	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAM5377	Chironomidae	*Cricotopus rufiventris [289]*	R/U/C/S	R/U/C/S	8	44		b5 b3	b5 b3	b3		b5 b3
BOLD:AAA5299	Chironomidae	*Cricotopus sylvestris [64]*	R/U/C/S	R/U/C/S	4	88		b5 b3	b5 b3			b5 b3
BOLD:AAU2576	Chironomidae	*Cricotopus trifascia [5]*	R/U/C/S	R/U/C/S	119	126		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAE4568	Chironomidae	*Eukiefferiella claripennis [185]*	R/U/C/-	R/U/C/-	164	116		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACT0982	Chironomidae	*Heterotrissocladius sp. 2SW [2]*	R/U/C/S	R/U/C/S	48	234	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAX3566	Chironomidae	*Macropelopia nebulosa [13]*	R/U/C/S	R/U/C/S	29	58	b5 b3	b5	b5 b3	b5 b3	b5 b3	b3
BOLD:AAB8862	Chironomidae	*Metriocnemus eurynotus [18]*	R/-/-/-	R/U/C/S	3	18		b5 b3		b5 b3	b5 b3	
BOLD:AAD4167	Chironomidae	*Micropsectra atrofasciata [26]*	R/U/C/S	R/U/C/S	826	596		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAC7823	Chironomidae	*Micropsectra contracta [14]*	R/U/C/S	R/U/C/S	62	98		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAD1527	Chironomidae	*Micropsectra lindrothi [18]*	R/U/C/S	R/-/-/-	5	4		b5 b3	b5 b3			b5 b3
BOLD:AAC7552	Chironomidae	*Micropsectra pallidula [24]*	R/U/C/S	R/U/C/S	91	53		b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAI1530	Chironomidae	*Micropsectra sp. 5ES [33]*	R/U/C/S	R/U/C/S	61	1641		b5 b3	b5 b3	b5 b3	b3	b5 b3
BOLD:ACR0263	Chironomidae	*Microtendipes pedellus [8]*	R/U/C/S	R/U/C/S	12	19		b5 b3		b5 b3		b5 b3
BOLD:AAW0928	Chironomidae	*Nanocladius rectinervis [6]*	R/U/C/S	R/U/C/S	593	166	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAD8971	Chironomidae	*Orthocladius oblidens [330]*	R/U/C/S	R/U/C/S	11359	5138	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAM5389	Chironomidae	*Orthocladius rubicundus [119]*	R/U/C/S	R/U/C/S	98	64	b5	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAW5449	Chironomidae	*Orthocladius rubicundus [25]*	R/U/C/S	R/U/C/S	39	145		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACX3335	Chironomidae	*Paracladius quadrinodosus [2]*	-/-/-/-	R/-/-/-	0	5		b3		b3		b3
BOLD:ACT5340	Chironomidae	*Paracladopelma camptolabis [2]*	R/U/C/S	R/U/C/S	13	15	b5 b3	b3	b5 b3	b3		b5 b3
BOLD:AAW4635	Chironomidae	*Paratanytarsus dissimilis [12]*	R/U/C/S	R/U/C/S	35	90		b5 b3	b5 b3			b5 b3
BOLD:AAL3267	Chironomidae	*Paratanytarsus lauterborni [3]*	R/U/C/S	R/U/C/S	157	239		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACM0242	Chironomidae	*Paratanytarsus sp.*	R/U/C/S	R/U/C/S	47	4		b5 b3		b5 b3		b5 b3
BOLD:AAO1037	**Chironomidae**	***Paratendipes albimanus [127]***	**R/U/C/S**	**R/U/C/S**	**161**	**159**		**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:AAU2481	Chironomidae	*Phaenopsectra flavipes [13]*	R/U/-/S	R/U/C/S	10	20	b5 b3	b3		b5 b3	b3	b5 b3
BOLD:AAL0178	Chironomidae	*Polypedilum albicorne [78]*	R/U/C/S	R/U/C/S	145	104		b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAM9576	Chironomidae	*Polypedilum albinodus [3]*	R/U/C/S	R/U/C/S	227	25		b5 b3	b5 b3	b5 b3		b5 b3

BIN	Family	Main species id of BIN	bc-5'	bc-3'	r-bc-5'	r-bc-3'	C	I	MC	MS	T1	T2
BOLD:AAW4728	Chironomidae	*Polypedilum pullum [2]*	R/-/-/-	-/-/-/-	4	0		b5		b5		b5
BOLD:AAD7458	Chironomidae	*Prodiamesa olivacea [46]*	R/U/C/S	R/U/C/S	79	16	b5 b3	b5 b3	b5	b5 b3		b5 b3
BOLD:ACQ1908	Chironomidae	*Rheocricotopus chalybeatus [6]*	R/U/C/S	-/-/-/-	58	0		b5		b5		b5
BOLD:AAV2322	Chironomidae	*Rheocricotopus fuscipes [20]*	R/U/C/S	R/U/C/S	11	31	b5 b3			b5 b3		b5 b3
BOLD:AAD0309	Chironomidae	*Stempellina bausei [10]*	R/U/C/S	R/U/C/S	141	438	b5 b3		b5	b5 b3	b5 b3	b5 b3
BOLD:AAU2625	Chironomidae	*Stempellinella edwardsi [10]*	R/U/-/S	R/U/C/S	7	9		b5 b3		b5 b3		b5 b3
BOLD:ACM5335	Chironomidae	*Synorthocladius semivirens [7]*	R/U/C/S	R/U/C/S	950	3924	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ACQ8988	**Chironomidae**	***Tanytarsus brundini [13]***	**R/U/-/S**	**R/U/C/S**	**2716**	**3155**		**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:AAB9119	**Chironomidae**	***Tanytarsus brundini [5]***	**R/-/C/-**	**R/-/-/-**	**1496**	**16**		**b5 b3**	**b5 b3**	**b5**	**b5**	**b5 b3**
BOLD:AAW1102	**Chironomidae**	***Tanytarsus ejuncidus [24]***	**R/U/C/-**	**R/U/C/S**	**2398**	**3230**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:AAU4439	**Chironomidae**	***Tanytarsus eminulus [124]***	**R/U/C/S**	**R/U/C/S**	**4847**	**2023**	**b5**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:ACF7553	Chironomidae	*Tanytarsus heusdensis [5]*	R/-/-/-	R/-/-/-	1	5		b5 b3		b5 b3	b5 b3	b3
BOLD:AAV3526	Chironomidae	*Tanytarsus heusdensis [6]*	R/U/C/S	R/U/C/S	2	34	b5 b3		b3	b5 b3		b5 b3
BOLD:ACR3318	**Chironomidae**	***Tanytarsus pallidicornis [10]***	**R/U/C/S**	**R/U/C/S**	**34**	**469**	**b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b3**	**b5 b3**
BOLD:ACD2995	Empididae	*Chelifera precatoria [6]*	R/U/C/S	R/U/C/S	156	385	b5 b3	b5 b3		b5 b3	b5 b3	b5 b3
BOLD:ACZ6583	Ephydridae	*Scatella tenuicosta [8]*	R/U/C/S	R/U/C/S	27	13		b5 b3		b5 b3	b5 b3	
BOLD:ACP1316	n.a	Diptera	R/U/C/S	R/U/C/S	106	95	b5 b3	b5 b3	b5	b5 b3		b5 b3
BOLD:ACY5064	n.a	Diptera	R/U/C/S	R/U/C/S	97	260	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ABA7297	Pediciidae	*Dicranota bimaculata [6]*	R/-/-/-	R/U/C/S	12	140	b5 b3		b5 b3			b5 b3
BOLD:AAL7819	Psychodidae	*Psychoda sp. [8]*	R/U/C/-	R/-/-/-	60	3		b5 b3		b5 b3	b5 b3	
BOLD:AAN3314	Simuliidae	*Simulium ornatum [41]*	R/U/C/S	R/U/C/S	524	178	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAA8323	Simuliidae	*Simulium silvestre [151]*	R/-/-/-	-/-/-/-	1	0	b5		b5			b5
BOLD:AAP9556	Simuliidae	*Simulium velutinum [11]*	R/U/C/S	R/U/C/S	191	191	b3	b5 b3	b5 b3		b3	b5 b3
BOLD:AAB8624	Simuliidae	*Simulium vernum [27]*	R/U/C/S	-/-/-/-	162	0	b5		b5			b5
BOLD:AAN6407	Sphaeroceridae	*Coproica ferruginata [126]*	R/U/C/S	-/-/-/-	11	0	b5			b5		b5
BOLD:AAJ5023	Tabanidae	*Chrysops caecutiens [8]*	R/U/C/S	-/-/-/-	18	0	b5		b5			b5
BOLD:ABV4656	Tipulidae	*Tipula benesignata [2]*	R/U/C/S	R/U/C/S	21	26	b5 b3			b5 b3		b5 b3
BOLD:AAE7386	Tipulidae	*Tipula paludosa [355]*	R/U/C/S	R/U/C/S	238	1516	b5 b3			b5 b3		b5 b3
BOLD:AAF6378	Tipulidae	*Tvetenia calvescens [14]*	-/-/-/-	R/-/-/-	0	4		b3		b3	b3	
BOLD:AAG1011	Tipulidae	*Tvetenia calvescens [186]*	R/U/C/S	R/U/C/S	14729	1930	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3

Notes: bc-5': *cox1* barcode 5' fragment; bc-3': *cox1* barcode 3' fragment; R/U/C/S: Indicates detection based on Reads, USEARCH, CROP and SWARM respectively. r-bc-5' and r-bc-3': Number of reads matched for bc-5' and bc-3' respectively. C: Control (upstream) sites; I: Impacted (downstream) sites; MC: Macrofauna subsamples; MS; Meiofauna subsamples; T1: Samples collected in time 1 (11 days after the spill); T2: Samples collected in time 2 (2.5 months after the spill). b5 and b3 indicates the detection of the OTU with the bc-5' and bc-3' fragments respectively based on the processed reads. In bold species identified with indval analyses as indicative for impacted sites. Named species are the most abundant within each BIN, in brackets the number of specimens identified to species level in the reference database.

Table 2. Betadiversity values, NMDS stress values and p-values for the comparison between control and impacted sites.

	Dataset		Betadiversity (Sorensen index)					Turnover (Simpson index)			Nestedness (Sorensen-Simpson index)		
Taxa	DNA fragment	Method	beta.sor	Adonis p-value	Adonis r^2	stress	*envfit p-value*	beta.sim	stress	*envfit p-value*	beta.sne	stress	*envfit p-value*
Metazoa	*cox1-3'*	*de novo* 3%	0.85	0.007	0.18	0.086	0.023	0.763	0.144	0.179	0.087	0.067	0.099
		de novo 10%	0.835	0.011	0.19	0.075	0.018	0.736	0.143	0.054	0.099	0.053	0.099
	cox1-5'	*de novo* 3%	0.833	0.004	0.17	0.122	0.016	0.743	0.149	0.089	0.09	0.042	0.158
		de novo 10%	0.821	0.005	0.18	0.12	0.036	0.731	0.118	0.144	0.09	0.035	0.107
	SSU	*de novo* 3%	0.775	0.007	0.21	0	0.107	0.726	0.108	0.036	0.049	0.032	0.35
Insecta	*cox1-3'*	*de novo* 3%	0.862	0.004	0.18	0.087	0.025	0.776	0.143	0.035	0.085	0.07	0.382
		de novo 10%	0.849	0.005	0.19	0.088	0.009	0.761	0.144	0.028	0.088	0.066	0.37
		BOLD ref	0.857	0.014	0.17	0.109	0.108	0.758	0.146	0.033	0.099	0.096	0.536
	cox1-5'	*de novo* 3%	0.841	0.03	0.16	0.14	0.153	0.749	0.154	0.328	0.092	0.063	0.369
		de novo 10%	0.818	0.019	0.17	0.108	0.066	0.739	0.115	0.226	0.079	0.058	0.314
		BOLD ref	0.846	0.023	0.16	0.154	0.089	0.759	0.162	0.041	0.087	0.058	0.196
	SSU	*de novo* 3%	0.805	0.037	0.16	0.173	0.132	0.739	0.161	0.057	0.066	0.059	0.62
Crustacea	*cox1-3'*	*de novo* 3%	0.801	0.002	0.23	0.107	0.013	0.684	0.122	0.001	0.118	0.069	0.656
		de novo 10%	0.773	0.004	0.28	0.078	0.009	0.67	0.133	0.006	0.103	0.081	0.779
		BOLD ref	0.826	0.004	0.29	0.089	0.001	0.753	0.077	0.001	0.073	0.181	0.669
	cox1-5'	*de novo* 3%	0.762	0.002	0.38	0.085	0.006	0.658	0.095	0.003	0.104	0.047	0.102
		de novo 10%	0.763	0.006	0.43	0.098	0.004	0.663	0.073	0.005	0.1	0.079	0.196
		BOLD ref	0.797	0.006	0.34	0.091	0.004	0.725	0.073	0.009	0.072	0.132	0.475
	SSU	*de novo* 3%	0.746	0.005	0.32	0.057	0.004	0.675	0.103	0.011	0.07	0.076	0.223
Arthropoda	*cox1-3'*	*de novo* 3%	0.845	0.005	0.19	0.113	0.011	0.77	0.136	0.179	0.075	0.061	0.26
		de novo 10%	0.832	0.007	0.20	0.092	0.017	0.758	0.135	0.015	0.074	0.043	0.317
		BOLD ref	0.856	0.005	0.19	0.104	0.113	0.779	0.128	0.006	0.077	0.095	0.569
	cox1-5'	*de novo* 3%	0.829	0.005	0.18	0.142	0.002	0.746	0.143	0.003	0.083	0.064	0.457
		de novo 10%	0.82	0.002	0.18	0.139	0.017	0.75	0.142	0.013	0.07	0	0.236
		BOLD ref	0.843	0.004	0.18	0.171	0.006	0.771	0.164	0.02	0.072	0.059	0.242
	SSU	*de novo* 3%	0.787	0.001	0.22	0.095	0.006	0.745	0.118	0.004	0.042	0.059	0.798

Notes: *de novo* 3% and *de novo* 10% refers to OTU clustering at these threshold values. BOLD ref indicates the results from the 'taxon dependent' approach of mapping reads against OTU clusters generated from BOLD data.

Table 3. Species with indicative value as identified by *indval* analyses based on the results of the reference database-dependent approach for the *cox1* gene fragments bc-5' and bc-3'.

BIN	Class	Order	Family	Species	GENE	T	Ind. value	P
BOLD:AAF2659	Branchiopoda	Diplostraca	Chydoridae	*Chydorus sphaericus*	bc-5'	I	0.6	0.046
BOLD:AAF2659	Branchiopoda	Diplostraca	Chydoridae	*Chydorus sphaericus*	bc-5'	T2-I	1	0.014
BOLD:AAP5931	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [1]*	bc-5'	T2-I	1	0.024
BOLD:ACU8677	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [1]*	bc-5'	T2-I	1	0.022
BOLD:AAI6018	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [123]*	bc-3'	I	0.6	0.048
BOLD:AAI6018	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [123]*	bc-3'	T2-I	1	0.023
BOLD:AAE4568	Insecta	Diptera	Chironomidae	*Eukiefferiella claripennis [185]*	bc-5'	T2-I	1	0.019
BOLD:AAO1037	Insecta	Diptera	Chironomidae	*Paratendipes albimanus [127]*	bc-3'	I	0.8	0.024
BOLD:AAO1037	Insecta	Diptera	Chironomidae	*Paratendipes albimanus [127]*	bc-5'	I	0.8	0.014
BOLD:ACQ8988	Insecta	Diptera	Chironomidae	*Tanytarsus brundini [13]*	bc-3'	I	0.8	0.02
BOLD:ACQ8988	Insecta	Diptera	Chironomidae	*Tanytarsus brundini [13]*	bc-5'	I	0.8	0.02
BOLD:AAB9119	Insecta	Diptera	Chironomidae	*Tanytarsus brundini [5]*	bc-5'	I	0.8	0.017
BOLD:AAW1102	Insecta	Diptera	Chironomidae	*Tanytarsus ejuncidus [24]*	bc-3'	I	0.75	0.049
BOLD:AAU4439	Insecta	Diptera	Chironomidae	*Tanytarsus eminulus [124]*	bc-3'	I	1	0.002
BOLD:AAU4439	Insecta	Diptera	Chironomidae	*Tanytarsus eminulus [124]*	bc-5'	I	0.86	0.015
BOLD:ACR3318	Insecta	Diptera	Chironomidae	*Tanytarsus pallidicornis [10]*	bc-3'	I	0.86	0.015
BOLD:AAU1007	Insecta	Ephemeroptera	Baetidae	*Centroptilum luteolum [17]*	bc-5'	T2	1	0.001
BOLD:ACH6832	Malacostraca	Amphipoda	Gammaridae	*Gammarus pulex [11]*	bc-3'	C	0.83	0.016
BOLD:ACH7960	Malacostraca	Amphipoda	Gammaridae	*Gammarus nekkensis [1]*	bc-5'	C	1	0.004
BOLD:ACG8343	Malacostraca	Amphipoda	Gammaridae	*Gammarus fossarum [37]*	bc-5'	C	0.83	0.025
BOLD:AAA1971	Malacostraca	Isopoda	Asellidae	*Asellus aquaticus [85]*	bc-3'	I	1	0.004
BOLD:ACV6778	Malacostraca	Isopoda	Asellidae	*Asellus aquaticus [7]*	bc-3'	I	0.8	0.01
BOLD:AAA1971	Malacostraca	Isopoda	Asellidae	*Asellus aquaticus [85]*	bc-5'	I	0.86	0.019

Notes: T: Treatment; C: Control sites, I: impacted sites; T2: collection period 2 (2.5 months after the spill); T2-I: Impacted sites at collection period 2. Named species are the most abundant within each BIN, in brackets the number of specimens identified to species level in the reference database.

FIGURES

Figure 1. The flotation method for extracting meio- and macro-fauna from the original Surber samples. The figure illustrates how after flotation (A) both fractions were separated by passage (B) through a 1 mm metal mesh sieve that retains the macrofauna (C), whereas a 0.45 micron sieve retains the meiofauna (D). At both sieving steps ample water was used to flush smaller items, including bacteria and other microorganisms that otherwise might also produce PCR products.

Figure 2. Shared OTUs from using alternative parameter settings of *de novo* OTU generation, for each of the four steps in Fig. S2. The diagrams show the proportion of shared OTUs in a list for any pair of parameter settings. Note that in most analyses the intersection of OTU lists indicates a large percentage common to all settings, except for the clustering with CROP in the *SSU* dataset and the BFC=20 in the denoising step.

Figure 3. Number of OTUs at 3% similarity thresholds at various hierarchical levels. Black: Usearch; dark grey: CROP; light grey: Swarm. The clustering of OTUs with each program was started from paired reads after quality filtering using the following parameters: BFC with *s*=0.5 for read denoising, *Maxee=1* for read filtering, and *mergepairs* in Usearch for read merging.

Figure 4. Total number of OTUs with the *de novo* generation and read mapping approaches for the two portions of *cox1*. The OTU count is based on BLAST+Megan for *de novo* generated OTUs and on the matches to the *BOLD-OTU* reference database for the read mapping approach. Black: Usearch at 3% sequence similarity threshold; dark grey: Usearch at 10% similarity threshold; light grey: read mapping to BOLD reference dataset.

Figure 5. NMDS ordinations for Metazoa, Arthropoda, Insecta and Crustacea based on presence/absence community matrices as obtained by read mapping against *de novo* generated OTUs at a 10% similarity threshold for the *bc-3′* gene fragment.

Figure 6. NMDS total betadiversity ordinations for Arthropoda, Insecta and Crustacea based on presence/absence community matrices as obtained by read mapping against *de novo* generated OTUs at 3% and 10% for the *bc-3′* and *bc-5′* gene fragments (*bc-5′* 3%; *bc-3′* 3%; *bc-5′* 10%; *bc-3′* 10%), at 3% for SSU (*SSU* 3%), and by read mapping against *BOLD-OTUS* (*bc-3′* BOLD and bc-5' BOLD)

Figure 7. NMDS ordinations for Arthropoda and the *cox1-5'*and *cox1-3′* datasets using the reference database approach with Macro (labelled with "M") and meiofauna (labelled with "m") subsamples considered independently.

References

Arribas P, Andújar C, Hopkins K *et al.* (2016) Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution*, **7**, 1071–1081.

Babraham Institute (2013) FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga A, Orme CDL (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Bisconti R, Canestrelli D, Tenchini R *et al.* (2016) Cryptic diversity and multiple origins of the widespread mayfly species group Baetis rhodani (Ephemeroptera: Baetidae) on northwestern Mediterranean islands. *Ecology and Evolution*, **6**, 7901–7910.

Bista I, Carvalho GR, Walsh K *et al.* (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications*, **8**, 14087.

Blaxter ML, De Ley P, Garey JR *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.

Blaxter M, Mann J, Chapman T *et al.* (2005) Defining operational taxonomic units using DNA barcode data. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1935–1943.

Bohan DA, Vacher C, Tamaddoni-Nezhad A *et al.* (2017) Next-Generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in Ecology and Evolution*, **32**, 477–487.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Brandon-Mong G-J, Gan H-M, Sing K-W *et al.* (2015) DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research*, **105**, 717–727.

Burgess R (2001) An improved protocol for separating meiofauna from sediments using colloidal silica sols. *Marine Ecology Progress Series*, **214**, 161–165.

Camargo J (1993) Macrobenthic surveys as a valuable tool for assessing freshwater quality in the Iberian Peninsula. *Environ. Monit. Assess*, **24**, 71–90.

Chessman BC, Trayler KM, Davis JA (2002) Family- and species-level biotic indices for macroinvertebrates of wetlands on the Swan Coastal Plain, Western Australia. *Marine and Freshwater Research*, **53**, 919–930.

Creer S, Deiner K, Frey S *et al.* (2016) The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, **7**, 1008–1018.

Creer S, Fonseca VG, Porazinska DL *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

Deagle BE, Jarman SN, Coissac E *et al.* (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 1789–1793.

Doublet V, Raimond R, Grandjean F *et al.* (2012) Widespread atypical mitochondrial DNA structure in isopods (Crustacea, Peracarida) related to a constitutive heteroplasmy in terrestrial species. *Genome*, **55**, 234–244.

Dufrêne M, Legendre P (1997) Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs*, **67**, 345–366.

Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**, 996–8.

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. *Plos One*, **10**, e0130324.

Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, **43**, 2513–2524.

European Commission (2000) Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy. *Official Journal (OJ L 327)*.

Fonseca VG, Carvalho GR, Sung W *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Friberg N, Bonada N, Bradley DC *et al.* (2011) Biomonitoring of Human Impacts in Freshwater Ecosystems. The Good, the Bad and the Ugly. *Advances in Ecological Research*, **44**, 1–68.

Gray C, Hildrew AG, Lu X *et al.* (2016) Recovery and nonrecovery of freshwater food webs from the effects of acidification. *Advances in Ecological Research*, **55**, 475–534.

Guardiola M, Uriz MJ, Taberlet P *et al.* (2015) Deep-sea, deep-sequencing: Metabarcoding extracellular DNA from sediments of marine canyons. *PLoS ONE*, **10**.

Gutiérrez-Cánovas C, Velasco J, Millán A (2008) SALINDEX: A macroinvertebrate index for assessing the ecological status of saline "ramblas" from SE of the Iberian Peninsula. *Limnetica*, **27**, 299316.

Hajibabaei M, Baird DJ, Fahner NA, Beiko R, Golding GB (2016) A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **371**, 20150330.

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics (Oxford, England)*, **27**, 611–8.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, **270**, 313–21.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Ingvason HR, Ólafsson JS, Gardarsson A (2004) Food selection of Tanytarsus gracilentus larvae (Diptera: Chironomidae): An analysis of instars and cohorts. *Aquatic Ecology*, **38**, 231–237.

Ji Y, Ashton L, Pedley SSM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–57.

Jones FC (2008) Taxonomic sufficiency: The influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Reviews*, **16**, 45–69.

Karaman GS, Pinkster S (1977) Freshwater Gammarus species from Europe, North Africa and adjacent regions of Asia (Crustacea-Amphipoda). Part I. Gammarus pilex-group and related species. *Bijdragen tot de Dierkunde*, **47**, 1–97.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Lanzén A, Lekang K, Jonassen I, Thompson EM, Troedsson C (2016) High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology*, **25**, 4392–4406.

Leese F, Altermatt F, Bouchez A *et al.* (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes*, **2**, e11321.

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*, **2014**, 201424997.

Li H (2015) BFC: correcting Illumina sequencing errors. *Bioinformatics*, 1–3.

Lobo J, Shokralla S, Costa MH, Hajibabaei M, Costa FO (2015) Stepwise implementation of high-throughput sequencing metabarcoding to estuarine macrobenthic communities. *Genome*, **58**, 248.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**, e593.

Marshall JC, Steward AL, Harch BD (2006) Taxonomic resolution and quantification of freshwater macroinvertebrate samples from an Australian dryland river: The benefits and costs of using species abundance data. *Hydrobiologia*, **572**, 171–194.

Meyer CP, Paulay G (2005) DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.

Oksanen J, Blanchet FG, Kindt R *et al.* (2013) Vegan: Community Ecology Package. R package version 2.0-10. http://cran.r-project.org/package=vegan. *R package ver. 2.0–8*, 254.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pauls SU, Alp M, Bálint M *et al.* (2014) Integrating molecular tools into freshwater ecology: Developments and opportunities. *Freshwater Biology*, **59**, 1559–1576.

Pons J, Barraclough T, Gomez-Zurita J *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.

Ratnasingham S, Hebert PDN (2007) BARCODING, BOLD : The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE*, **8**.

Roberts D (2007) Labdsv: Ordination and multivariate analysis for ecology. R package version 1, 3-1.

Rutschmann S, Gattolliat JL, Hughes SJ *et al.* (2014) Evolution and island endemism of morphologically cryptic Baetis and Cloeon species (Ephemeroptera, Baetidae) on the Canary Islands and Madeira. *Freshwater Biology*, **59**, 2516–2527.

Sánchez-Montoya MDM, Puntí T, Suárez ML *et al.* (2007) Concordance between ecotypes and macroinvertebrate assemblages in Mediterranean streams. *Freshwater Biology*, **52**, 2240–2255.

Schirmer M, Ijaz UZ, D'Amore R *et al.* (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, **43**, e37.

Schloss PD, Westcott SL (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology*, **77**, 3219–3226.

Schmidt-Kloiber A, Hering D (2015) Www.freshwaterecology.info - An online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. *Ecological Indicators*, **53**, 271–282.

Schmidt-Kloiber A, Nijboer RC (2004) The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiologia*, **516**, 269–283.

Shokralla S, Gibson JF, Nikbakht H *et al.* (2014) Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, **14**, 892–901.

Shokralla S, Porter TM, Gibson JF *et al.* (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, **5**, 9687.

Staton S (2013) Pairfq: sync paired-end FASTA/Q files and keep singleton reads.

Stubauer I, Moog O (2000) Taxonomic sufficiency versus need for information- comments based on Austrian experience in biological water quality monitoring. In: *Proceedings of the 27th Congress of the International Association of Theoretical and Applied Limnology. Vol. 27*, p. 5.

Sworobowicz L, Grabowski M, Omasz Mamos T *et al.* (2015) Revisiting the phylogeography of Asellus aquaticus in Europe: insights into cryptic diversity and spatiotemporal diversification. *Freshwater Biology*, **60**, 1824–1840.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–50.

Tachet H, Richoux P, Bournaud M, Usseglio-Polatera P (2002) *Invertebres D'eau Douce. Systematique, Biologie, Ecologie* (C Editions, Ed,). Paris.

Tang CQ, Leasi F, Obertegger U *et al.* (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences*, **109**, 16208–16212.

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, **12**, 377–88.

Thompson MSA, Bankier C, Bell T *et al.* (2016) Gene-to-ecosystem impacts of a catastrophic pesticide spill: testing a multilevel bioassessment approach in a river ecosystem. *Freshwater Biology*, **61**, 2037–2050.

United States (1972) Federal Water Pollution Control Act Amendments of 1972. Pub.L. 92-500, October 18.

Wang Y, Naumann U, Wright ST, Warton DI (2012) Mvabund- an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474.

Wheeler QD (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **359**, 571–583.

Wickham H (2013) stringr: simple, consistent wrappers for common string operations. R package.

Williams HC, Ormerod SJ, Bruford MW (2006) Molecular systematics and phylogeography of the cryptic species complex Baetis rhodani (Ephemeroptera, Baetidae). *Molecular Phylogenetics and Evolution*, **40**, 370–382.

Woodward G, Gray C, Baird DJ (2013) Biomonitoring for the 21st Century: new perspectives in an age of globalisation and emerging environmental threats. *Limnetica*, **32**, 159–174.

Wright JF, Sutcliffe DW, Furse MT (2000) *Assessing the biological quality of fresh waters*. Freshwater Biological Association, Ambleside, Cumbria, UK.

Yang C, Wang X, Miller JA *et al.* (2014) Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, **46**, 379–389.

Yu D, Ji Y, Emerson B *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics (Oxford, England)*, **29**, 2869–76.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, **30**, 614–20.

Zinger L, Chave J, Coissac E *et al.* (2016) Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology and Biochemistry*, **96**, 16–19.

Table 1. BINs of Diptera from BOLD identified based on *usearch_global* searches under a 3% similarity threshold of (i) processed reads and (ii) OTUs clustered at 3% (details in text).

BIN	Family	Main species id of BIN	bc-5'	bc-3'	r-bc-5'	r-bc-3'	C	I	MC	MS	T1	T2
BOLD:AAJ7051	Agromyzidae	*Agromyza pseudoreptans [19]*	-/-/-/-	R/U/C/S	0	5		b3	b3		b3	
BOLD:ACI4790	Bibionidae	*Dilophus febrilis [21]*	R/U/C/S	R/U/C/S	478	428	b5 b3		b5 b3	b5	b5	b5 b3
BOLD:ACP0608	Cecidomyiidae	Cecidomyiidae *[3]*	R/U/C/S	R/-/-/S	455	31		b5 b3		b5 b3		b5 b3
BOLD:ACS1169	Ceratopogonidae	*Palpomyia flavipes [6]*	R/U/-/-	R/U/C/S	20	41	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ACD1957	Chironomidae	*Apsectrotanypus trifascipennis [9]*	R/U/C/S	R/U/C/S	81	1446	b5 b3	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ADE2432	Chironomidae	*Brillia bifida [6]*	R/-/-/-	-/-/-/-	1	0		b5		b5	b5	
BOLD:ACR1089	Chironomidae	Chironomidae	R/U/C/S	R/U/C/S	550	93	b5 b3	b5 b3	b5	b5 b3	b5 b3	b5 b3
BOLD:ACP8764	Chironomidae	Chironomidae *[13]*	R/U/C/S	-/-/-/-	130	0		b5	b5	b5		b5
BOLD:ACP6740	Chironomidae	Chironomidae *[60]*	R/-/-/-	R/U/C/S	11	9		b5 b3	b3	b5 b3		b5 b3
BOLD:ACP2182	Chironomidae	Chironomidae *[92]*	R/U/C/S	R/U/C/S	33	26		b5 b3		b5 b3		b5 b3
BOLD:AAW5799	Chironomidae	*Conchapelopia hittmairorum [3]*	R/-/-/-	R/-/-/-	1	4		b5 b3		b5 b3		b5 b3
BOLD:AAP5886	Chironomidae	*Conchapelopia melanops [11]*	R/U/C/S	R/U/C/S	1018	331	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ACQ3496	Chironomidae	*Conchapelopia pallidula [1]*	R/U/C/S	-/-/-/-	3	0		b5		b5		b5
BOLD:ACD1670	Chironomidae	*Corynoneura sp.*	R/U/C/S	R/U/C/S	1178	1263	b5 b3	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACT8698	Chironomidae	*Corynoneura sp. [6]*	R/U/C/S	R/U/C/S	21	72	b5 b3	b5 b3		b5 b3		b5 b3
BOLD:AAW5785	Chironomidae	*Cricotopus albiforceps [139]*	R/-/-/-	R/-/-/-	2	4	b5 b3			b5 b3		b5 b3
BOLD:AAF2345	Chironomidae	*Cricotopus annulator [19]*	R/U/C/S	R/U/C/S	30	14		b5 b3	b5 b3	b5		b5 b3
BOLD:AAP5931	Chironomidae	*Cricotopus bicinctus [1]*	R/-/-/-	-/-/-/-	60	0		b5	b5	b5		b5
BOLD:ACU8677	Chironomidae	*Cricotopus bicinctus [1]*	R/-/-/-	-/-/-/-	5	0		b5	b5	b5		b5
BOLD:AAI6018	Chironomidae	*Cricotopus bicinctus [123]*	R/U/C/-	R/-/C/S	5751	4759	b5	b5 b3	b5 b3	*b5 b3	b5	b5 b3
BOLD:AAT9677	Chironomidae	*Cricotopus bicinctus [27]*	R/U/-/S	R/U/C/-	123	1991	b3	b5 b3	b5 b3	b5 b3		b5 b3

BIN	Family	Main species id of BIN	bc-5'	bc-3'	r-bc-5'	r-bc-3'	C	I	MC	MS	T1	T2
BOLD:AAM5377	Chironomidae	*Cricotopus rufiventris [289]*	R/U/C/S	R/U/C/S	8	44		b5 b3	b5 b3	b3		b5 b3
BOLD:AAA5299	Chironomidae	*Cricotopus sylvestris [64]*	R/U/C/S	R/U/C/S	4	88		b5 b3	b5 b3			b5 b3
BOLD:AAU2576	Chironomidae	*Cricotopus trifascia [5]*	R/U/C/S	R/U/C/S	119	126		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAE4568	Chironomidae	*Eukiefferiella claripennis [185]*	R/U/C/-	R/U/C/-	164	116		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACT0982	Chironomidae	*Heterotrissocladius sp. 2SW [2]*	R/U/C/S	R/U/C/S	48	234	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAX3566	Chironomidae	*Macropelopia nebulosa [13]*	R/U/C/S	R/U/C/S	29	58	b5 b3	b5	b5 b3	b5 b3	b5 b3	b3
BOLD:AAB8862	Chironomidae	*Metriocnemus eurynotus [18]*	R/-/-/-	R/U/C/S	3	18		b5 b3		b5 b3	b5 b3	
BOLD:AAD4167	Chironomidae	*Micropsectra atrofasciata [26]*	R/U/C/S	R/U/C/S	826	596		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAC7823	Chironomidae	*Micropsectra contracta [14]*	R/U/C/S	R/U/C/S	62	98		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAD1527	Chironomidae	*Micropsectra lindrothi [18]*	R/U/C/S	R/-/-/-	5	4		b5 b3	b5 b3			b5 b3
BOLD:AAC7552	Chironomidae	*Micropsectra pallidula [24]*	R/U/C/S	R/U/C/S	91	53		b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAI1530	Chironomidae	*Micropsectra sp. 5ES [33]*	R/U/C/S	R/U/C/S	61	1641		b5 b3	b5 b3	b5 b3	b3	b5 b3
BOLD:ACR0263	Chironomidae	*Microtendipes pedellus [8]*	R/U/C/S	R/U/C/S	12	19		b5 b3		b5 b3		b5 b3
BOLD:AAW0928	Chironomidae	*Nanocladius rectinervis [6]*	R/U/C/S	R/U/C/S	593	166	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAD8971	Chironomidae	*Orthocladius oblidens [330]*	R/U/C/S	R/U/C/S	11359	5138	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAM5389	Chironomidae	*Orthocladius rubicundus [119]*	R/U/C/S	R/U/C/S	98	64	b5	b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAW5449	Chironomidae	*Orthocladius rubicundus [25]*	R/U/C/S	R/U/C/S	39	145		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACX3335	Chironomidae	*Paracladius quadrinodosus [2]*	-/-/-/-	R/-/-/-	0	5		b3		b3		b3
BOLD:ACT5340	Chironomidae	*Paracladopelma camptolabis [2]*	R/U/C/S	R/U/C/S	13	15	b5 b3	b3	b5 b3	b3		b5 b3
BOLD:AAW4635	Chironomidae	*Paratanytarsus dissimilis [12]*	R/U/C/S	R/U/C/S	35	90		b5 b3	b5 b3			b5 b3
BOLD:AAL3267	Chironomidae	*Paratanytarsus lauterborni [3]*	R/U/C/S	R/U/C/S	157	239		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:ACM0242	Chironomidae	*Paratanytarsus sp.*	R/U/C/S	R/U/C/S	47	4		b5 b3		b5 b3		b5 b3
BOLD:AAO1037	**Chironomidae**	***Paratendipes albimanus [127]***	**R/U/C/S**	**R/U/C/S**	**161**	**159**		**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:AAU2481	Chironomidae	*Phaenopsectra flavipes [13]*	R/U/-/S	R/U/C/S	10	20	b5 b3	b3		b5 b3	b3	b5 b3
BOLD:AAL0178	Chironomidae	*Polypedilum albicorne [78]*	R/U/C/S	R/U/C/S	145	104		b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAM9576	Chironomidae	*Polypedilum albinodus [3]*	R/U/C/S	R/U/C/S	227	25		b5 b3	b5 b3	b5 b3		b5 b3
BOLD:AAW4728	Chironomidae	*Polypedilum pullum [2]*	R/-/-/-	-/-/-/-	4	0		b5		b5		b5
BOLD:AAD7458	Chironomidae	*Prodiamesa olivacea [46]*	R/U/C/S	R/U/C/S	79	16	b5 b3	b5 b3	b5	b5 b3		b5 b3
BOLD:ACQ1908	Chironomidae	*Rheocricotopus chalybeatus [6]*	R/U/C/S	-/-/-/-	58	0		b5		b5		b5
BOLD:AAV2322	Chironomidae	*Rheocricotopus fuscipes [20]*	R/U/C/S	R/U/C/S	11	31	b5 b3			b5 b3		b5 b3
BOLD:AAD0309	Chironomidae	*Stempellina bausei [10]*	R/U/C/S	R/U/C/S	141	438	b5 b3		b5	b5 b3	b5 b3	b5 b3
BOLD:AAU2625	Chironomidae	*Stempellinella edwardsi [10]*	R/U/-/S	R/U/C/S	7	9		b5 b3		b5 b3		b5 b3
BOLD:ACM5335	Chironomidae	*Synorthocladius semivirens [7]*	R/U/C/S	R/U/C/S	950	3924	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ACQ8988	**Chironomidae**	***Tanytarsus brundini [13]***	**R/U/-/S**	**R/U/C/S**	**2716**	**3155**		**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:AAB9119	**Chironomidae**	***Tanytarsus brundini [5]***	**R/-/C/-**	**R/-/-/-**	**1496**	**16**		**b5 b3**	**b5 b3**	**b5**	**b5**	**b5 b3**
BOLD:AAW1102	**Chironomidae**	***Tanytarsus ejuncidus [24]***	**R/U/C/-**	**R/U/C/S**	**2398**	**3230**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:AAU4439	**Chironomidae**	***Tanytarsus eminulus [124]***	**R/U/C/S**	**R/U/C/S**	**4847**	**2023**	**b5**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b5 b3**
BOLD:ACF7553	Chironomidae	*Tanytarsus heusdensis [5]*	R/-/-/-	R/-/-/-	1	5		b5 b3		b5 b3	b5 b3	b3
BOLD:AAV3526	Chironomidae	*Tanytarsus heusdensis [6]*	R/U/C/S	R/U/C/S	2	34	b5 b3		b3	b5 b3		b5 b3
BOLD:ACR3318	**Chironomidae**	***Tanytarsus pallidicornis [10]***	**R/U/C/S**	**R/U/C/S**	**34**	**469**	**b3**	**b5 b3**	**b5 b3**	**b5 b3**	**b3**	**b5 b3**
BOLD:ACD2995	Empididae	*Chelifera precatoria [6]*	R/U/C/S	R/U/C/S	156	385	b5 b3	b5 b3		b5 b3	b5 b3	b5 b3
BOLD:ACZ6583	Ephydridae	*Scatella tenuicosta [8]*	R/U/C/S	R/U/C/S	27	13		b5 b3		b5 b3	b5 b3	
BOLD:ACP1316	n.a	Diptera	R/U/C/S	R/U/C/S	106	95	b5 b3	b5 b3	b5	b5 b3		b5 b3
BOLD:ACY5064	n.a	Diptera	R/U/C/S	R/U/C/S	97	260	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:ABA7297	Pediciidae	*Dicranota bimaculata [6]*	R/-/-/-	R/U/C/S	12	140	b5 b3		b5 b3			b5 b3
BOLD:AAL7819	Psychodidae	*Psychoda sp. [8]*	R/U/C/-	R/-/-/-	60	3		b5 b3		b5 b3	b5 b3	
BOLD:AAN3314	Simuliidae	*Simulium ornatum [41]*	R/U/C/S	R/U/C/S	524	178	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3
BOLD:AAA8323	Simuliidae	*Simulium silvestre [151]*	R/-/-/-	-/-/-/-	1	0	b5		b5			b5
BOLD:AAP9556	Simuliidae	*Simulium velutinum [11]*	R/U/C/S	R/U/C/S	191	191	b3	b5 b3	b5 b3		b3	b5 b3
BOLD:AAB8624	Simuliidae	*Simulium vernum [27]*	R/U/C/S	-/-/-/-	162	0	b5		b5			b5
BOLD:AAN6407	Sphaeroceridae	*Coproica ferruginata [126]*	R/U/C/S	-/-/-/-	11	0	b5			b5		b5
BOLD:AAJ5023	Tabanidae	*Chrysops caecutiens [8]*	R/U/C/S	-/-/-/-	18	0	b5		b5			b5
BOLD:ABV4656	Tipulidae	*Tipula benesignata [2]*	R/U/C/S	R/U/C/S	21	26	b5 b3			b5 b3		b5 b3
BOLD:AAE7386	Tipulidae	*Tipula paludosa [355]*	R/U/C/S	R/U/C/S	238	1516	b5 b3			b5 b3		b5 b3
BOLD:AAF6378	Tipulidae	*Tvetenia calvescens [14]*	-/-/-/-	R/-/-/-	0	4		b3		b3	b3	
BOLD:AAG1011	Tipulidae	*Tvetenia calvescens [186]*	R/U/C/S	R/U/C/S	14729	1930	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3	b5 b3

Notes: bc-5': *cox1* barcode 5' fragment; bc-3': *cox1* barcode 3' fragment; R/U/C/S: Indicates detection based on Reads, USEARCH, CROP and SWARM respectively. r-bc-5' and r-bc-3': Number of reads matched for bc-5' and bc-3' respectively. C: Control (upstream) sites; I: Impacted (downstream) sites; MC: Macrofauna subsamples; MS; Meiofauna subsamples; T1: Samples collected in time 1 (11 days after the spill); T2: Samples collected in time 2 (2.5 months after the spill). b5 and b3 indicates the detection of the OTU with the bc-5' and bc-3' fragments respectively based on the processed reads. In bold species identified with indval analyses as indicative for impacted sites. Named species are the most abundant within each BIN, in brackets the number of specimens identified to species level in the reference database.

Table 2. Betadiversity values, NMDS stress values and p-values for the comparison between control and impacted sites.

	Dataset		Betadiversity (Sorensen index)					Turnover (Simpson index)			Nestedness (Sorensen-Simpson index)		
Taxa	DNA fragment	Method	beta.sor	Adonis p-value	Adonis r^2	stress	*envfit p-value*	beta.sim	stress	*envfit p-value*	beta.sne	stress	*envfit p-value*
Metazoa	*cox1-3'*	*de novo* 3%	0.85	0.007	0.18	0.086	0.023	0.763	0.144	0.179	0.087	0.067	0.099
		de novo 10%	0.835	0.011	0.19	0.075	0.018	0.736	0.143	0.054	0.099	0.053	0.099
	cox1-5'	*de novo* 3%	0.833	0.004	0.17	0.122	0.016	0.743	0.149	0.089	0.09	0.042	0.158
		de novo 10%	0.821	0.005	0.18	0.12	0.036	0.731	0.118	0.144	0.09	0.035	0.107
	SSU	*de novo* 3%	0.775	0.007	0.21	0	0.107	0.726	0.108	0.036	0.049	0.032	0.35
Insecta	*cox1-3'*	*de novo* 3%	0.862	0.004	0.18	0.087	0.025	0.776	0.143	0.035	0.085	0.07	0.382
		de novo 10%	0.849	0.005	0.19	0.088	0.009	0.761	0.144	0.028	0.088	0.066	0.37
		BOLD ref	0.857	0.014	0.17	0.109	0.108	0.758	0.146	0.033	0.099	0.096	0.536
	cox1-5'	*de novo* 3%	0.841	0.03	0.16	0.14	0.153	0.749	0.154	0.328	0.092	0.063	0.369
		de novo 10%	0.818	0.019	0.17	0.108	0.066	0.739	0.115	0.226	0.079	0.058	0.314
		BOLD ref	0.846	0.023	0.16	0.154	0.089	0.759	0.162	0.041	0.087	0.058	0.196
	SSU	*de novo* 3%	0.805	0.037	0.16	0.173	0.132	0.739	0.161	0.057	0.066	0.059	0.62
Crustacea	*cox1-3'*	*de novo* 3%	0.801	0.002	0.23	0.107	0.013	0.684	0.122	0.001	0.118	0.069	0.656
		de novo 10%	0.773	0.004	0.28	0.078	0.009	0.67	0.133	0.006	0.103	0.081	0.779
		BOLD ref	0.826	0.004	0.29	0.089	0.001	0.753	0.077	0.001	0.073	0.181	0.669
	cox1-5'	*de novo* 3%	0.762	0.002	0.38	0.085	0.006	0.658	0.095	0.003	0.104	0.047	0.102
		de novo 10%	0.763	0.006	0.43	0.098	0.004	0.663	0.073	0.005	0.1	0.079	0.196
		BOLD ref	0.797	0.006	0.34	0.091	0.004	0.725	0.073	0.009	0.072	0.132	0.475
	SSU	*de novo* 3%	0.746	0.005	0.32	0.057	0.004	0.675	0.103	0.011	0.07	0.076	0.223
Arthropoda	*cox1-3'*	*de novo* 3%	0.845	0.005	0.19	0.113	0.011	0.77	0.136	0.179	0.075	0.061	0.26
		de novo 10%	0.832	0.007	0.20	0.092	0.017	0.758	0.135	0.015	0.074	0.043	0.317
		BOLD ref	0.856	0.005	0.19	0.104	0.113	0.779	0.128	0.006	0.077	0.095	0.569
	cox1-5'	*de novo* 3%	0.829	0.005	0.18	0.142	0.002	0.746	0.143	0.003	0.083	0.064	0.457
		de novo 10%	0.82	0.002	0.18	0.139	0.017	0.75	0.142	0.013	0.07	0	0.236
		BOLD ref	0.843	0.004	0.18	0.171	0.006	0.771	0.164	0.02	0.072	0.059	0.242
	SSU	*de novo* 3%	0.787	0.001	0.22	0.095	0.006	0.745	0.118	0.004	0.042	0.059	0.798

Notes: *de novo* 3% and *de novo* 10% refers to OTU clustering at these threshold values. BOLD ref indicates the results from the 'taxon dependent' approach of mapping reads against OTU clusters generated from BOLD data.

Table 3. Species with indicative value as identified by *indval* analyses based on the results of the reference database-dependent approach for the *cox1* gene fragments bc-5' and bc-3'.

BIN	Class	Order	Family	Species	GENE	T	Ind. value	P
BOLD:AAF2659	Branchiopoda	Diplostraca	Chydoridae	*Chydorus sphaericus*	bc-5'	I	0.6	0.046
BOLD:AAF2659	Branchiopoda	Diplostraca	Chydoridae	*Chydorus sphaericus*	bc-5'	T2-I	1	0.014
BOLD:AAP5931	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [1]*	bc-5'	T2-I	1	0.024
BOLD:ACU8677	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [1]*	bc-5'	T2-I	1	0.022
BOLD:AAI6018	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [123]*	bc-3'	I	0.6	0.048
BOLD:AAI6018	Insecta	Diptera	Chironomidae	*Cricotopus bicinctus [123]*	bc-3'	T2-I	1	0.023
BOLD:AAE4568	Insecta	Diptera	Chironomidae	*Eukiefferiella claripennis [185]*	bc-5'	T2-I	1	0.019
BOLD:AAO1037	Insecta	Diptera	Chironomidae	*Paratendipes albimanus [127]*	bc-3'	I	0.8	0.024
BOLD:AAO1037	Insecta	Diptera	Chironomidae	*Paratendipes albimanus [127]*	bc-5'	I	0.8	0.014
BOLD:ACQ8988	Insecta	Diptera	Chironomidae	*Tanytarsus brundini [13]*	bc-3'	I	0.8	0.02
BOLD:ACQ8988	Insecta	Diptera	Chironomidae	*Tanytarsus brundini [13]*	bc-5'	I	0.8	0.02
BOLD:AAB9119	Insecta	Diptera	Chironomidae	*Tanytarsus brundini [5]*	bc-5'	I	0.8	0.017
BOLD:AAW1102	Insecta	Diptera	Chironomidae	*Tanytarsus ejuncidus [24]*	bc-3'	I	0.75	0.049
BOLD:AAU4439	Insecta	Diptera	Chironomidae	*Tanytarsus eminulus [124]*	bc-3'	I	1	0.002
BOLD:AAU4439	Insecta	Diptera	Chironomidae	*Tanytarsus eminulus [124]*	bc-5'	I	0.86	0.015
BOLD:ACR3318	Insecta	Diptera	Chironomidae	*Tanytarsus pallidicornis [10]*	bc-3'	I	0.86	0.015
BOLD:AAU1007	Insecta	Ephemeroptera	Baetidae	*Centroptilum luteolum [17]*	bc-5'	T2	1	0.001
BOLD:ACH6832	Malacostraca	Amphipoda	Gammaridae	*Gammarus pulex [11]*	bc-3'	C	0.83	0.016
BOLD:ACH7960	Malacostraca	Amphipoda	Gammaridae	*Gammarus nekkensis [1]*	bc-5'	C	1	0.004
BOLD:ACG8343	Malacostraca	Amphipoda	Gammaridae	*Gammarus fossarum [37]*	bc-5'	C	0.83	0.025
BOLD:AAA1971	Malacostraca	Isopoda	Asellidae	*Asellus aquaticus [85]*	bc-3'	I	1	0.004
BOLD:ACV6778	Malacostraca	Isopoda	Asellidae	*Asellus aquaticus [7]*	bc-3'	I	0.8	0.01
BOLD:AAA1971	Malacostraca	Isopoda	Asellidae	*Asellus aquaticus [85]*	bc-5'	I	0.86	0.019

Notes: T: Treatment; C: Control sites, I: impacted sites; T2: collection period 2 (2.5 months after the spill); T2-I: Impacted sites at collection period 2. Named species are the most abundant within each BIN, in brackets the number of specimens identified to species level in the reference database.