Author/s:
Carew, ME;Metzeling, L;St Clair, R;Hoffmann, AA

Title:
Detecting invertebrate species in archived collections using next-generation sequencing

Date:
2017-09-01

Citation:
Carew, M. E., Metzeling, L., St Clair, R. & Hoffmann, A. A. (2017). Detecting invertebrate species in archived collections using next-generation sequencing. Molecular Ecology Resources, 17 (5), pp.915-930. https://doi.org/10.1111/1755-0998.12644.

Persistent Link:
https://hdl.handle.net/11343/292493

Received Date : 06-Sep-2016
Revised Date : 15-Nov-2016
Accepted Date : 06-Dec-2016
Article type : Resource Article

Detecting invertebrate species in archived collections using next generation sequencing

Melissa E. Carew*, Leon Metzeling †, Rosalind St Clair †, and·Ary A. Hoffmann*

*Victorian Centre for Aquatic Pollution Identification and Management (CAPIM), School of BioSciences, Bio21 institute, 30 Flemington Rd, The University of Melbourne 3010, Victoria, Australia
†Environment Protection Authority Victoria, Ernest Jones Drive, Macleod, 3085, Victoria, Australia

Keywords: Invertebrates, DNA barcoding, Insect, Aquatic, Metagenomic
Correspondence: Melissa Carew
Postal Address: Bio21 institute, 30 Flemington Rd, The University of Melbourne, Victoria, Australia, 3010.
Phone: +61 3 8344 2314
Fax: +61 3 8344 2279
E-mail: mecarew@unimelb.edu.au

Running title: Detect invertebrates with NGS

Abstract

Invertebrate biodiversity measured at mostly family level is widely used in biological monitoring programs to assess anthropogenic impacts on ecosystems. However, next generation sequencing (NGS) could allow development of new more sensitive

biomonitoring tools by allowing rapid species identification. This could be accelerated if archived invertebrate collections and environmental information from past programs are used to understand species distributions and their environmental responses. In this study, we take archived macroinvertebrate samples from two sites collected on multiple occasions and test if NGS can successfully detect species. Samples had been stored in 70% ethanol at room temperature for up to 12 years. Three amplicons ranging from 197-274 bps within the DNA barcode region were amplified from samples and compared to DNA barcoding libraries to identify species. We were able to amplify partial DNA barcodes from most samples, and species were often detected with multiple amplicons. However, some singletons and taxa poorly covered by DNA barcoding were missed. This suggests additional DNA barcodes will be required to fill 'gaps' in current DNA barcode libraries for aquatic macroinvertebrates and/or that it may not be possible to detect all taxa in a sample. Furthermore, older samples often detected fewer taxa and were less reliable for amplification, suggesting NGS is best used on samples within eight years of collection. Nevertheless, many common taxa with existing DNA barcodes were reliably identified with NGS and were often present at sites across multiple years, showing the potential of NGS for detecting common and abundant species in archived material.

Introduction

Invertebrate biodiversity is widely used in monitoring programs to assess impacts of anthropogenic activities on terrestrial (Kremen *et al.* 1993; Cortet *et al.* 1999; Hilty & Merenlender 2000; Andersen *et al.* 2004) and aquatic (Rosenberg & Resh 1993; Bunn 1995; Fore *et al.* 1996; Bonada *et al.* 2006; Chessman *et al.* 2007; Kellar *et al.* 2014) ecosystems.

Most monitoring programs gather environmental information along with invertebrates to assist in understanding factors leading to changes in biodiversity. Generally, invertebrate communities are only identified to order or family levels in terrestrial studies (e.g. Lovell *et al.* 2007; Neville & Yen 2007; Thomson & Hoffmann 2007; Sharley *et al.* 2008) or mostly to family level in aquatic studies, (e.g. Barbour *et al.* 1992; Chessman 1995; Resh *et al.* 1995; Wright *et al.* 2000; Wells *et al.* 2002) due to the time, expertise, expense and difficultly involved in routinely identifying all taxa to genus or species levels. Once programs are concluded, invertebrate samples can be archived by storing in 70% ethanol at room temperature.

DNA sequencing of invertebrates in archived collections could allow species to be easily identified, giving insight into species temporal variation where sites have been

visited on multiple occasions, and into species distributions along environmental gradients where there is also environmental information available. Given that genera and species from the same family can show different environmental responses (Resh & Unzicker 1975; Lenat & Resh 2001; Mandaville 2002; Carew *et al.* 2007b; Carew *et al.* 2011b), identifying species in archived collections could enable development of new species-based biomonitoring tools to provide a deeper understanding of environmental condition.

Currently, DNA barcodes based on the mitochondrial Cytochrome Oxidase I (COI) sequences are widely used in species identification (Hebert *et al.* 2003). In archived material, DNA barcodes could be bulk sequenced using next generation sequencing (NGS) and these sequences then compared to reference DNA barcodes generated from voucher specimens found in DNA barcode databases for identification (see Zimmermann *et al.* 2014). In recent years, the number of invertebrate DNA barcodes available has rapidly expanded (Ratnasingham & Hebert 2007; Stockle & Hebert 2008). Currently, there are over 5.6 million invertebrate DNA barcodes representing over 266,000 species available through the BOLD system v3 database (http://www.boldsystems.org/ - Accessed on 30-Aug-2016), which could be used for routinely identifying species in bulk extracted invertebrate samples sequenced with NGS.

Generating DNA barcodes from freshly collected material is often successful for many taxa, but DNA sequencing of archived material can be problematic, often due to storage at room temperature in 70% ethanol that can result in DNA degradation (Hajibabaei *et al.* 2005; Zimmermann *et al.* 2008; Baird *et al.* 2011). In reality, many archived samples may encounter even lower ethanol percentages due to insufficient ethanol being added when samples are collected, evaporation of ethanol over time or storing too many individuals together, leading to further DNA degradation. To overcome this problem, 'mini DNA barcodes' (Hajibabaei *et al.* 2006b) or DNA barcodes of <300 bps (e.g. Zeale *et al.* 2011; Hajibabaei *et al.* 2012; Leray *et al.* 2013; Clarke *et al.* 2014) can be used to produce partial DNA barcodes for species identification. Partial DNA barcodes have been shown to be successful for identifying taxa with NGS (e.g. Hajibabaei *et al.* 2011; Carew *et al.* 2013; Zhou *et al.* 2013; Kermarrec *et al.* 2014), but have not been tested for bulk species identification from archived material.

Here we test if macroinvertebrate samples identified to family level from two sites collected over a 12-year period as part of a long-term monitoring program can be identified to species with NGS. Three short amplicons (<300 bp) within the DNA barcode region were amplified due the likely degraded condition of the DNA in samples. We use the

Illumina MiSeq platform to test whether species from different orders and families can be detected in samples of different ages.

Methods

Study design

Our study focused on two locations with macroinvertebrate samples collected on five occasions spanning 12 years. The samples were collected by Environment Protection Authority (EPA) Victoria as part of their long-term monitoring program. The first site was located in the Wimmera River at Riverside upstream of the city of Horsham, Victoria, Australia (IHC), while the second site was in the Franklin River at Toora, Victoria, Australia (TUF) (Table 1). These sites were chosen based on the availability of samples from multiple sampling events and macroinvertebrate diversity. We selected the IHC site as being typical of a low family diversity site, and the TUF site as being typical of a high family diversity site. Samples were identified mostly to family level according to the Rapid Bioassessment Protocols developed by EPA Victoria as part of their Guidelines for Environmental Management (available at: http://www.epa.vic.gov.au/~/media/Publications/604%201.pdf). In addition, three samples from the Wimmera River (IHC04, IHC08, IHC10) were also identified to lowest taxonomic unit, which included some species level identifications. All samples were collected as part of a spring sampling program with the exception of the 2009 and 2011 samples from Franklin River, which came from an autumn sampling program. We also included a chironomid sample (RL09) as a control because all individuals in this sample had previously been identified to species level and bulk sequenced using 454 pyrosequencing (see Carew *et al.* 2013). The purpose of this control was to examine data quality and to help in testing a bioinformatic pipeline for data analysis.

Sample processing and DNA extraction

For each sample, taxa were dissected with the aim of maximising the detection of species diversity. Therefore, where possible we dissected a similar amount of tissue from each individual in an attempt to prevent large taxa out-competing small taxa in PCR amplifications. Depending on the taxonomic group, we used legs (if present) or small amounts abdominal/body tissue. In some taxa, such as Chironomidae and Oligochaeta, we were unable to avoid including the gut in DNA extractions. Each sample was dissected in a new petri dish to avoid cross contamination between samples. Tissues from each individual were placed into a 1.5 ml microcentrifuge tube for bulk extraction and PCR, and

subsequent Illumina MiSeq sequencing. A Qiagen DNeasy® Tissue Kit (Qiagen, Hilden, Germany) was used to extract total genomic DNA from tubes of tissues following the manufacturer's protocol. All DNA extractions were performed in an area separated from post PCR processing, to reduce the risk of contamination from PCR amplicons.

We included a biological replicate of one sample (TUF11) to act as a control for the robustness of species identification of NGS in our dataset. We used sample TUF11 due to its high family diversity and the fact that it was the most recently collected sample. Therefore, this sample was likely to indicate differences in species detection within the same sample and also indicate any contamination by PCR amplicons.

Amplicon library construction and sequencing

A two-step PCR process was used to obtain amplicons for illumina MiSeq sequencing to increase primer specificity and yield of amplicons from our samples. The first PCR involved amplifying the three amplicons of <300 bp within the Cytochrome Oxidase I DNA barcode region (Hebert *et al.* 2003). For this purpose, we used three primer sets (Table 2). The first set of primers, COIBfor/HCOI, amplifies the 3' end of the DNA barcode fragment, while the second set of primers, ZBJ-ArtF1c/ZBJ-ArtR2c and LCOI/MLepR21, overlap and amplify the 5' end of the DNA barcode fragment. These primers were selected based on their ability to amplify a broad range of taxa when screened against a taxonomically diverse panel of aquatic macroinvertebrates (data not shown). Template specific primers had Illumina adaptors incorporated onto the 5' end for the attachment of Nextera-XT Illumina indexes (Illumina Corporation, San Diego, CA, USA) in the second round PCR. All PCRs were set up in a pre-PCR laminar flowhood, including negative controls (no DNA), with template DNA added just prior to PCR amplification to minimise the risk of contamination.

First round PCR reactions contained 2 µl of DNA template, 16.4 µl molecular biology grade water, 2.5 µl PCR buffer, 1 µl $MgCl_2$ (50 mM), 2 µl dNTPs mix (25 mM), 0.5 µl forward primer (10 mM), 0.5 µl reverse primer (10 mM), and 0.1 µl Platinum Taq polymerase (5 U/ml) (Invitrogen, Carlsbad, CA, USA) in a total volume of 25 µl. First round PCR's for the COIBfor/HCOI primer set used the following conditions: 94°C for 3 min followed by 25 cycles of 94°C for 20 sec, 48°C for 45 sec, 72°C for 30 sec, then 1 cycle of 72°C for 5 min For the ZBJ-ArtF1c/ZBJ-ArtR2c and LCOI/MLepR21 primer sets the same conditions were used except that an annealing temperature of 46°C was used.

Amplifications were performed as three replicates. All amplicons were checked by agarose gel electrophoresis. The PCR replicates were pooled and cleaned using Sera-

mag customised magnetic beads mix (GE Healthcare Life Sciences, Little Chalfont, Buckinghamshire, United Kingdom), where 1.5 volumes of bead mix were added to pooled PCR mixes. Samples were mixed by gentle pipetting and then incubated at room temperature for 15 mins. After incubation, samples were placed on a magnetic stand for 5-10 mins until the solution had cleared. The supernatant was removed and the two washes using 200 µl of 80% ethanol were performed. Samples were then air dried to remove residual ethanol and were re-suspended in 30 µl of DPEC water. They were placed onto the magnetic stand until the solution had cleared and then the supernatant was transferred to a fresh tube for second round PCRs. Cleaned amplicons were quantified with a Qubit Fluorometer (Life Technologies, Carlsbad, CA, USA).

The second round PCRs were performed after pooling three amplicons in approximately equal molar ratios. Second round PCR reactions used 5-15 µl of the first round amplicons, 25 µl BIO-X-ACT short mix (Bioline, London, United Kingdom), 5 µl forward Nextera-XT index primer (10 mM), and 5 µl reverse Nextera-XT primer (10 mM). PCR conditions were as follows: 94°C for 5 mins followed by 12 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 30 sec, then 1 cycle of 72°C for 5 mins. All amplicons were cleaned and quantified ensuring concentrations were > 5 ng/µl as described above.

Library quantification, normalization, pooling and the Illumina MiSeq run using a 600 cycle flow cell MiSeq sequencing kit V3 (300 bp x 2) (Illumina, San Diego, CA) were performed by Australian Genome Research Facility Ltd (AGRF) according to the manufacturer's protocols.

Bioinformatic processing

Samples were first examined using FastaQC to determine their quality, and were then uploaded into Geneious version 8.1.6 (Kearse *et al.* 2012). Sequences from each sample were trimmed to <197 bp, paired reads were set and FLASH version 1.2.9 (Magoc & Salzberg 2011) was used to merge paired reads with default settings. Merged reads were annotated with each set of forward and reverse primer pairs and then extracted to isolate each of the three amplicons. Only amplicons of appropriate size (+/- 6 bp of the amplicon's expected size) with both forward and reverse primers on each end were retained. Primers were then trimmed and a custom *de novo* assembly at 98% similarity was used to reduce redundancy in the dataset. The number of reads that contributed to each contig (group of sequences with > 98% sequence similarity) from the *de novo* assembly was recorded. Singletons and contigs containing ten or fewer reads were discarded (see Bokulich *et al.* 2012). The remaining contigs were aligned, edited and checked for an open reading

frame. Contigs were then BLAST searched against a DNA barcoding reference database of freshwater macroinvertebrates (see below), and species with > 97.5% match were identified. The ten best matches were also checked to ensure that species matches were specific. Chimeric sequences were identified by using BLAST alignments to the most common sequences in a sample (those with higher than 200 reads) and full length DNA barcodes from the reference library using a '*de novo*' approach. Those that showed clear chimeric patterns were removed.

DNA barcoding reference database

A DNA barcoding reference database of freshwater macroinvertebrates was constructed by downloading and combining DNA barcodes from three sources; GenBank (http://www.ncbi.nlm.nih.gov/genbank/), BOLD systems v3 (http://www.boldsystems.org/) and private databases for the Chironomidae (M. Carew), the Trichoptera (courtesy of M. Shackleton, D. Cartwright, J. Dean and R. St Clair) and several macroinvertebrate specimens from the Maribyrnong River, Melbourne, Australia (M. Carew). Only aquatic families were included in library construction to reduce the size of the database, improve specificity and shorten search times. The database was uploaded into Geneious version 8.1.6 and used in custom BLAST searches; any taxa that did not have a match to sequences in our freshwater macroinvertebrate DNA barcoding reference database were searched on the entire GenBank and BOLD system v3 sites. Due to taxonomic limitations or reference specimens that were not identified to species, our database contains some taxa that are identified to family/order. This included taxa from the Maribyrnong River which are listed at family/order level followed by "undet." given they have not been formally identified. The BOLD system v3 contained DNA barcodes with multiple species names and specimens that were not identified to species level. For these taxa we have also cited their BOLD: BIN URI code (which consist of three letters and four numbers) as this suggests possible species grouping and avoids taxonomic confusion. Chironomidae were identified to mostly species; some species are designated as "sp." followed by a letter or number to represent a species that could not be identified using taxonomic keys.

Results

Presence of macroinvertebrate families

The archived macroinvertebrate samples varied in their family composition. Macroinvertebrate diversity at the Wimmera River site was often less diverse (<24 families) than at the Franklin River site (>22 families) (Table 1). At the Wimmera River site

14 families were found on a single occasion, compared to 11 families for the Franklin River samples (Fig 1). Families not found on multiple occasions tended to be represented by one or a few individuals. Families like the Dytiscidae (Coleoptera), Chironomidae (Diptera), Corixidae (Hemiptera) and Leptoceridae (Trichoptera) were found at both sites in all samples, while the Physidae (Gastropoda), Atyidae (Crustacean) and Veliidae (Hemiptera) occurred at <4 samples at both sites. In addition, the Notonectidae (Hemiptera) and Coenagrionidae (Odonata) were present in all samples from the Wimmera River and the Hydrophilidae and Elmidae (both Coleoptera), and Baetidae and Leptophlebiidae (both Ephemeroptera) were present in all samples from the Franklin River. In some samples certain families were missing as they had been removed for morphological examination prior to this study (Fig 2). These taxa are also indicated in Table S1 for the Wimmera River site and Table S2 for the Franklin River site. Therefore, tissues from these families were not available for NGS.

Amplification success of archived macroinvertebrate samples

We obtained amplicons for eight of the ten archived macroinvertebrate samples - the Wimmera River 2006 (IHC06) and Franklin River 2002 (TUF02) samples did not produce amplicons (Table 1). Furthermore, the longest amplicon (LCOI/MLepR2) did not amplify from the Wimmera River 2004 (IHC04) sample. Similarly, a very low number of LCOI/MLepR2 amplicons were recovered from the 2009 Franklin River (TUF09) sample.

Reads (individual sequences) recovered for each sample varied with amplicon size. The smallest COIBfor/HCOI amplicon produced the most reads (Table 3). Following *de novo* assembly of sequences, each amplicon produced between 3 and 100 contigs per sample. Generally, more contigs were recovered from the smaller amplicons.

Species identification using NGS

In the chironomid control sample (RL09), all expected chironomid species were identified at >97.5% sequence match using BLAST searches, except for *Chironomus tepperi* which was represented by a single individual (Table 4). We found a large number of chimeric sequences in RL09 (data not shown), which were identified and removed by manually checking BLAST alignments. Additional species which were not found in our samples, a microcrustacea, the oligochaete *Chaetogaster diaphansous,* and an additional chironomid taxon, *Paratrichocladius* sp. 3*,* were also detected with one primer set. Typically, species confirmed to be in the sample based on individual examination were detected with multiple amplicons and a high number of reads.

From the archived collections, we found most contigs showed >97.5% sequence match to species in our freshwater DNA barcode library, but some contigs from each amplicon could not be matched at >97.5% (Table 3). For all samples except IHC04 where only two primer sets amplified, matches for most species involved multiple amplicons (Fig 3). Almost all matches of >97.5% were to a single taxon. However, a few contigs matched more than one taxon. In samples IHC08, IHC10, TUF08 and TUF11 all primer sets produced a contig that had a 100% match to an undescribed Hemipteran species from the BOLD database and *Procladius paludicola* (Chironomidae). This also occurred in samples IHC12 and RL09 with a second undescribed Hemipteran species from the BOLD database and *Chironomus februarius* (Chironomidae). We also found the percentage match between the COIBfor/HCOI amplicon differed to those found with the other two amplicons for a few species. For example, a COIBfor/HCOI contig had a 95% match and the Zeale and LCOI/MLepR2 amplicons showed >98% match to a *Gerris* sp. (Gerridae, Hemiptera), suggesting amplicons from different parts of the DNA barcode region can show different sequence similarity. Furthermore, NGS sequences matching the *Gerris* sp. were more likely to be from *Microvelia oceanica* (Veliidae, Hemiptera), as there were no Gerridae in these samples. This suggests a possible error in morphological identification in the DNA barcode library, or species from these similar Hemipteran families that share related DNA barcodes. We also found matches using the COIBfor/HCOI amplicon to the endosymbiont, *Wolbachia,* in samples TUF07, TUF08 and TUF09, and to zooplankton in sample IHC04, suggesting this primer set will amplify non-target species. A summary of the taxa with matches to those from BOLD system v3 database can be found in Table S3.

When we compared morphological identifications at the lowest taxonomic level of the Wimmera River samples IHC04, IHC08 and IHC10 to those from NGS analysis, we found that almost all species identified morphologically were detected if they had a matching DNA barcode in our DNA barcode library (Table 5). Many of the morphologically identified genera were identified to species using DNA barcodes, and in some instances particular genera, such as *Ischnura* and *Polypedilum,* were represented by multiple species. We were also able to determine that some morphologically identified species matched DNA barcodes from the BOLD database even though these did not have species names. For example *Hemicordulia tau* matched the DNA barcode 'Corduliidae sp.ACA3305' from the BOLD system v3 database. Species, such as *Microvelia oceanica*, Ceratopogonidae spp., *Austrogomphus* sp. and *Oecetis* sp., showed matches of <97.5% to other related taxa in our database, indicating that even though we did not have these species in our DNA barcoding database, we could match them to related species.

The NGS proved reliable, with similar results from biological replicate samples from site TUF11 (Fig 1, Table S2). The detection of one species of the 40 species found in the TUF11 samples differed between the replicates; the chironomid, *Cricotopus parbicinctus*, was detected in only one replicate with a single primer set with a low number of reads (Fig 2, Table S2).

The amplification of additional taxa in our control suggests it was possible that primer sets targeting smaller amplicons may amplify DNA from other sources such as gut content. As in the case of the control sample, we found two instances where the COIBfor/HCOI primer set amplified *Paratya australiensis* (Atyidae) in sample IHC08, as well as *Procladius paludicola* (Tanypodinae) and *Tanytarsus inextentus* (Tanytarsini) in sample IHC10, despite these families/tribes not being recorded as present based on morphological examination. However, these accounted for < 5% of taxa identified with a single amplicon, showing that most species identified by a single amplicon are likely to represent species in samples (Fig 3). Taxa that amplified for multiple amplicons were always representative of families found in the sample, except in the oldest sample (IHC04). In this sample we found the presence of chironomid sequences, despite these animals being removed from this sample prior to our study. There were sequences from eight chironomid species that matched genera found during earlier morphological identification (Table 5), suggesting residual DNA from these species may have remained in the sample. However, we also detected a *Cricotopus* species from the Orthocladiinae (not recorded for this sample based on morphological identification) with two amplicons. We were unable to verify whether this represented a misidentification in the original morphological study or a possible gut contamination. We did not detect any other taxa that had been removed from a sample.

Species diversity and temporal persistence in samples

The greatest species diversity was found in the Chironominae (Diptera) with 20 species detected; the Leptoceridae (Trichoptera) were also diverse with 11 species detected (Fig 4). These groups were amongst the most abundant, with greater than 100 individuals in samples. We also had high numbers of Corixidae in samples, but species from *Micronecta* were only identified to genus due the cryptic diversity that was not represented in our freshwater DNA barcode library. We found multiple species in the Leptophlebiidae and Baetidae (Ephemeroptera); Tanypodinae, Orthocladiinae and Simulidae (Diptera); Coenagrionidae (Odonata); Calamoceratidae, Conoesucidae, Ecnomidae, Hydrobiosidae, Hydroptilidae and Philorheithridae (Trichoptera) and the order Oligochaeta.

We often found different species within the same families when we compared the two sites. For example, the Chiltoniidae (formerly the Ceinidae in the Amphipoda) from the Wimmera River was represented by *Austrochiltonia subtenuis,* while at the Franklin River, *Austrochiltonia australis* and an unknown *Austrochiltonia* species were found. Similarly, species composition of Trichoptera, Ephemeroptera and Chironomidae differed between the two sites (Table S1 and S2). However, for some families the same species were found at both sites. The introduced *Physa acuta,* the only known member of the Physidae (Gastropoda) in Australia, was found at both sites (Table S1 and Table S2).

'Gaps' in the DNA barcode libraries

We did not detect species with NGS from all families identified by morphological examination. Samples from the Wimmera River had substantially better coverage with taxa detected from >80% of families present compared to the Franklin River, which had greater family diversity and often more individuals (Table 1). The oldest samples from each site had the least number of species from families detected, suggesting fewer taxa amplified in older samples.

Species could only be detected and identified if a corresponding DNA barcode was present in our freshwater macroinvertebrate DNA barcode library. To determine if some taxa were missing from our DNA barcode library, we attempted to go back to our archived samples and amplify individuals for Sanger sequencing from families where we did not detect a species with NGS. Unfortunately, we could not recover full or partial DNA barcodes from any of these individuals. DNA barcodes from individuals would have allowed us to fill 'gaps' in our freshwater macroinvertebrate DNA barcode library for the two sites. We repeatedly failed to detect species from the Acarina, Coleoptera (Hydrophilidae and Elmidae), Hemiptera (Notonectidae and Veliidae), Bivalvia (Sphaeriidae), Amphipoda (Chiltoniidae), Emphemeroptera (Caenidae), Gastropoda (Lymnaeidae), Diptera (Ceratopogonidae, Dixidae and Tipulidae), Plectoptera (Gripopterygidae), Oligochaeta, and Odonata (Synlestidae) (Fig 2). It is possible that these taxa and other families where we did not detect a species represent broader gaps in freshwater DNA barcoding library in Australia. To examine the extent of these gaps, we examined the BINs (Barcode Index Numbers) which represent the possible species based on grouping of individuals with high sequence similarity and the number of DNA barcoded species identified from the BOLD system v3 database, and compared this to the estimated species diversity for the orders found in this study from Australia (Fig 5). We found species DNA barcode coverage was low compared to estimated species diversity for the Acarina,

Coleoptera, Diptera, Hemiptera, Plecoptera, Odonata and Turbellaria, indicating it was possible species from study sites belonging to these orders were not represented in our DNA barcoding library. This was supported when we directly searched our DNA barcode library for species names from morphologically identified individuals from the Wimmera samples IHC04, IHC08 and IHC10 (Table 5).

While gaps in DNA barcode libraries explain some missing taxa, we also examined whether particular primer pairs may have failed to produce amplicons. We found the greatest number of taxa were detected with the COIBfor/HCOI primer pair which detected 80% of species found in the study followed by the Zeale primer pair which detected 77% of species. The LCOI/MLepR2 primer pair detected the least number of taxa (56% of species) and appeared not to amplify some species, particularly those from the Amphipoda and some Oligochaeta. It also failed to amplify some insect species detected which the other two amplicons detected, highlighting the importance of using multiple primer pairs. Without complete DNA barcode libraries to identify all contigs produced from NGS, it is difficult to determine if some taxa were not detected due to lack of amplification or representation in DNA barcode libraries. However, the presence of unmatched contigs does suggest some additional species are present, possibly from families where we did not detect any species. Some of these taxa showed multiple matches to species at < 97.5% from particular families (noted in Table S1 and Table S2).

Discussion

Our study showed that NGS using the Illumina MiSeq platform successfully and reliably detected macroinvertebrate species in archived material. We were able to detect species from most families in samples and most of these species were detected with multiple amplicons. In all but one instance, species were detected with multiple amplicons and were representative of families found via morphological examination. Our study confirms the effectiveness of NGS for detecting multiple species in mixed environmental samples (Tedersoo *et al.* 2010; Hajibabaei *et al.* 2011; Carew *et al.* 2013; Aylagas *et al.* 2014; Gibson *et al.* 2014; Kermarrec *et al.* 2014) and the effectiveness of the MiSeq platform for sequencing DNA barcodes (Shokralla *et al.* 2015).

While more recently-collected archived material amplified well with the primers used in this study, amplification of material older than eight years proved less reliable. This corresponds with Zimmermann *et al.* (2008) who showed amplifications of up to 500 bps were often unsuccessful on archived material greater than ten years old due to rapid

fragmentation of DNA under poor preservation conditions. Furthermore, Baird *et al.* (2011) found that samples stored for <1 to 23 years by formalin fixation and subsequent long-term preservation in 70% ethanol showed poor amplification success, with <17.5% of samples producing DNA barcodes. Most of these were mini DNA barcodes (Hajibabaei *et al.* 2006b) of 137 bp with only 2.9% producing full length DNA barcodes. While the material collected in our study was not subject to formalin fixation, long-term storage in 70% ethanol at room temperature appeared to have a negative effect on DNA quality. This was evident when we attempted to amplify individuals for short DNA barcodes of <300 bp with the three amplicons used for NGS. This shows that bulk extraction, amplification and sequencing using NGS is better able to produce partial DNA barcodes from archived material subject to long-term storage using poor DNA preservation methods. Unlike Sanger DNA sequencing, NGS technologies can DNA sequence individual PCR amplicons (Shokralla *et al.* 2012) making NGS far less sensitive to low amplicon concentrations common in poorly preserved material. However, if material stored in 70% ethanol was less than six months old it is possible for full DNA barcodes to be reliably produced (Stein *et al.* 2013). This means that more recently archived material may be successfully used for both NGS and producing full length DNA barcodes for individuals not in current DNA barcode libraries.

While we were able to detect species that were either large or represented by multiple individuals, we found it difficult to detect taxa that were small, singletons or not well represented in our DNA barcoding library. Small taxa and/or singletons are more likely to be missed during PCR amplification and sequencing (Hajibabaei *et al.* 2011; Hajibabaei *et al.* 2012). This was evident in our control sample where we did not detect the chironomid, *Chironomus tepperi,* which was represented by a single individual. Using 454 pyrosequencing of this sample, Carew *et al.* (2013) were able to detect this taxon but with a very low number of reads. However, the remaining six chironomid species represented by singletons were all reliably detected using the Illumina MiSeq platform. The failure to detect small and/or singleton individuals may be affected by the overall species diversity in samples. For example, we achieved poorer coverage of taxa from families in the more diverse Franklin River samples compared to the Wimmera River. While this result may also be attributed to missing taxa in DNA barcoding libraries, it is possible that some small and/or singleton taxa may not have amplified due to PCR bias or a lack of template to initiate PCR (Hajibabaei *et al.* 2011) which may have been exacerbated by high diversity and individual numbers. Currently multi-amplicon NGS studies, where DNA barcoding libraries are largely complete, have shown that ~90% of species can be detected from

invertebrate environmental samples where species diversity involves >40 species (Hajibabaei *et al.* 2012; Gibson *et al.* 2014), while species can be detected at >99% for environmental samples with <15 species (Carew *et al.* 2013). However, the effect of high sample diversity on detection rates remains to be empirically tested and is likely to be affected by primers selection and bias, the number of primers used, sequencing depth and the composition and abundance of species in samples.

Furthermore, it was difficult in some instances to determine whether a small and/or singleton species were present or represented DNA from other sources. While family identifications can be useful for determining if a taxon found using NGS is likely to be present in a sample, it is not always reliable when a species is detected with a low number of reads. Non-target species can be detected over a target species in a sample from the same family, as we found with the detection of *Paratrichocladius* sp.3 but not *C. tepperi* with a single primer pair in the chironomid control sample. Given the short length of our amplicons it is possible that non-target taxa may be amplified from gut content, given one primer sets was developed for gut content analysis (Zeale *et al.* 2011). Similarly, amplification of chironomid taxa from the IHC04 sample indicates DNA leaching from co-stored taxa (Shokralla *et al.* 2010) as the species found mostly corresponded to the genera found during morphological sorting. We only found four instances where sequences from macroinvertebrate species were detected in archived samples from families/tribes not found during morphological examination and these were mostly detected using the COIBfor/HCOI primer set.

Primer selection has a strong influence on the ability to detect taxa with NGS (Clarke *et al.* 2014). The use of multiple primer sets improves the ability to detect species (Hajibabaei *et al.* 2012; Gibson *et al.* 2014). This is increasingly important when NGS needs to capture a broad taxonomic cross-section of species. Taxa currently considered in aquatic monitoring include those from nine phyla, with the Arthropoda representing the most common and abundant group of freshwater macroinvertebrates (Chessman 1995). Taxonomic diversity can present a challenge for bulk amplification of DNA barcodes, but primers that amplify a diverse array of taxa for the DNA barcode region in environmental samples are widely available (Hajibabaei *et al.* 2012; Carew *et al.* 2013; Gibson *et al.* 2014; Shokralla *et al.* 2014; Shokralla *et al.* 2015). The primers used in this study were able to detect taxa from three phyla, but it is likely additional primer design will be needed for some groups and new universal primer sets developed specifically for 'difficult' taxa would be useful. Testing of DNA barcoding primers by Clarke *et al* (2014) suggested a

minimum of three amplicons would be advisable for capturing most taxonomic diversity in samples.

The lack of comprehensive DNA barcode libraries for Australian freshwater invertebrates restricted our ability to detect taxa from some families. Our inability to consistently detect species from some Dipteran, Hemipteran and Coleopteran families and the Oligochaeta and Acarina, likely resulted from species in these groups being absent from our freshwater DNA barcoding library. In Australia, taxonomic and geographic coverage of DNA barcoding of aquatic macroinvertebrates is still relatively poor with only some groups, such as the Trichoptera (e.g. Shackleton & Webb 2014) and Chironomidae (e.g. Carew *et al.* 2007a; Carew *et al.* 2011a; Carew *et al.* 2013; Carew & Hoffmann 2015) well covered in our study area of south eastern Australia. Currently, a single study from south-eastern Australia of over 500 individuals from 11 orders contributes the only broad taxonomic coverage of DNA barcoding for Australian aquatic macroinvertebrates (Shackleton & Rees 2015). Expansion of DNA barcoding coverage will be essential when performing routine species identification using NGS, as matching taxa to DNA barcoded reference specimens enables linkage to current taxonomic and environmental information (Baird & Hajibabaei 2012; Carew *et al.* 2013; Gibson *et al.* 2014; Zimmermann *et al.* 2014). In this study, we have highlighted taxonomic gaps in DNA barcoding libraries and these taxa can now be targeted for individual DNA barcoding. In particular, this study suggests that aquatic Coleoptera should be a focus in future DNA barcoding efforts. Not only are aquatic Coleoptera common and abundant in macroinvertebrate samples, they often show substantial variability in environmental response (Chessman *et al.* 2002).

While the DNA extraction method used in this study was slow as we aimed to maximise the detection of taxa without the destruction of specimens in samples, a rapid extraction method would be needed if samples are to be utilised on a larger scale (see Hajibabaei *et al.* 2012; Gibson *et al.* 2015). A rapid extraction method together with NGS would facilitate development of new biological monitoring tools based on species responses to environmental change (Hajibabaei *et al.* 2011; Baird & Hajibabaei 2012; Carew *et al.* 2013; Aylagas *et al.* 2014; Zimmermann *et al.* 2014). Environmental responses of taxa at family and order levels have been routinely used in environmental monitoring. For example, family biotic indices based on the sensitivities of taxa to pollution have been widely used to grade streams and river conditions (Hilsenhoff 1982; Armitage *et al.* 1983; Chessman & McEvoy 1997) and in some cases developed at lower taxonomic levels (Chessman *et al.* 2002). Identifying species with NGS from hundreds of archived macroinvertebrate collections, which are linked to substantial environmental information,

will allow species responses to be rapidly assessed. Archived collections allow this to be performed at both the spatial and temporal level for at least several years, particularly if collections like those in this study form part of long-term monitoring programs. This allows patterns of species occurrence to be understood and changes based on past events to be evaluated using existing samples.

Acknowledgements

The authors thank Michael Shackleton, David Cartwright and John Dean for the use of their DNA barcode information. We also thank Rahul Rane for his advice in the laboratory and on data analysis, and Isabel Valenzuela for assistance with primer selection. This study was funded primarily by the Australian Research Council through their Linkage and Fellowship schemes, with additional support from Melbourne Water Corporation, the Victorian Government and EPA Victoria.

References

Andersen AN, Fisher A, Hoffmann BD, Read JL, Richards R (2004) Use of terrestrial invertebrates for biodiversity monitoring in Australian rangelands, with particular reference to ants. *Austral Ecology* **29**, 87-92.

Armitage PD, Moss D, Wright JF, Furse MT (1983) The performance of new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running waters. *Water Research* **17**, 333-347.

Aylagas E, Borja A, Rodriguez-Ezpeleta N (2014) Environmental status assessment using DNA metabarcoding: Towards a genetics based Marine Biotic Index (gAMBI). *PLoS ONE* **9**, e90529.

Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology* **21**, 2039-2044.

Baird DJ, Pascoe TJ, Zhou X, Hajibabaei M (2011) Building freshwater macroinvertebrate DNA-barcode libraries from reference collection material: Formalin preservation vs specimen age. *Journal of the North American Benthological Society* **30**, 125-130.

Barbour MT, Plafkin JL, Bradley BP, Graves CG, Wisseman RW (1992) Evaluation of EPA's rapid bioassessment benthic metrics: Metric redundancy and variability among reference stream sites. *Environmental Toxicology and Chemistry* **11**, 437-449.

554 Bokulich N, Subramanian S, Faith J, Gevers D, Gordon J, Knight R, Mills D, Caporaso G
555 (2012) Quality-filtering vastly improves diversity estimates from Illumina amplicon
556 sequencing. *Nature Methods* **10**, 57-59.
557 Bonada N, Prat N, Resh VH, Statzner B (2006) Developments in aquatic insect
558 biomonitoring: A comparative analysis of recent approaches. *Annual Review of*
559 *Entomology* **51**, 495-523.
560 Bunn S (1995) Biological monitoring of water quality in Australia: Workshop summary and
561 future direction. *Australian Journal of Ecology* **20**, 220-227.
562 Carew ME, Hoffmann AA (2015) Delineating closely related species with DNA barcodes
563 for routine biological monitoring. *Freshwater Biology* **60**, 1545-1560.
564 Carew ME, Marshall SE, Hoffmann AA (2011a) A combination of molecular and
565 morphological approaches resolves species in the taxonomically difficult genus
566 *Procladius* Skuse (Diptera: Chironomidae) despite high intra-specific morphological
567 variation. *Bulletin of Entomological Research* **101**, 1-15.
568 Carew ME, Miller AD, Hoffmann AA (2011b) Phylogenetic signals and ecotoxicological
569 responses: potential implications for aquatic biomonitoring. *Ecotoxicology* **20**, 595-
570 606.
571 Carew ME, Pettigrove V, Cox RL, Hoffmann AA (2007a) DNA identification of urban
572 Tanytarsini chironomids (Diptera: Chironomidae). *Journal of the North American*
573 *Benthological Society* **24**, 586-599.
574 Carew ME, Pettigrove V, Cox RL, Hoffmann AA (2007b) The response of Chironomidae to
575 sediment pollution and other environmental characteristics in urban wetlands.
576 *Freshwater Biology* **52**, 2444-2462.
577 Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013) Environmental monitoring
578 using next generation sequencing: Rapid identification of macroinvertebrate
579 bioindicator species. *Frontiers in Zoology* **10**, 45-60.
580 Chessman B, Williams S, Besley C (2007) Bioassessment of streams with
581 macroinvertebrates: effect of sampled habitat and taxonomic resolution. *Journal of*
582 *the North American Benthological Society* **26**, 546-565.
583 Chessman BC (1995) Rapid assessment of rivers using macroinvertebrates: A procedure
584 based on habitat-specific sampling, family level identification and a biotic index.
585 *Australian Journal of Ecology* **20**, 122-129.
586 Chessman BC, McEvoy PK (1997) Towards diagnostic biotic indices for river
587 macroinvertebrates. *Hydrobiologia* **364**, 169-182.

588 Chessman BC, Traylor KM, Davis JA (2002) Family- and species-level biotic indices for
589 macroinvertebrates of wetlands on the Swan Coastal Plain, Western Australia.
590 *Marine and Freshwater Research* **53**, 919-930.
591 Clarke LJ, Soubrier J, Weyrich LS, Cooper A (2014) Environmental metabarcodes for
592 insects: in silicoPCR reveals potential for taxonomic bias. *Molecular Ecology*
593 *Resources* **14**, 1160-1170.
594 Cortet J, Vauflery AG-D, Poinsot-Balaguer N, Gomot L, Texier C, Cluzeau D (1999) The
595 use of invertebrate soil fauna in monitoring pollutant effects. *European Journal of*
596 *Soil Biology* **35**, 115-134.
597 Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of
598 mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.
599 *Molecular Marine Biology and Biotechnology* **3**, 294-299.
600 Fore LS, Karr JR, Wisseman RW (1996) Assessing invertebrate responses to human
601 activities: Evaluating alternative approaches. *Journal of the North American*
602 *Benthological Society* **15**, 212-231.
603 Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W,
604 Hajibabaei M (2014) Simultaneous assessment of the macrobiome and microbiome
605 in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings*
606 *of the National Academy of Sciences* **111**, 8007-8012.
607 Gibson JF, Shokralla S, Curry C, Baird DJ, Monk WA, King I, Hajibabaei M (2015) Large-
608 scale biomonitoring of remote and threatened ecosystems via high-throughput
609 sequencing. *PLoS ONE* **10**, e0138432.
610 Hajibabaei M, DeWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM,
611 Hebert PDN (2005) Critical factors for assembling a high volume of DNA barcodes.
612 *Philosophical Transactions of the Royal Society B - Biological Sciences* **360**, 1959-
613 1967.
614 Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006a) DNA barcodes
615 distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of*
616 *Sciences of the United States of America* **103**, 968-971.
617 Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental
618 barcoding: A next-generation sequencing approach for biomonitoring applications
619 using river benthos. *PLoS ONE* **6**, e17497.
620 Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006b) A
621 minimalist barcode can identify a specimen whose DNA is degraded. *Molecular*
622 *Ecology Notes* **6**, 959-964.

623 Hajibabaei M, Spall J, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a
624 freshwater benthic macroinvertebrate community through non-destructive
625 environmental barcoding of DNA from preservative ethanol. *BMC Ecology* **12**, 28-
626 38.
627 Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identification through
628 DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological*
629 *Sciences* **270**, 313-321.
630 Hilsenhoff WL (1982) Using a biotic index to evaluate water quality in streams. Wisconsin
631 Department of Natural Resources, Madison, Wisconsin, U.S.A.
632 Hilty J, Merenlender A (2000) Faunal indicator taxa selection for monitoring ecosystem
633 health. *Biological Conservation* **92**, 185-197.
634 Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A,
635 Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012)
636 Geneious Basic: An integrated and extendable desktop software platform for the
637 organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.
638 Kellar CR, Hassell KL, Long SM, Myers JH, Golding L, Rose G, Kumar A, Hoffmann AA,
639 Pettigrove V (2014) Ecological evidence links adverse biological effects to pesticide
640 and metal contamination in an urban Australian watershed. *Journal of Applied*
641 *Ecology* **51**, 426-439.
642 Kermarrec L, Franc A, Rimet F, Chaumeil P, Frigerio J-M, Humbert J-F, Bouchez A (2014)
643 A next-generation sequencing approach to river biomonitoring using benthic
644 diatoms. *Freshwater Science* **33**, 349-363.
645 Kremen C, Colwell R, Erwin T, Murphy D, Noss R, Sanjayan M (1993) Terrestrial
646 arthropod assemblages: their use in conservation planning. *Conservation biology*,
647 **7,** 796-808.
648 Lenat DR, Resh VH (2001) Taxonomy and stream ecology -The benefits of genus - and
649 species - level identifications. *Journal of the North American Benthological Society*
650 **20**, 287-289.
651 Leray M, Yang J, Meyer C, Mills S, Agudelo N, Ranwez V, Boehm J, Machida R (2013) A
652 new versatile primer set targeting a short fragment of the mitochondrial COI region
653 for metabarcoding metazoan diversity: Application for characterizing coral reef fish
654 gut contents. *Frontiers in Zoology* **10**, 34.
655 Lovell S, Hamer M, Slotow R, Herbert D (2007) Assessment of congruency across
656 invertebrate taxa and taxonomic levels to identify potential surrogates. *Biological*
657 *Conservation* **139**, 113-125.

Magoc T, Salzberg SL (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963.

Mandaville SM (2002) Benthic macroinvertebrates in freshwaters- Taxa tolerance values, metrics and protocols, p. 120 Soil and Water Conservation Society of Metro Halifax, Halifax, Canada.

Neville PJ, Yen AL (2007) Standardising terrestrial invertebrate biomonitoring techniques across natural and agricultural systems. *Australian Journal of Experimental Agriculture* **47**, 384-391.

Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355-364.

Resh VH, Norris RN, Barbour MT (1995) Design and implementation of rapid assessment approaches for water resource monitoring using benthic macroinvertebrates. *Australian Journal of Ecology* **20**, 108-121.

Resh VH, Unzicker JD (1975) Water quality monitoring and aquatic organisms: The importance of species identification. *Journal of the Water Pollution Control Federation* **47**, 9-19.

Rosenberg DM, Resh VH (1993) Introduction to freshwater biomonitoring and benthic macroinvertebrates. In: *Freshwater biomonitoring and benthic macroinvertebrates* (eds. Rosenberg DM, Resh VH), pp. 1-9. Chapman and Hall, New York, NY, USA.

Shackleton M, Rees G (2015) DNA barcoding Australian macroinvertebrates for monitoring programs: Benefits and current short comings. *Marine and Freshwater Research*. **67**, 80-390

Shackleton ME, Webb JM (2014) Two new species of *Calocoides* Neboiss 1984 (Trichoptera: Calocidae) from eastern Australia, with descriptions of the immature stages. *Austral Entomology* **53**, 444-457.

Sharley DJ, Hoffmann AA, Thomson LJ (2008) The effects of soil tillage on beneficial invertebrates within the vineyard. *Agricultural and Forest Entomology* **10**, 233-243.

Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M (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, Dobosz R, Janzen DH, Hallwachs W, Golding GB, Hajibabaei M (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports* **5**, 9687.

692 Shokralla S, Singer GA, Hajibabaei M (2010) Direct PCR amplification and sequencing of
693 specimens' DNA from preservative ethanol. *Biotechniques* **48**, 233-234.
694 Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing
695 technologies for environmental DNA research. *Molecular Ecology* **21**, 1794-1805.
696 Stein ED, White BP, Mazor RD, Miller PE, Pilgrim EM (2013) Evaluating ethanol-based
697 sample preservation to facilitate use of DNA barcoding in routine freshwater
698 biomonitoring programs using benthic macroinvertebrates. *PLoS ONE* **8**, e51273.
699 Stockle MY, Hebert PD (2008) Barcode of life. *Scientific American* **299**, 82-88.
700 Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E,
701 Chuyong G, Kõljalg U (2010) 454 Pyrosequencing and Sanger sequencing of
702 tropical mycorrhizal fungi provide similar results but reveal substantial
703 methodological biases. *New Phytologist* **188**, 291-301.
704 Thomson LJ, Hoffmann AA (2007) Effects of ground cover (straw and compost) on the
705 abundance of natural enemies and soil macro invertebrates in vineyards.
706 *Agricultural and Forest Entomology* **9**, 173-179.
707 Wells F, Metzeling L, Newall P (2002) Macroinvertebrate regionalisation for the use in the
708 management of aquatic ecosystems. *Environmental Monitoring and Assessment*
709 **74**, 271-294.
710 Wright JF, Sutcliffe DW, Furse MT (2000) Assessing the biological quality of fresh waters:
711 RIVPACS and other techniques. Freshwater Biological Association, Ambelside, UK.
712 Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G (2011) Taxon-specific PCR for
713 DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources* **11**,
714 236-244.
715 Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013) Ultra-deep
716 sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples
717 without PCR amplification. *GigaScience* **2**, 4.
718 Zimmermann J, Abarca N, Enke N, Skibbe O, Kusber WH, Jahn R (2014) Taxonomic
719 reference libraries for environmental barcoding: a best practice example from
720 diatom research. *PLoS ONE* **9**, e108793.
721 Zimmermann J, Hajibabaei M, Blackburn DC, Hanken J, Cantin E, Posfai J, Evans TC, Jr.
722 (2008) DNA damage in preserved specimens and tissue samples: A molecular
723 assessment. *Frontiers in Zoology* **5**, 18-31.
724
725 **Data Accessibility**

DNA sequences: NCBI SRA accession no. SRR3049836, SRR3049877, SRR3049892, SRR3049913, SRR3049926, SRR3049977, SRR304998 - SRR3049984.
NCBI BioProject: PRJNA306791

Author Contributions

MEC performed the molecular lab work, carried out the data analysis, developed the analysis pipeline, participated in the study design and drafted the manuscript. LM assisted in sourcing samples for the DNA reference database, participated in the study design and MiSeq experiment and helped to draft the manuscript. RS verified the taxonomic identifications from the BOLD database and helped to draft the manuscript. AAH refined the study design, performed statistical analysis and helped to draft the manuscript. All authors read and approved the final manuscript.

Supporting information

Table S1. Species identified using NGS from site IHC
Table S2. Species identified using NGS from site TUF
Table S3. Taxa from the BOLD Systems v3 database

Table 1. The site location, site code and date from which the samples in this study were collected. The number of taxa morphologically identified to mostly families (no. of families), the number of individuals (n) and the taxa recovered with NGS from the families where Chironomidae are identified mostly to subfamily and Oligochaeta and Acarina to order are given.

Site	Latitude	Longitude	Site code	Date	No. of Families	n	Taxa detected from families with NGS (% detected)
Wimmera River	-36.6948	142.2380	IHC04	09-Nov-04	17	95	12[a] (80%)
			IHC06	05-Dec-06	19	100	No amplification
			IHC08	07-Oct-08	19	91	18 (95%)
			IHC10	18-Oct-10	15	125	13 (86%)
			IHC12	07-Dec-12	24	179	21 (87%)
Franklin River	-38.6295	146.3088	TUF02	10-Dec-02	22	423	No amplification
			TUF07	08-Oct-07	23	145	10[b] (58%)
			TUF08	29-Oct-08	39	323	27 (69%)
			TUF09	24-Mar-09	36	254	24 (66%)
			TUF11	09-Mar-11	33	157	25[c] (78%)

[a] three families removed from sample previously

750 [b] five families removed from sample previously

751 [c] one family not sampled

752

753 **Table 2.** Template specific primer used for amplification of macroinvertebrate samples in

754 this study.

Primer set	Name	Sequence (5' to 3')	Reference	Amplicon size
1	COIBfor	GAYCGWATRCCWYTRTTTGTYTGATC	This study	197 bp
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer *et al*. (1994)	
2	ZBJ-ArtF1c	AGATATTGGAACWTTATATTTTATTTTTGG	Zeale *et al*. (2011)	211 bp
	ZBJ-ArtR2c	WACTAATCAATTWCCAAATCCTCC	Zeale *et al.* (2011)	
3	LCO1490	GGTCAACAAATCATAAAGATATTGG	Folmer *et al*. (1994)	278 bp
	MLepR2	CTTATATTATTTATTCGTGGGAAAGC	Hajibabaei *et al.* (2006a)	

755

Table 3. The number of reads, contigs and sequences matched to taxa in the freshwater macroinvertebrate database for the three amplicons used in this study. The number of 'reads' given includes only those used to construct contigs. The number of 'contigs' includes only contigs of 10 or greater reads after chimera removal. The species identified by % match given the number of MegaBLAST matches to species in the freshwater invertebrate database at >97.5% (a species match) and 90-97.4% (a possible genus/family match) and <89.9% (unmatched contigs). Note that some contigs produced matches to the same species, hence there are often more contigs than 'total' matches.

sample	COIBfor/HCOI amplicon						LCOI/MLepR2 amplicon						Zeale amplicon					
			Species identified % match						Species identified % match						Species identified % match			
	reads	contigs	>97.5	90-97.4	<89.9	Total	reads	contigs	>97.5	90-97.4	<89.9	Total	reads	contigs	>97.5	90-97.4	<89.9	Total
IHC04	50800	41	23	5	4	32	0	0	0	0	0	0	48408	32	13	8	3	24
IHC08	239445	83	22	13	13	48	21329	22	10	1	2	13	50002	34	16	6	0	22
IHC10	310781	65	20	8	6	34	22231	19	11	2	1	14	68256	35	13	3	0	16
IHC12	19271	47	27	13	6	46	9057	23	13	4	1	19	8319	32	19	7	3	29
TUF07	131592	53	20	7	3	30	40493	35	14	8	0	22	60462	44	14	6	0	20
TUF08	180036	100	42	16	13	71	19764	47	23	8	1	32	88406	65	33	10	1	44
TUF09	42424	67	31	9	9	49	92	3	3	0	0	3	38261	51	28	11	0	39
TUF11	139915	93	29	19	15	63	46974	52	26	11	0	37	65522	65	31	13	3	47
TUF11 rep	173394	75	27	16	9	52	50898	54	27	13	0	40	87039	75	30	15	4	49
RL09	85885	86	14	0	0	14	46271	16	8	0	0	8	175924	41	16	0	1	17

Table 4. The detection of taxa present in the chironomid control sample (RL09) showing the species present, their abundance (n), and the number of reads for each amplicons detected with NGS. Additional taxa are not recorded as being present in the sample using individual identification but found using NGS.

Species	n	Amplicons (reads)			Total reads
		COIBfor/HCOI	LCOI/MLepR2	Zeale	
Chironomus australis	1	516	1494	55866	57876
Chironomus cloacalis	1	35		714	749
Chironomus duplex	3	1659		87	1746
Chironomus februarius	20	3384	3748	66877	74009
Chironomus oppositus	22	637	4017		4654
Chironomus pseudoppositus	3	100		28	128
Chironomus tepperi	1				0
Cladopelma sp.2	8	771	125	14609	15505
Dicrotendipes pseudoconjunctus	7	593	560	10674	11827
Dicrotendipes sp.4	1	54	50	845	949
Kiefferulus martini	1	85	57	576	718
Microchironomus forcipatus	1	43		100	143
Procladius villosimanus	30	15605	36220	25050	76875
Tanytarsus inextentus	1	35		218	253
Additional taxa					
Chaetogaster diaphanous (Oligochaeta)				15	15
Karualona (Branchiopoda)				22	22
Podocopida (Ostracoda)				189	189
Paratrichocladius sp.3 (Chironomidae)				54	54

768 **Table 5.** Lowest taxonomic level identifications based on morphological examination of samples IHC04, IHC08 and IHC10 from the
769 Wimmera River sites IHC04, IHC08 and IHC10 compared to species detected with NGS. The presence of DNA barcodes for species and
770 genera are also indicated. BOLD BIN numbers (codes with 3 letters and 4 numerals) are used for some species.

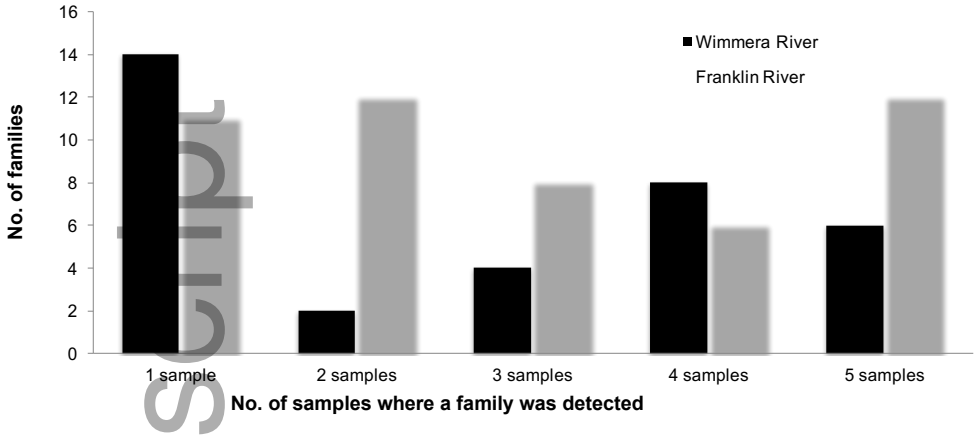
Morphological identification	DNA barcode available (y/n)	IHC04		IHC08		IHC10	
		n	NGS species identification	n	NGS species identification	n	NGS species identification
ACARINA				2		1	Acarina sp. (undet.)
Dugesiidae		2*					
DECOPODA							
Aytidae							
Paratya australiensis	y	4	*P. australiensis*			4	*P. australiensis*
Chiltoniidae							
Austrochiltonia sp.	y	3	*A. subtenuis*	18	*A. subtenuis*	11	*A. subtenuis*
GASTROPODA							
Physidae							
Physa acuta	y	13	*P. acuta*			8	*P. acuta*
OLIGOCHAETA		2	*Chaetogaster diaphanus*	8	*Branchiura sowerbyi, Dero digitata*	1	*Dero* sp.AAL1506
EMPHEMEROPTERA							
Caenidae		3		4	*Tasmanocoenis* sp. (96% match)		
COLEOPTERA							
Dytiscidae						1	
Allodessus sp.	y					2	*Allodessus* sp. ACG1311
Antiporus gilberti	n			1		1	
Antiporus sp.	n			2			
Limbodessus sp.	y					1	
Sternopriscus sp.	n	4		4			
Hydraenidae							
Ochthebius sp.	n			2	Hydraenidae sp. (undet.)		
Hydrophilidae		2				2	
HEMIPTERA							
Corixidae		3					

Micronecta sp.	y	4	*Micronecta* spp.	7	*Micronecta* spp.	21	*Micronecta* spp.
Sigara truncatipala	y	1	*S. truncatipala*				
Notonectidae							
Anisops thienemanni	n			1		1	
Anisops sp.	n	2		3			
Enithares woodwardi	y			1	*E. woodwardi*		
Veliidae							
Microvelia oceanica	n	3	*Gerris*	2	*Gerris*	1	*Gerris*
Microvelia peramoena	n					1	
DIPTERA							
Ceratopogonidae							
Ceratopogonidae EPA sp. 47	n					3	Ceratopogonidae sp. (undet.)
Ceratopogonidae EPA sp.17	n	2		6	Ceratopogonidae (92.4% match)		
Ceratopogonidae EPA sp.28	n	2					
Chironomidae							
Cardiocladius sp.	y			1			
Chironomidae (Pupa)				3	*P. grimmii*, *C.* sp.2		
Chironomus sp.	y	6	*C. australis*			32	*C.* sp. *'februarius'*
Cricotopus sp.	y		*C* sp.2, *C* sp.B			1	*C.* sp.2
Dicrotendipes sp.	y	2	*D. pseudoconjunctus*	2	*D. pseudoconjunctus*	2	*D. pseudoconjunctus*
Harrisius sp.	n			1*			
Kiefferulus martini	y					1	*K. martini*
Kiefferulus sp.	y	1	*K. intertinctus*				
Larsia sp.	y	1	Tanypodinae sp.9				
Parachironomus sp.	y	1*		1*			

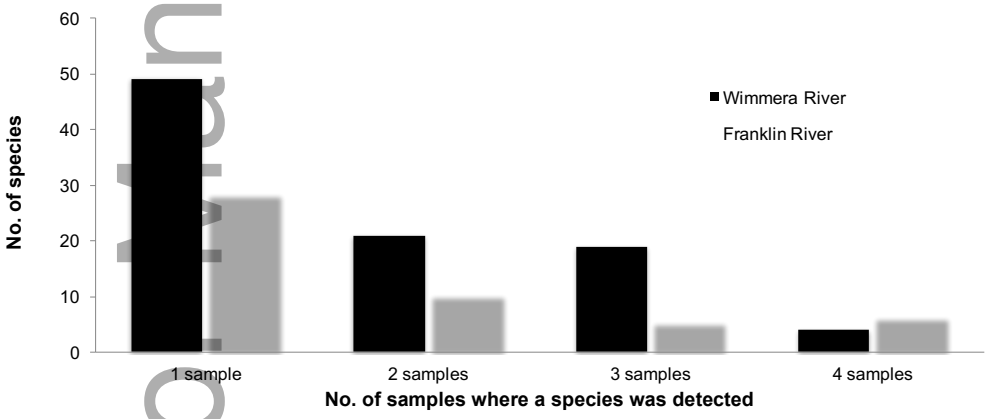
Paratanytarsus sp.	y	15	*P. grimmii*				
Polypedilum sp.	y	3	*P. vespertinus, P.* S1				
Procladius sp.	y	2	*P. paludicola*	3	*P. paludicola*		
Tanytarsus sp.	y	1	*T. inextentus*	2	*T. inextentus*		
ODONATA							
Coenagrionidae		2	*Xanthagrion erythroneurum*	7	*Xanthagrion erythroneurum*	14	*I. aurora,* CoenagrionidGC sp. AAO8233
Ischnura heterosticta	y					4	
Ischnura sp.	y	9	*I. heterosticta*	6	*I. aurora, I. heterosticta*		*I. heterosticta*
Gomphidae							
Austrogomphus sp.	y			1	Gomphidae (94% match)		
Hemicorduliidae							
Hemicordulia tau	y			1	Corduliidae sp.ACA3305	1	Corduliidae sp.ACA3305
TRICHOPTERA							
Leptoceridae							
Oecetis sp.	y			1	Leptoceridae sp. (92% match)		
Triplectides australis	y	2	*T. australis*	1	*T. australis*	11	*T. australis*

771

a)



b)



		SITES								
		IHC04	IHC08	ICH09	IHC12	TUF07	TUF08	TUF09	TUF11	TUF11 rep
	ACARINA				1		1		1	1
BIVALVIA	Sphaeriidae									
DECOPODA	Atyidae	1		1	1	1	1			
AMPHIPODA	Chiltoniidae	1	1	1	1		1		1	1
TURBELLARIA	Dugesiidae									
GASTROPODA	Hydrobiidae									
	Planorbidae							1	1	1
	Lymnaeidae							1		
	Physidae	1		1	1	1	1	1	1	1
	OLIGOCHAETA	1	1	1						
EMPHEMEROPTERA	Baetidae				1		4	2	3	3
	Caenidae		1		1		1	1		
	Leptophlebiidae				1		8	4	4	4
	Oniscigastridae									
COLEOPTERA	Dytiscidae		1	2	1	1	2	1		
	Elmidae									
	Hydraenidae		1		1					
	Hydrochidae									
	Hydrophilidae									
	Gyrinidae						1			
	Scirtidae				1	2	1		1	1
HEMIPTERA	Corixidae	2	1	1	1	1	1	2	1	1
	Gerridae						1			
	Notonectidae		1				1	1	1	1
	Curculionidae									
	Pleidae								1	1
	Veliidae	1	1	1	1	1	1			
DIPTERA	Ceratopogonidae		1	1	1					
	Chironominae					3	8	7	4	4
	Chironomini	7*	1	3	3					
	Tanytarsini	2*	2	1						
	Orthocladiinae	1	2	1	1	3	5	2	3	4
	Tanypodinae	2*	1		1	6	5			
	Dixidae									
	Simuliidae							2	1	1
	Tipulidae									
TRICHOPTERA	Atriplectididae						1		1	1
	Calamoceratidae						1	1	1	1
	Conoesucidae						1		1	1
	Ecnomidae				2		3	2	2	2
	Helicopsychidae							1	1	1
	Hydropsychidae									
	Hydrobiosidae						1	1	1	1
	Hydroptilidae						1	1	1	1
	Leptoceridae	1	2	1	2	4	8	4	4	4
	Philorheithridae							1	2	2
	Polycentropodidae								1	1
	Odontoceridae						1	1	1	1
ODONATA	Coenagrionidae	2	2	1	4			1	1	1
	Gomphidae		1				1	1		
	Hemicorduliidae		1	1						
	Lestidae									
	Synlestidae									
PLECOPTERA	Gripopterygidae						1