MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1365–1377 doi: 10.1111/1755-0998.12539

Lessons from genome skimming of arthropod-preserving ethanol

B. LINARD,*[a1] P. ARRIBAS,*†[b1] C. ANDÚJAR,*† A. CRAMPTON-PLATT*‡ and A. P. VOGLER*†
*Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK, †Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK, ‡Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK

Abstract

Field-collected specimens of invertebrates are regularly killed and preserved in ethanol, prior to DNA extraction from the specimens, while the ethanol fraction is usually discarded. However, DNA may be released from the specimens into the ethanol, which can potentially be exploited to study species diversity in the sample without the need for DNA extraction from tissue. We used shallow shotgun sequencing of the total DNA to characterize the preservative ethanol from two pools of insects (from a freshwater habitat and terrestrial habitat) to evaluate the efficiency of DNA transfer from the specimens to the ethanol. In parallel, the specimens themselves were subjected to bulk DNA extraction and shotgun sequencing, followed by assembly of mitochondrial genomes for 39 of 40 species in the two pools. Shotgun sequencing from the ethanol fraction and read-matching to the mitogenomes detected ~40% of the arthropod species in the ethanol, confirming the transfer of DNA whose quantity was correlated to the biomass of specimens. The comparison of diversity profiles of microbiota in specimen and ethanol samples showed that 'closed association' (internal tissue) bacterial species tend to be more abundant in DNA extracted from the specimens, while 'open association' symbionts were enriched in the preservative fluid. The vomiting reflex of many insects also ensures that gut content is released into the ethanol, which provides easy access to DNA from prey items. Shotgun sequencing of DNA from preservative ethanol provides novel opportunities for characterizing the functional or ecological components of an ecosystem and their trophic interactions.

Keywords: bacterial symbionts, Coleoptera, genome skimming, mitochondrial metagenomics, preservative ethanol

Received 25 February 2016; revision received 6 May 2016; accepted 6 May 2016

Introduction

The exploration of biodiversity using high-throughput sequencing (HTS) opens a path to new questions and novel empirical approaches. Although initially focusing on microbial diversity (Sogin *et al.* 2011), more recent HTS studies have tackled the characterization of complex communities of macroscopic organisms (e.g. Fonseca *et al.* 2010; Ji *et al.* 2013; Andújar *et al.* 2015). The high sensitivity of these methods also permits the study of DNA isolated directly from the environment (eDNA), such as soil (e.g. Andersen *et al.* 2012) and water (e.g. Jerde *et al.* 2011; Thomsen *et al.* 2012), or ingested DNA from the gut of predators (Paula *et al.* 2015) or blood-sucking invertebrates (iDNA) (e.g. Schnell *et al.* 2012). Most studies have used PCR amplification for targeting particular gene regions and taxonomic groups (metabarcoding) and result in a set of sequences used for profiling the species mixture (Ji *et al.* 2013). As an alternative to metabarcoding, the DNA of such mixtures can also be characterized by metagenomic shotgun sequencing, in a procedure commonly referred to as 'genome skimming' (GS) (Straub *et al.* 2012) and its extension to metagenomes ('metagenome skimming', MGS) (Linard *et al.* 2015). Shallow sequencing of the total DNA and subsequent assembly of reads with genome assemblers preferentially extracts the high-copy number fraction of a sample including the mitochondrial genomes (Gillett *et al.* 2014; Andújar *et al.* 2015; Crampton-Platt *et al.* 2015; Tang *et al.* 2015). In addition, MGS can provide useful information about the species' nuclear genomes and

Correspondence: Benjamin Linard, Fax: +44 (0)207 942 5229; E-mail: b.linard@nhm.ac.uk and Paula Arribas, Fax: +44 (0) 207 942 5229; E-mail: p.arribas@nhm.ac.uk
[a]Present address: LIRMM (Laboratoire d'Informatique de Robotique et de Microelectronique de Montpellier) CNRS University of Montpellier Montpellier France
[b]Present address: Island Ecology and Evolution Research Group IPNA-CSIC La Laguna 38206 Spain

[1]Equal contribution.

© 2016 John Wiley & Sons Ltd

concomitant biodiversity such as bacterial symbionts or gut content (e.g. Paula *et al.* 2015; Linard *et al.* 2015).

Assemblages of invertebrates, which may be a primary target of such HTS efforts, are frequently collected into ethanol as preservative in the field until DNA extraction is performed at some later point. Frequently, multiple conspecific or heterospecific individuals and even complete communities are stored together in a single container, under the assumption that cross-contamination is too low to be detectable in the Sanger sequencing of the individual specimens. However, reports of PCR amplification of arthropod genes from ethanol and even from alcoholic beverages indicate that traces of DNA are transferred from the specimen to the preservative (e.g. Shokralla *et al.* 2010; Hajibabaei *et al.* 2012), and with the much greater sensitivity of single-molecule sequencing, the question about the magnitude of cross-contamination takes on a new significance. In addition, detecting low concentration DNAs in the preservative opens exciting new opportunities for the study of bulk biodiversity samples, as extractions directly from the ethanol may avoid the need for tissue preparations and the resulting damage to specimens caused by standard methods. This would be particularly useful for the sequencing of spirit-preserved collections in the world's natural history museums.

In a recent metabarcoding study of benthic arthropods, the set of species obtained directly from the specimen mixture were reported to be detectable also in the ethanol in which these specimens had been stored (Hajibabaei *et al.* 2012). However, these PCR-based studies did not provide a quantitative measure of the amount of transferred DNA. The great sequencing depth achievable with Illumina sequencing now permits a more direct approach to address the question about DNA transfer to the ethanol with PCR-free methods by shotgun sequencing of DNA from the preservative ethanol. This approach could be a straightforward, nondestructive way to study bulk-collected arthropods. In addition, the nontargeted sequencing of total DNA could also be used to explore specific fractions of the associated biodiversity that are released into the preservative, for example from the gut or attached to the exoskeleton, which may be different in composition from the directly sequenced specimen. Therefore, shallow metagenomic sequencing of preservative ethanol could be used as an alternative tool to study species diversity and biotic associations.

Here, we conducted shotgun sequencing on DNA extracted from ethanol used as a killing agent and preservative in field collecting of mixed arthropods (one freshwater pool and one terrestrial pool). We also extracted DNA from the ethanol-preserved specimens and assembled complete mitochondrial genome sequences from shotgun sequencing thereof. These assemblies served as reference sequences to map the reads from the ethanol fraction, as a measure of the magnitude of DNA transfer from the specimens to the preservative medium. In addition, we extensively explored the concomitant biodiversity detectable in the preservative fluid, with special attention to potential gut content released from the live specimens when placed in the ethanol. The collection fluid therefore may be enriched for food items and gut bacteria, but may be impoverished for internal parasites and bacterial endosymbionts if compared with specimen DNA extractions. Considering that field collection of bulk arthropod communities into preservative ethanol remains the primary step in most biodiversity surveys, sequencing of ethanol-derived DNA may be a powerful approach for the study of species diversity and ecology.

Materials and methods

Specimen collection

Two arthropod pools were generated with specimens collected from terrestrial and aquatic environments in Richmond Park, Surrey, UK (coordinates: 51.456083, −0.264840). Aquatic arthropods were collected along the edge of a pond using a 5-mm mesh. Live specimens were transferred to a 100-mL sterile vial containing 80 mL of 100% (pure) ethanol to generate a pooled 'aquatic' sample (Fig. 1A). A 'terrestrial' sample was obtained by hand collection of beetles under stones and logs in the area surrounding the pond. Both were conserved for less than a day at ambient temperature and maintained at −18 °C for two weeks before DNA extraction was performed. The specimens occupied up to half of the volume of the collecting vial, reducing the final concentration of the ethanol to an unknown degree.

Mitochondrial metagenomics of voucher specimens

Specimens from each pool (*vouchers*) were individually removed from the ethanol using sterilized forceps, identified to genus level, grouped by morphospecies, and their body length measured (Fig. 1B). Individual nondestructive DNA extraction was performed on up to four specimens of each morphospecies using the DNeasy Blood & Tissue Spin-Column Kit (Qiagen). The 5′ half of the *cox1* gene (barcode fragment) was PCR-amplified using the *FoldF* and *FoldR* primers (see Appendix S1, Supporting information for details), and the PCR products were Sanger-sequenced with ABI technology. Morphological identifications were validated by BLAST searches against the NCBI and BOLD databases (accessed on 29-04-2015). DNA concentrations of specimen

© 2016 John Wiley & Sons Ltd

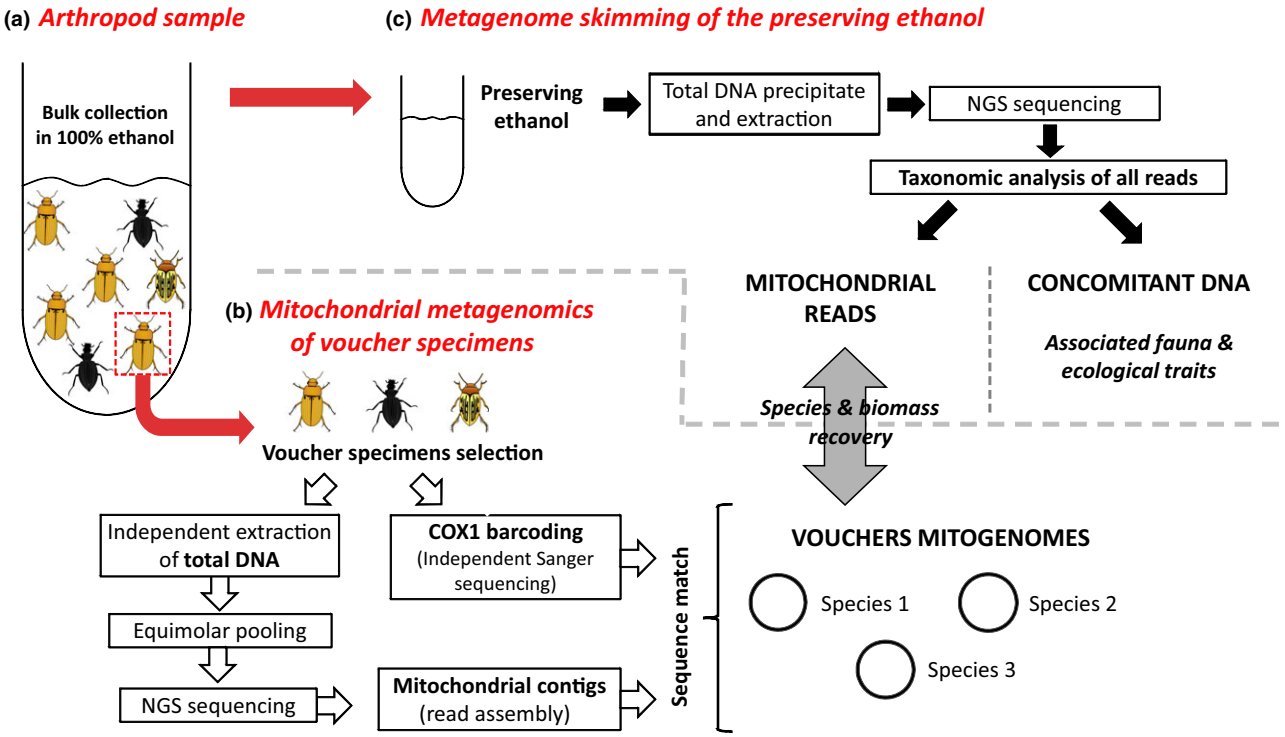

Fig. 1 Schematic representation of the experimental design and bioinformatics pipeline followed in this study.

extractions were estimated using the Qubit dsDNA HS Assay Kit (Invitrogen), and equimolar pooled aliquots were used to prepare two specimen pools: *Terrestrial Vouchers* (TV) and *Aquatic Vouchers* (AV). Two Illumina TruSeq DNA PCR-free libraries were prepared and sequenced on an Illumina MiSeq sequencer (2 × 250 bp paired-end reads).

Raw paired reads were trimmed to remove residual library adaptors with TRIMMOMATIC v0.32 (Bolger *et al.* 2014), and PRINSEQ v0.20.4 (Schmieder & Edwards 2011) was used for filtering low-quality reads. Filtered reads from each pool were then assembled using four different assemblers: CELERA ASSEMBLER v7.0 (Myers 2000), IDBA-UD v1.1.1 (Peng *et al.* 2012), NEWBLER v2.7 (Miller *et al.* 2010) and RAY-META v1.6.5 (Boisvert *et al.* 2012). Contigs with regions of high similarity produced by the different assemblers were merged with the *'De Novo Assembly'* function of GENEIOUS v7.1.8 (minimum overlap = 500 bp; minimum overlap identity = 99%). The resulting mitogenomes were first annotated with the MITOS server (Bernt *et al.* 2013), then manually curated to validate all protein-coding, rRNA and tRNA genes. Finally, mitogenomes were matched with the corresponding Sanger *cox1* sequences for species assignment. For further details on the mitochondrial metagenomics pipeline, see Crampton-Platt *et al.* (2015) and Appendix S1 (Supporting information).

Metagenomics of voucher specimens and preservative ethanol

The preservative ethanol from the terrestrial and aquatic pools was decanted and centrifuged (Fig. 1C) at 14000 *g* for 30 min at 6 °C to allow for sedimentation of precipitated DNA (Tréguier *et al.* 2014). The supernatant was discarded, the precipitate was dried, and DNA was extracted using the DNeasy Blood & Tissue Spin-Column Kit (Qiagen). Concentrations of total DNA extracts were estimated using the Qubit dsDNA HS Assay Kit (Invitrogen), and the two pools representing the terrestrial and aquatic specimens, respectively, in equal concentrations were used to prepare TruSeq DNA PCR-free libraries, referred to as *Terrestrial Ethanol* (TE) and *Aquatic Ethanol* (AE), and Illumina sequenced (2 × 250 bp paired-end reads for AE; 2 × 300 bp paired-end reads for TE) using 5% and 4% of a flow cell on the MiSeq. Adapter removal and quality control followed the same protocol as described above for the *vouchers* (TV and AV; also see Appendix S1, Supporting information).

Voucher species recovery from the preservative ethanol—Species recovery from the preservative ethanol was assessed by matching the filtered TE and AE reads against the voucher sequences using BLAST (≥97% similarity over

© 2016 John Wiley & Sons Ltd

≥150 bp). Sanger sequences, full-length assembled mitogenomes and the protein-coding genes only (i.e. excluding the less variable rRNA genes) were used as references to check for differences in species recovery depending on the voucher information used. The biomass of each species in the pools was estimated using specimen length as a proxy for body size, multiplied by number of specimens, and was subsequently correlated with the number of matching reads from the *ethanol* libraries.

Phylogenetic profile of the vouchers and the preservative ethanol—The diversity of concomitant DNA (reads presumed not to be derived from the genomes of voucher specimens) was estimated for each library (Fig. 1C) by (i) a general taxonomic characterization of the paired reads and (ii) a more precise assignment of the reads to mitochondria, plastids, nuclear rRNAs and putative bacterial symbionts. The general taxonomic characterization is based on a custom database combining the whole content of the preformatted NCBI *nt* (nucleotides) database and all coleopteran assemblies currently available in the NCBI *wgs* database (Appendix S1, Supporting information for the reason motivating this choice). Each library was aligned to this custom database with MEGABLAST from the BLAST+ package (Camacho *et al.* 2009), retaining only hits with a maximum E-value of 1e-15. BLAST outputs were then analysed with MEGAN 5.10.3 (Huson *et al.* 2007). The MEGAN LCA (Lowest Common Ancestor) clustering was set to consider paired reads as belonging to the same entity and only the top 20% of BLAST hits were considered for taxonomic assignments, with all other MEGAN clustering parameters kept at default values. Pie charts describing the taxonomic content of the *voucher* and *ethanol* libraries were also generated with MEGAN.

Assignment of reads to four specific categories of DNA markers was based on read matches to four custom reference databases, including (i) 'Mitochondria' containing all complete and partial mitochondrial genomes (minimum 10 kb) from the NCBI *nt* database (downloaded on 05-05-2015); (ii) 'Plastids' obtained by retrieving all complete and fragmented plastid genomes (minimum 10 kb) from the NCBI Nucleotide database (downloaded on 04-05-2015); (iii) 'Symbionts' based on all complete genomes available from NCBI for a panel of bacterial genera known for their symbiotic interactions in different arthropod lineages, including 27 bacterial genera reported in Russell *et al.* (2012) (retrieved from the NCBI Genome database on 08-07-2014; details in Appendix S1, Supporting information); and (iv) 'Nuclear rRNAs' corresponding to the whole content of the SILVA database (Quast *et al.* 2013) (release 119, containing manually curated 18S and 28S rRNAs for 2 100 000 bacteria, 49 000 archaea, 95 000 eukaryotes and 44 000 unclassified cultured organisms). Reads of all libraries were aligned to these databases with MEGABLAST, and the taxonomic classification of the BLAST best hit was assigned based on stringent similarity thresholds (Appendix S1, Supporting information). Mitochondrial and plastid reads were then grouped according to high taxonomic levels (Arthropods, Plants, Fungi, etc.), while bacterial symbionts and rRNA reads were assigned to genera when more than 99% similar to a reference for >90% of the read. Only taxa supported by more than five matching reads in one of the libraries were considered for further analyses.

The proportion of reads assigned to the above four classes of DNA markers in different taxa were compared between the *vouchers* (AV, TV) and the *ethanol* (AE, TE) libraries. For a single library, a marker proportion is reported as the ratio of base pairs assigned to a particular taxon over the total number of base pairs sequenced in the library. The percentage difference (increase or decrease) of this proportion in the ethanol compared with the voucher libraries was calculated. Formally, in a library L of size S (bp) we define a pair $\{C, M\}$ representing a clade C and a DNA marker M. In L, the number of bp n associated with M and identified as belonging to C is noted $n^L_{\{C,M\}}$ and is then converted to a library proportion $P^L_{\{C,M\}}$ with the formula:

$$P^L_{\{C,M\}} = \frac{n^L_{\{C,M\}}}{S^L}$$

The percentage change (% change) observed for a pair $\{C,M\}$ in a library L_2 compared with a library L_1, as well as the magnitude of change corresponding to this increase (when positive) or decrease (when negative), is then defined as:

$$\%\,\text{change}_{L_2/L_1} = \frac{P^{L_2}_{\{C,M\}} - P^{L_1}_{\{C,M\}}}{P^{L_1}_{\{C,M\}} \times 100}$$

Typically, L_2 will correspond to an *ethanol* library (E) that is compared to L_1 constituting a *voucher* library (V) and a pair of clade and marker could be for instance {Bacterial symbiont, rRNAs}. Then, the differential recovery obtained from the ethanol is reported as the order of magnitude ($\log_{10}$) of the difference $\Delta F_{E/V}$ in nucleotide counts between both libraries, that is

$$\Delta F_{L_2/L_1} = \log_{10}\left(\left|\%\,\text{change}_{L_2/L_1}\right|\right)$$

For instance, for the pair {Bacterial symbiont, rRNAs} a $\Delta F_{E/V} = 2$ indicates a recovery of symbionts

© 2016 John Wiley & Sons Ltd

rRNA base pairs 100 times higher in the *ethanol* (preservative) compared with the *voucher* (the specimen itself).

Results

Assembly of mitogenomes from voucher specimens

A total of 126 and 49 specimens were collected, respectively, in the aquatic and terrestrial habitats, which in total represented 38 morphospecies from the order Coleoptera and one morphospecies each of Trichoptera and Megaloptera encountered as larval stages in the freshwater pool. Representatives of all morphospecies were selected as vouchers, and depending on body size and where possible, up to four specimens were subjected to DNA extractions (to standardize the amount of DNA for improved assembly), for a total of 72 specimens (see Table 1). Sanger sequencing generated successful *cox1* barcodes for 37 of the 40 morphospecies (Table 1). BLAST matches of these voucher *cox1* sequences against the NCBI and BOLD databases showed good agreement with the morphospecies identifications (Table 1). The voucher DNA extracts were pooled in equal concentrations to generate two mixtures, one terrestrial (TV) and one aquatic (AV). Illumina MiSeq sequencing on these pools produced, respectively, 10 782 446 and 26 867 180 paired reads after quality control and resulted in successful assembly of complete or nearly complete mitochondrial genomes for 39 of the 40 morphospecies (Table 1).

Metagenomics of voucher specimens and preservative ethanol

Voucher species recovery from the preservative ethanol—The TE and AE libraries built from the preservative ethanol produced a total of 1 960 740 and 1 772 094 paired reads, respectively. Matching these reads against the voucher *cox1* sequences recovered only four species, while using the full-length and protein-coding genes of the assembled mitogenomes recovered 15 and 13 species. The species with highest recovery were those with high biomass in the samples, including the larval specimens of *Sialis sp.* (Neuroptera) and *Dorcus sp.* (Coleoptera:Lucanidae) (see Table 1), and a strong correlation was found between the log transformed number of reads in the preservative ethanol and the estimated biomass of each species (Pearson $R = 0.88$, P-value = 0.0001; Fig. 2).

Phylogenetic profile of the vouchers and the preservative ethanol—The general taxonomic characterization of the paired reads showed that in all libraries a large proportion of reads has no BLAST hits to our custom reference databases, with 95.3%, 95.5%, 93.0% and 95.2% of reads unmatched in AV, TV, AE and TE, respectively. The inclusion of coleopteran genome assemblies (from NCBI *wgs* data) in the reference database contributed significantly to the MEGAN identification of arthropod nuclear DNA (compared with using NCBI nucleotide reference set alone; see Appendix S2, Supporting information). This was particularly striking for the aquatic pool, for which the number of identified coleopteran reads increased by a factor 4.4 in AV and 14.1 in AE, while this factor was 1.8 and 1.3 in the terrestrial TV and TE pools.

Identified reads showed different profiles in the voucher and ethanol libraries, but also between the two habitats (Fig. 3). In the *voucher* libraries, the great majority of these reads were apparently derived from the target specimens, with 78.6 and 77.4% identified as arthropod reads in AV and TV. This proportion was reduced in the *ethanol* libraries to 17.2 and 7.1% in AE and TE. Other DNAs were present in low proportions in the vouchers but dominant in the preservative ethanol. In both *voucher* libraries, Proteobacteria were the 2nd most dominant clade. In AV, Proteobacteria are followed by Nematoda, Platyhelminthes and Chordata reads in decreasing proportions, with more than half of the Chordata reads identified as sequences of *Cyprinus carpio* (common Eurasian carp). Within Platyhelminthes, 10 158 reads were assigned at the species level to the tapeworm *Hymenolepsis diminuta*. No species-level identifications were obtained for Nematoda, which produced scattered matches to numerous subtaxa. TV showed a similar profile with a dominance of Proteobacteria, followed by a more diverse pattern of various bacterial phyla.

The *ethanol* libraries were characterized by a high diversity of bacterial taxa. Again, Proteobacteria were prevalent but the TE sample clearly differed from all others by showing a large proportion of reads matching Firmicutes (36.5%). In addition, a high diversity of eukaryotic clades was recovered. Ascomycota (fungi) were observed in both habitats with a greater prevalence in TE (6.2%). Chordata and Streptophyta (land plants and green algae) were identified in AE.

Further analyses allowed the assignment of the reads to three main groups, including (i) arthropods, (ii) taxa potentially associated with the gut or the environment, and (iii) bacterial endosymbionts. Their relative proportion was compared in the *voucher* and *ethanol* libraries (Fig. 4, Table S3, Supporting information). Generally, DNA reads were recovered, in decreasing order of abundance, from plastids, mitochondria and rRNA genes in eukaryotes, and from complete genomes and rRNAs in bacterial symbionts, reflecting that longer markers produced more read matches. In agreement with Fig. 3, the proportion of Arthropoda reads in the *ethanol* was much lower than in the *vouchers* for both habitats. On average, a two orders of magnitude ($F = 2.0$) loss was observed

© 2016 John Wiley & Sons Ltd

Table 1 Data set description and voucher species recovery from the preservative ethanol. Ethanol reads correspond to the number of quality filtered reads from the ethanol libraries matching vouchers sequences. Shaded cells show detection of each voucher species from the ethanol libraries

Species	Community	Stage	Total specimens	Specimens used as vouchers	Total estimated biomass	*cox1*_Sanger	Mitogenome	Ethanol reads matching *cox1*	Ethanol reads matching complete mitogenomes	Ethanol reads matching protein-coding mitogenes
Acilius sulcatus BMNH1425211	Aquatic	Adult	2	1	36	X	X	0	0	0
Berosus affinis BMNH1425169	Aquatic	Adult	3	2	13.5	X	X	0	0	0
Colymbetes fuscus BMNH1425212	Aquatic	Adult	5	2	90	X	X	0	15	15
Dryops luridus BMNH1425163	Aquatic	Adult	4	3	20	X	X	0	2	1
Haliplus immaculatus BMNH1425121	Aquatic	Adult	3	2	9	X	X	0	0	0
Haliplus lineatocollis BMNH1425118	Aquatic	Adult	5	3	15	X	X	0	2	0
Helochares sp. BMNH1425100	Aquatic	Adult	10	4	60	X	X	0	0	0
Hydrochus sp. BMNH1425167	Aquatic	Adult	2	2	6	X	X	0	0	0
Hydroporus planus BMNH1425115	Aquatic	Adult	1	2	4.5	X	X	0	0	0
Hydroporus discretus BMNH1425116	Aquatic	Adult	2	2	8	X	X	0	0	0
Hydroporus gyllenhalii BMNH1425127	Aquatic	Adult	2	2	7	X	X	0	2	0
Hydroporus obscurus BMNH1425129	Aquatic	Adult	1	2	3.5	X	X	0	0	0
Hydroporus erythrocephalus BMNH1425131	Aquatic	Adult	27	3	81	X	X	0	2	2
Hydropsyche pellucidulla BMNH1425186	Aquatic	Larva	4	2	56	X	X	2	55	25
Hygrobia hermanni BMNH1425190	Aquatic	Adult	3	1	30	X	X	0	0	0
Hygrotus inaequalis BMNH1425126	Aquatic	Adult	1	1	3	X	X	0	1	1
Hygrotus impressopunctatus BMNH1425158	Aquatic	Adult	5	3	25	X	X	0	0	0
Hygrotus confluens BMNH1425172	Aquatic	Adult	1	1	3.5	X	X	0	0	0
Liopterus haemorrhoidalis BMNH1425193	Aquatic	Adult	6	2	42	X	X	0	0	0
Noterus clavicornis BMNH1425090	Aquatic	Adult	22	3	99	X	X	0	9	5
Sialis lutaria BMNH1425199	Aquatic	Larva	11	2	154	NO	X	24	476	432
Abax parallelepipedus BMNH1425236	Terrestrial	Adult	2	1	40	X	X	0	0	0
Agriotes obscurus BMNH1425233	Terrestrial	Larva	2	1	30	X	X	0	0	0
Anisosticta novemdecimpunctata BMNH1425231	Terrestrial	Adult	1	1	3.5	NO	X	0	0	0
Athous haemorrhoidalis BMNH1425235	Terrestrial	Larva	1	1	9	X	X	0	1	1
Atrecus affinis sp. BMNH1425232	Terrestrial	Adult	1	1	7	X	X	0	2	2
Calathus melanocephalus BMNH1425227	Terrestrial	Adult	1	1	7	NO	X	0	0	0
Cyphon variabilis BMNH1425225	Terrestrial	Adult	2	2	9	X	X	0	0	0
Dorcus parallelipipedus BMNH1425260	Terrestrial	Larva	7	1	175	X	X	17	478	360
Melanotus villosus BMNH1425245	Terrestrial	Larva	8	4	45	X	X	0	6	4
Nalassus laevioctostriatus BMNH1425217	Terrestrial	Adult	5	2	42.5	X	X	0	0	0
Nebria brevicollis BMNH1425256	Terrestrial	Adult	1	1	14	X	X	0	0	0
Ocypus olens BMNH1425259	Terrestrial	Larva	1	1	16	X	X	0	0	0
Pterostichus niger BMNH1425241	Terrestrial	Adult	4	1	84	X	X	0	12	5
Pterostichus madidus BMNH1425238	Terrestrial	Adult	4	2	64	X	X	0	2	2
Stenus clavicornis BMNH1425222	Terrestrial	Adult	3	2	18	X	X	0	0	0
Stenus boops BMNH1425230	Terrestrial	Larva	1	1	5	X	X	0	0	0
Stomis pumicatus BMNH1425229	Terrestrial	Adult	1	1	6.5	X	X	0	0	0
Tasgius sp. BMNH1425251	Terrestrial	Adult	2	1	34	X	NO	7	0	0
Uloma sp. BMNH1425257	Terrestrial	Larva	2	2	26	X	X	0	0	0

© 2016 John Wiley & Sons Ltd

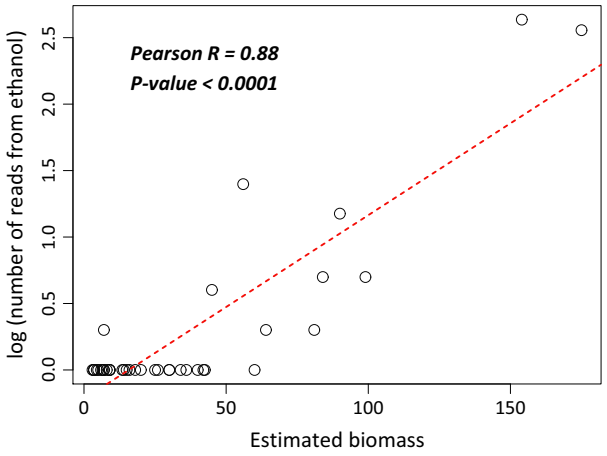

Fig. 2 Relationship between numbers of metagenomic reads from the preservative ethanol for each species and its estimated biomass in the samples.

for both the mitochondrial and the rRNA sequences (Fig. 4A). In contrast, read numbers for some taxa potentially associated with the environment and gut content (Fig. 4B) were increased in the *ethanol* by between 2.2 (Fungi rRNA) to 4.6 (Annelida rRNA) orders of magnitude. Following Douglas (2015), the symbiont species were divided into those with 'closed associations' representing strict bacterial symbionts confined to bacteriocytes or specific host tissues, and those in 'open associations' representing bacterial infections, loose symbiotic interactions or commensals of the gut. All genera in closed associations (*Wolbachia, Rickettsia, Regiella*) showed a lower recovery from the *ethanol* compared with the *vouchers*, and *Wolbachia* and *Rickettsia*, respectively, were absent altogether in TE and AE, despite their strong signal in the *vouchers* (Fig. 4C). On the other hand, symbiont genera with open associations showed more complex patterns, but in general recovery was higher or at least at similar levels in the *ethanol* than in the *vouchers*. Interestingly, in both TV and TE we noticed the presence of rRNA genes from endosymbionts typically associated with Collembola, possibly providing indirect evidence for predation on arthropod microfauna in some of the voucher specimens of the terrestrial pool (Fig. 4C).

Discussion

Species recovery and shotgun metagenomic sequencing from preservative ethanol

Earlier PCR-based studies have demonstrated that specimen DNA can be obtained from the preservative ethanol (e.g. Shokralla *et al.* 2010; Hajibabaei *et al.* 2012), while here we established the power of direct shotgun sequencing, for a broader characterization of the sampled specimens. PCR-based approaches are effective for detection of low DNA concentration templates and thus have been successful for generating fairly complete species inventories from the ethanol fraction (Hajibabaei *et al.* 2012). We show that the number of DNA reads pertaining to the specimens themselves is rather low and, at the selected sequencing depth, less than half of species present in the samples could be identified from the reads, despite the availability of complete reference mitogenomes. If it is the aim of a study to detect all species in the sample, PCR amplification may be the more efficient approach, but with the proviso that the specific primers used in the assay limit the outcome of the detected taxa (only *cox1* was used in previous studies). Alternatively, a combination of primer sets (Hajibabaei *et al.* 2012) can be used but holds the risk of cross-sample contamination, in particular if samples differ greatly in the concentration of DNA. In addition, the PCR approach may not be universally successful. In our attempts to replicate the *cox1* results on the ethanol samples generated here, we experienced a complete failure of amplification despite the use of various primers and PCR protocols (data not shown). The DNA concentration and level of preservation were sufficient for metagenomic libraries, which generally require much more DNA template than the PCR, ruling out issues affecting the quality or quantity of the template for PCR failure. Instead, PCR inhibitors from the environment or the gut may be enriched in the ethanol fraction, which apparently affects the PCR, but less so the library construction and direct sequencing of the DNA.

In addition, the shotgun approach provides a better quantitative measure of the DNA concentrations for each species, as it is not affected by uneven amplification of templates in the mixture. We find that the DNA pool was dominated by two large-bodied species present in multiple individuals (*Dorcus* sp. in TE and *Sialis* sp. in AE) that accounted for >23% of all mitochondrial reads. Both species were encountered in the larval stages, whose soft cuticle may have facilitated the release of DNA into the ethanol. Some species with low biomass (body size x specimen number) or hard cuticle remain in below the detection limit but should be recovered with deeper sequencing of ethanol libraries beyond the ~5% of a MiSeq flow cell used here. Similarly, recovery of low-biomass species could be improved if great differences in DNA concentration are avoided by sorting according to body size or life stage during field collecting.

The availability of reference sequences was a key requirement for the shotgun approach. We generated an almost-complete reference set of mitogenomes following

© 2016 John Wiley & Sons Ltd

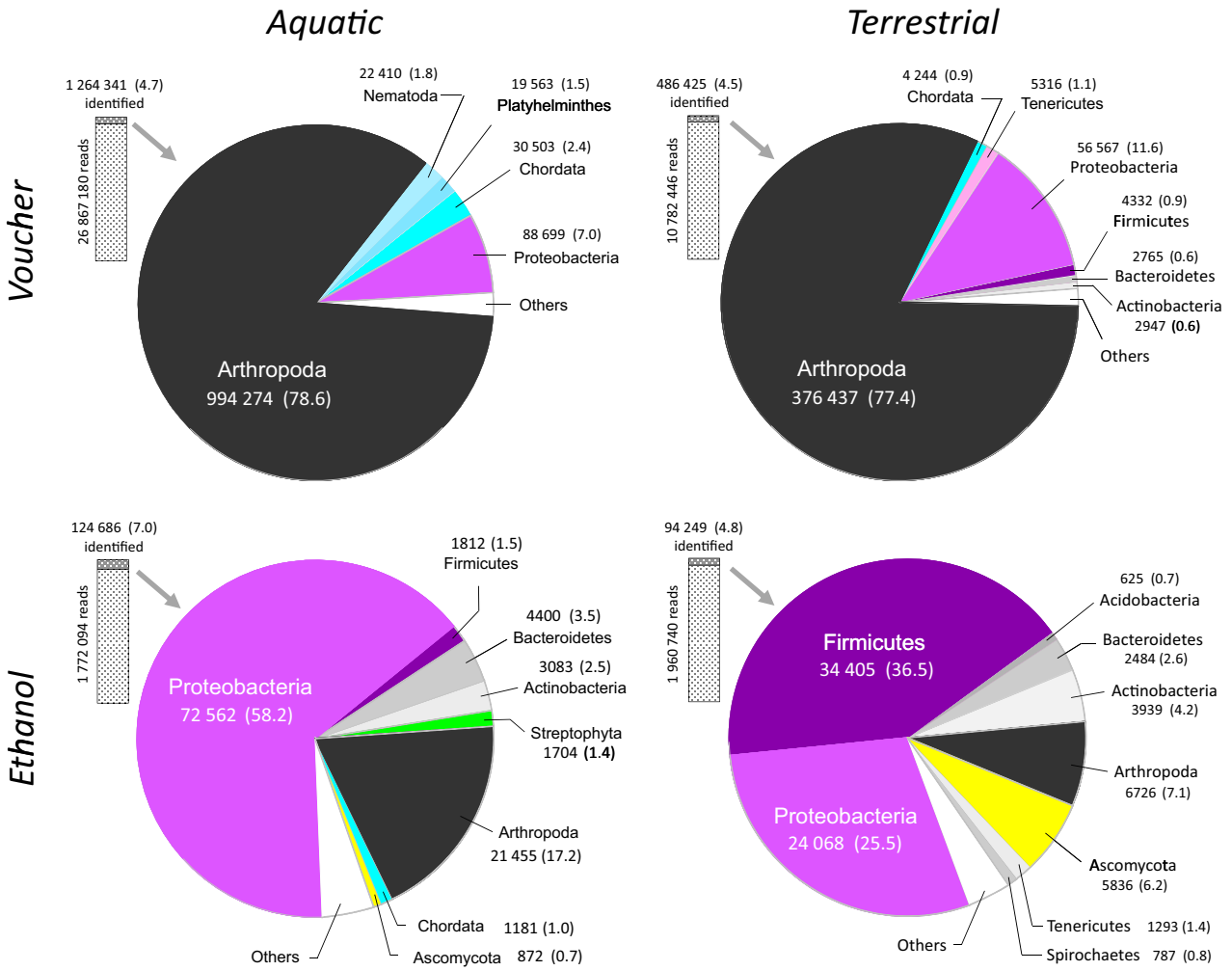

Fig. 3 Taxonomic composition of the identified DNA reads. MEGAN-based identifications are reported for the four libraries. The names of the most abundant taxa are reported while all minor taxa are grouped in the 'other' fraction. The pie charts represent the DNA reads identified as the given taxonomic group and their percentage of the total number of identified reads is given in parentheses. The bars next to each pie chart indicate the number of reads in the library identified to a taxonomic group and their proportion of total reads in parentheses.

an established protocol (Crampton-Platt *et al.* 2015, 2016). At the read depth used here (approximately 1% of a MiSeq flow cell per species), this procedure was highly efficient and even exceeded the species identification rate of *cox1* PCR-based Sanger sequencing of the same specimens. In addition, the *ethanol* libraries produced many matches to arthropod nuclear DNA, including rRNA genes that could be identified against external databases (Fig. 4A). Although complementing mitochondrial references with rRNA markers would greatly increase the sensitivity of species recovery, the assembly of rRNA genes remains challenging. In our tests, no unequivocal contigs were produced in both TV and AV, despite the use of four different assemblers (Table S4, Supporting information). While present in high copy number in metazoan genomes, alternating highly conserved and rapidly evolving expansion segments in the primary sequence of rRNA genes (Stage & Eickbush 2007) currently prevent the assembly from short sequence reads.

Exploration of concomitant biodiversity from the preservative ethanol

The *ethanol* libraries may be considered as complex 'environmental DNA' (eDNA) mixtures that include the DNA released from the focal specimens, together with organisms associated with these specimens and potentially unconnected organisms carried over from the wider ecosystem (Bohmann *et al.* 2014). Bacteria are expected

© 2016 John Wiley & Sons Ltd

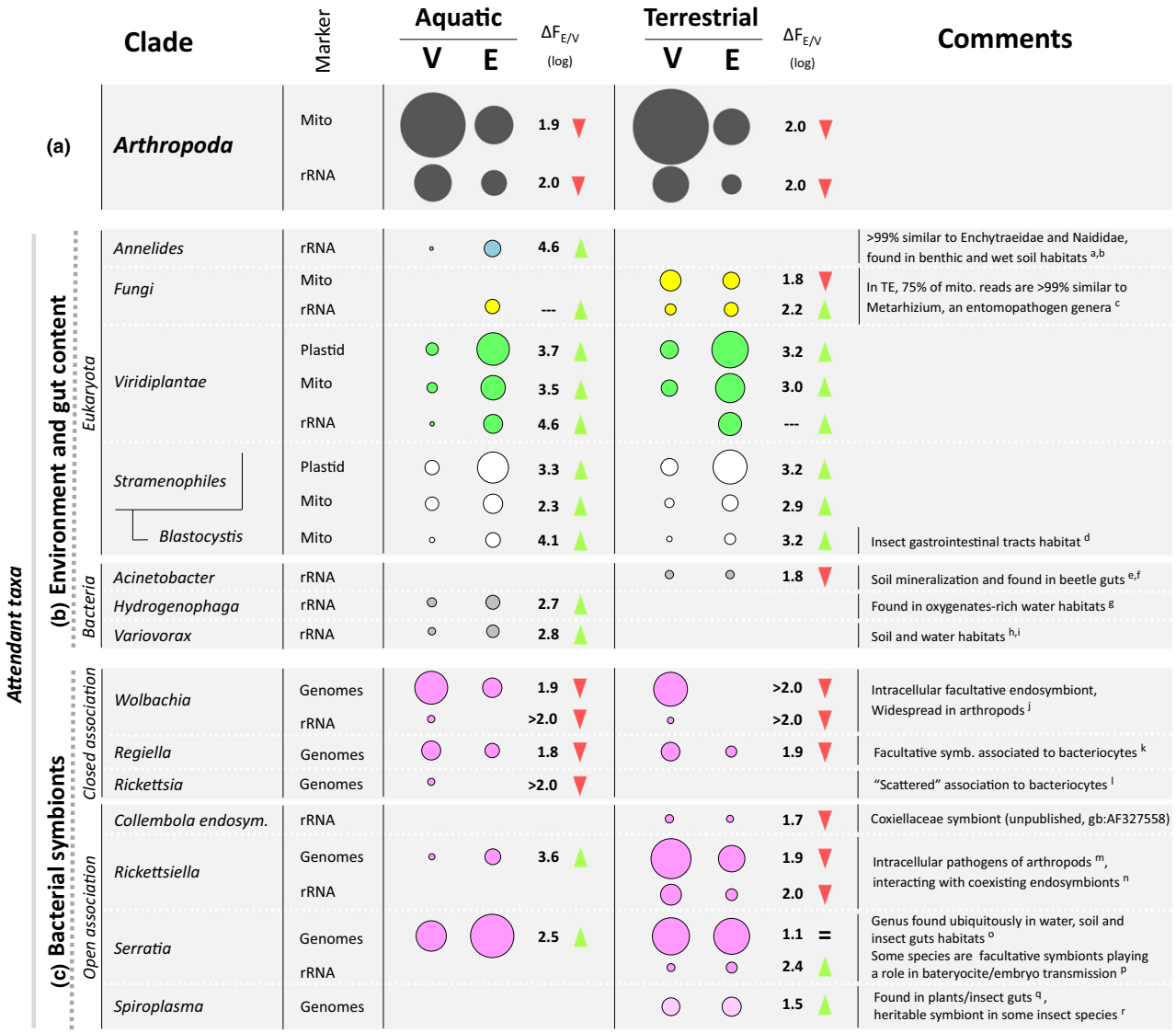

Fig. 4 Ethanol recovery for concomitant DNA. The number of base pairs identified for four types of markers (plastids, mitochondria, rRNAs and symbiont genomic DNA) in different taxa was quantified in the *vouchers* and *ethanol* metagenomes and normalized by library size. Taxa (1st column) are grouped in Arthropoda (A), Environment and Gut (B) and Bacterial symbionts categories (C) based on literature information about the identified taxa ('Comment'). Circle areas represent the square root of the relative proportion of each taxon/marker combination detected in the *vouchers* library (V columns) and the *ethanol* libraries (E columns) in both habitats and their colours are matching taxa in Fig. 3. The increased or reduced recovery in the *ethanol* relative to the *vouchers* libraries is indicated by green or red arrows, and the magnitude of change is given as the $\log_{10}$ of the factor change ($\Delta F_{E/V}$, see Methods). For instance, a $F = 2.0$ lower recovery for a selected taxon/marker indicates that 100 times fewer base pairs were recovered in *ethanol* compared to *vouchers*. References in the last column are as follows: a. Caspers (1986) b. Envall *et al.* (2006) c. Jackson & Jaronski (2009) d. Yoshikawa *et al.* (2007) e. Morales-Jiménez *et al.* (2009) f. Dijkshoorn & Nemec (2008) g. Willems (2014) h. Carbajal-Rodríguez *et al.* (2011) i. Carrino-Kyker & Swanson (2008) j. Sicard *et al.* (2014) k. Moran *et al.* (2005) l. Caspi-Fluger *et al.* (2011) m. Cordaux *et al.* (2007) n. Tsuchida *et al.* (2014) o. Grimont & Grimont (2006) p. Koga *et al.* (2012) q. Gasparich *et al.* (2004) r. Haselkorn *et al.* (2009).

to have a high chance of recovery in the DNA reads, as they are present in high copy numbers and they are detected by read matching against full genomes. Some bacterial genera detected in the ethanol are known to be associated with specific habitats (e.g. *Acinetobacter*, *Hydrogenophaga*; Fig. 4B). These were present in small proportions (Fig. 3), as would be expected in specimens collected manually from the environment, which limits these contaminants. A larger proportion of the ethanol-enriched clades seems to be associated with gut content

© 2016 John Wiley & Sons Ltd

such as *Proteobacteria* or *Firmicutes*, which are generally dominant microbiota of insect guts, followed by *Bacteroidetes*, *Actinobacteria* and *Tenericutes*. The libraries recovered very similar profiles to those obtained in a recent study of insect gut microbiomes (see Figure S2; Yun *et al.* 2014). Bacterial clades known to be gut specific are part of this profile in both habitats, that is high proportions of Enterobacteriales (*Proteobacteria*) and 'open associations' symbionts (*Serratia*, *Rickettsiella*, etc.). Hence, the vomiting of many arthropods at the moment of being immersed in the ethanol (which is seen in many insects but particularly in predatory beetles) appears to be an effective mechanism for the release of gut content to the preservative medium. These DNA profiles from specimen mixtures reflect compound microbiota that are determined by the species composition and relative abundance of the insect communities and their habitat, diet and developmental stage. A case in point are the *Firmicutes* that include the obligatory anaerobic Clostridiales known to be present primarily during larval stages (Yun *et al.* 2014). This group dominated in particular the terrestrial sample with 55% of all reads compared with 34% in the aquatic sample (Table 1, Fig. 3), which is consistent with the higher biomass of larvae in the former.

Other 'closed association' bacterial endosymbionts show the reverse pattern, that is a higher DNA proportion in the vouchers than in the preservative ethanol. These species reside in the bacteriocytes, specialized intracellular compartments that are not expected to be released into the preservative medium. Specifically, *Wolbachia*, *Regiella* and *Rickettsia* are present in most arthropod communities (Werren *et al.* 2008) and in our samples are easily detectable in the voucher libraries but are poorly, if at all, recovered from the ethanol (Fig. 4C). By contrast, several bacterial genera implicated in 'open' symbiotic associations as commensals outside the bacteriocytes (Moran *et al.* 2005) show more mixed patterns. This category of bacteria appears to be the main candidate if one intends to use the preservative ethanol for the study of insect symbiont communities. Finally, some eukaryotic species relevant to insect biology were also detected (Fig. 4). The Viridiplantae and Stramenopiles were greatly enriched in the ethanol (Fig. 4) and may represent ingested food items. Potential infectious agents, such as the entomopathogenic fungus *Metharizium* (Jackson & Jaronski 2009), represented as much as 75% of fungal reads in TE. In contrast, the fungal genus *Hymenolepis* known to have parasitic life cycles using insects as intermediary hosts (Shostak 2014) is strongly detected in AV (10 160 reads identified to genus level) and its absence in AE suggests an association with internal tissues but not the gut content.

The value of the preservative ethanol

The increasing depth of modern sequencing technology is changing the analysis of field-collected preserved samples. Each specimen can be seen as an ecosystem in its own right harbouring microbiota, parasites and ingested food. Deep sequencing therefore shifts the focus of metagenomic studies of bulk specimen samples, which were initially geared towards the analysis of species and phylogenetic diversity of a local insect community (e.g. Andújar *et al.* 2015; Crampton-Platt *et al.* 2015; Gómez-Rodríguez *et al.* 2015; Tang *et al.* 2015), but now can take a holistic view that provides new opportunities for research.

For bulk samples, the interactions cannot be ascribed to any particular species in the mixture, but the information is still highly valuable to characterize the functional or ecological components of an ecosystem *in toto*, for example through the parallel study of macro- and microbiomes of bulk samples. For higher precision, the methodology can be modified to include only members of a single species or possibly individually preserved specimens, allowing comparisons among codistributed species for analyses of resource segregation or the turnover in feeding source for a given species or assemblage among different sites. Additionally, the regurgitation of gut content into the ethanol provides a procedure for noninvasive DNA isolation for identification of food items, and it overcomes the problem that the degraded DNA of the gut content makes up only a small proportion of sequence reads compared with the well-preserved gut tissue that cannot be removed even with careful dissections (e.g. Paula *et al.* 2015). The greatest value of these techniques lies in the possibility for making comparison of numerous samples, each of them surveyed for multiple types of trophic interactions, given a different ecological context in which the target taxa are found. The high cost of shotgun sequencing relative to PCR-based metabarcoding may be a deterrent for such studies, but due to the emergence of cheaper methods for library construction (e.g. Baym *et al.* 2015) and the limited amount of sequencing required (e.g. 5% of MiSeq per sample in the current study), these costs are not prohibitive. Thus, the use of the preservative ethanol extends the metasystematic approach to biodiversity assessment and environmental monitoring, for more effective analysis and management of complex ecosystems (Gibson *et al.* 2014). The biomass dependence of shotgun sequencing is another strength of this approach, to provide abundance estimates for ecological studies, while also recovering rare components without PCR biases. Increased sequencing depth and/or biomass preprocessing of the samples could be useful strategies when

© 2016 John Wiley & Sons Ltd

recovering low-biomass entities is required. At the same time, the extension of reference databases, including complete mitochondrial genomes or nuclear genomes, will also increase the reliability of these approaches, reducing their dependency on the completeness of existing public databases.

Beyond the study of freshly collected samples, the significance of bulk sampling and preservative sequencing may arise from the molecular analysis of historical spirit collections. Museum collections provide enormous resources as a baseline against which modern observations can be compared, helping us to build predictive models in a world increasingly influenced by human activities (Suarez & Tsutsui 2004). A holistic approach to the study of preservative ethanol (specimen + eDNA) should reconsider specimen collection and storage practices. A widespread practice to obtain 'cleaner' samples from field collections is the replacement of the original ethanol fraction, which is usually discarded, but this procedure loses valuable information and efforts should be made to store this initial preservative (as volume can easily be reduced through evaporation). Ethanol should also be carefully considered in the management and maintenance of these collections, such as following protocols based on a 'topping-up' of the ethanol (e.g. Notton 2010) instead of replacement.

Long-term microbiota characterization appears to be a potential outcome from insect spirit collections. The ability to quantify the microbiotas in insect specimen vs. ethanol fractions can establish their relationships with the 'host' specimens, while the coexistence of similar organisms within samples from different ecosystems may uncover the pathogenic or ecological role played by the insect microbiome (Mira *et al*. 2010). Similarly, organisms attached to the surface of specimens, such as pollen in the leg baskets of bees or fungi contained in the mycangia of wood-boring beetles, may be present in the preservative medium. Such molecular information can complement the information associated with collection records making the ethanol metagenome itself a record from which more associations may be identified in the future when more DNA reads will be identified against the growing genome reference set. Further studies on the dynamics of DNA transfer from specimens to ethanol under different conditions and how this DNA degrades through time are needed to uncover the full potential of the preserving ethanol into which specimens are collected. But it appears that preservative ethanol is an unexpected source of molecular knowledge: it will contain both the specimen and concomitant biodiversity and can provide valuable biological information when subjected to shallow metagenomic sequencing.

Acknowledgements

This research was funded by the Leverhulme Trust (grant F/00696/P to APV) and the NHM Biodiversity Initiative. PA was supported by two postdoctoral grants from the Royal Society (Newton International Program, UK) and the Spanish Ministry of Economy and Competitiveness (Juan de la Cierva Formación Program, Spain). ACP was funded by a NHM/UCL joint PhD studentship. CA received additional support of a Synthesys grant (GB-TAF-2966) and a postdoctoral NERC grant (NE/L013134/1). Thanks are due to Richmond Park managers for collection permission and assistance, Alex Aitken, Stephen Russell, Kevin Hopkins and Peter Foster (all NHM) for their technical assistance and Sergio Pérez and Félix Picazo for help on the specimen collection and identification, respectively.

References

Andersen K, Bird KL, Rasmussen M *et al*. (2012) Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Andújar C, Arribas P, Ruzicka F *et al*. (2015) Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*, **24**, 3603–3617.

Baym M, Kryazhimskiy S, Lieberman TD *et al*. (2015) Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*, **10**, 1–15.

Bernt M, Donath A, Jühling F *et al*. (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, **69**, 313–319.

Bohmann K, Evans A, Gilbert MTP *et al*. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, **29**, 358–367.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, **13**, R122.

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

Camacho C, Coulouris G, Avagyan V *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Carbajal-Rodríguez I, Stöveken N, Satola B, Wübbeler JH, Steinbüchel A (2011) Aerobic degradation of mercaptosuccinate by the gram-negative bacterium variovorax paradoxus strain B4. *Journal of Bacteriology*, **193**, 527–539.

Carrino-Kyker SR, Swanson AK (2008) Temporal and spatial patterns of eukaryotic and bacterial communities found in vernal pools. *Applied and Environmental Microbiology*, **74**, 2554–2557.

Caspers H (1986) Aquatic Oligochaeta. Proceedings of the second international symposium on aquatic oligochaete biology, held in Pallanza, Italy, September 1982. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, **71**, 583–583.

Caspi-Fluger A, Inbar M, Mozes-Daube N *et al*. (2011) Rickettsia "in" and "out": two different localization patterns of a bacterial symbiont in the same insect species. *PLoS ONE*, **6**, e21096.

Cordaux R, Paces-Fessy M, Raimond M *et al*. (2007) Molecular characterization and evolution of arthropod-pathogenic Rickettsiella bacteria. *Applied and Environmental Microbiology*, **73**, 5045–5047.

Crampton-Platt A, Timmermans MJTN, Gimmel ML *et al*. (2015) Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, **32**, 2302–2316.

Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, **5**, 15.

© 2016 John Wiley & Sons Ltd

Dijkshoorn L, Nemec A (2008) The diversity of the genus Acinetobacter. In: *Acinetobacter Molecular Microbiology* (Eds. Ulrike Gerischer), pp. 1–34. Caister Academic Press, Norfolk, UK.

Douglas AE (2015) Multiorganismal insects: diversity and function of resident microorganisms. *Annual Review of Entomology*, **60**, 17–34.

Envall I, Källersjö M, Erséus C (2006) Molecular evidence for the non-monophyletic status of Naidinae (Annelida, Clitellata, Tubificidae). *Molecular Phylogenetics and Evolution*, **40**, 570–584.

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

Gasparich GE, Whitcomb RF, Dodge D *et al.* (2004) The genus *Spiroplasma* and its non-helical descendants: phylogenetic classification, correlation with phenotype and roots of the *Mycoplasma mycoides* clade. *International Journal of Systematic and Evolutionary Microbiology*, **54**, 893–918.

Gibson J, Shokralla S, Porter TM *et al.* (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8007–8012.

Gillett CPDT, Crampton-Platt A, Timmermans MJTN *et al.* (2014) Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, Baselga A, Vogler AP (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, **6**, 883–894.

Grimont F, Grimont PD (2006) The genus Serratia. In: *The Prokaryotes SE - 11* (eds Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E), pp. 219–244. Springer, New York.

Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, **12**, 28.

Haselkorn TS, Markow TA, Moran NA (2009) Multiple introductions of the *Spiroplasma* bacterial endosymbiont into *Drosophila*. *Molecular Ecology*, **18**, 1294–1305.

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

Jackson MA, Jaronski ST (2009) Production of microsclerotia of the fungal entomopathogen *Metarhizium anisopliae* and their potential for use as a biocontrol agent for soil-inhabiting insects. *Mycological Research*, **113**, 842–850.

Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, **4**, 150–157.

Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Koga R, Meng X-Y, Tsuchida T, Fukatsu T (2012) Cellular mechanism for selective vertical transmission of an obligate insect symbiont at the bacteriocyte-embryo interface. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E1230–E1237.

Linard B, Crampton-Platt A, Timmermans MJTN, Vogler AP (2015) Metagenome skimming of insect specimen pools: potential for comparative genomics. *Genome Biology and Evolution*, **7**, 1474–1489.

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.

Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F (2010) The bacterial pan-genome:a new paradigm in microbiology. *International Microbiology*, **13**, 45–57.

Morales-Jiménez J, Zúñiga G, Villa-Tanaca L, Hernández-Rodríguez C (2009) Bacterial community and nitrogen fixation in the red turpentine beetle, *Dendroctonus valens* LeConte (Coleoptera: Curculionidae: Scolytinae). *Microbial Ecology*, **58**, 879–891.

Moran NA, Russell JA, Koga R, Fukatsu T (2005) Evolutionary relationships of three new species of Enterobacteriaceae living as symbionts of aphids and other insects. *Applied and Environmental Microbiology*, **71**, 3302–3310.

Myers EW (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Notton DG (2010) Maintaining concentration: a new practical method for profiling and topping up alcohol-preserved collections. *Collection Forum*, **24**, 1–27.

Paula DP, Linard B, Andow Da *et al.* (2015) Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. *Molecular Ecology Resources*, **15**, 880–892.

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, **28**, 1420–1428.

Quast C, Pruesse E, Yilmaz P *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.

Russell JA, Funaro CF, Giraldo YM *et al.* (2012) A veritable menagerie of heritable bacteria from ants, butterflies, and beyond: broad molecular surveys and a systematic review. *PLoS ONE*, **7**, e5102.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schnell IB, Thomsen PF, Wilkinson N *et al.* (2012) Screening mammal biodiversity using DNA from leeches. *Current Biology: CB*, **22**, R262–R263.

Shokralla S, Singer GA, Hajibabaei M (2010) Direct PCR amplification and sequencing of specimens' DNA from preservative ethanol. *BioTechniques*, **48**, 233–234.

Shostak AW (2014) *Hymenolepis diminuta* infections in tenebrionid beetles as a model system for ecological interactions between helminth parasites and terrestrial intermediate hosts: a review and meta-analysis. *The Journal of Parasitology*, **100**, 46–58.

Sicard M, Dittmer J, Grève P, Bouchon D, Braquart-Varnier C (2014) A host as an ecosystem: Wolbachia coping with environmental constraints. *Environmental Microbiology*, **16**, 3583–3607.

Sogin ML, Morrison HG, Huber JA *et al.* (2016) Microbial Diversity in the Deep Sea and the Underexplored "Rare Biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–12120.

Stage DE, Eickbush TH (2007) Sequence variation within the rRNA gene loci of 12 *Drosophila* species. *Genome Research*, **17**, 1888–1897.

Straub SCK, Parks M, Weitemier K *et al.* (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

Suarez AV, Tsutsui ND (2004) The value of museum collections for research and society. *BioScience*, **54**, 66.

Tang M, Hardman CJ, Ji Y *et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics (M Gilbert, Ed,). *Methods in Ecology and Evolution*, **6**, 1034–1043.

Thomsen PF, Kielgast J, Iversen LL *et al.* (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Tréguier A, Paillisson J-M, Dejean T *et al.* (2014) Environmental DNA surveillance for invertebrate species: advantages and technical limitations to detect invasive crayfish *Procambarus clarkii* in freshwater ponds (E Crispo, Ed,). *Journal of Applied Ecology*, **51**, 871–879.

Tsuchida T, Koga R, Fujiwara A, Fukatsu T (2014) Phenotypic effect of "Candidatus Rickettsiella viridis", a facultative symbiont of the pea aphid (*Acyrthosiphon pisum*), and its interaction with a coexisting symbiont. *Applied and Environmental Microbiology*, **80**, 525–533.

Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology. *Nature reviews. Microbiology*, **6**, 741–751.

Willems A (2014) The family Comamonadaceae. In: *The Prokaryotes SE - 238*(eds Rosenberg E, DeLong E, Lory S, Stackebrandt E, Thompson F), pp. 777–851. Springer, Berlin Heidelberg.

Yoshikawa H, Wu Z, Howe J *et al.* (2007) Ultrastructural and phylogenetic studies on Blastocystis isolates from cockroaches. *The Journal of Eukaryotic Microbiology*, **54**, 33–37.

© 2016 John Wiley & Sons Ltd

Yun J-H, Roh SW, Whon TW *et al*. (2014) Insect gut bacterial diversity determined by environmental habitat, diet, developmental stage, and phylogeny of host. *Applied and Environmental Microbiology*, **80**, 5254–5264.

B.L., P.A. and C.A. conceived the study; B.L., P.A., C.A. and A.C.P. collected the specimen; P.A. obtained the molecular data; B.L., P.A., C.A. and A.C.P. analysed the data; B.L., P.A., A.P.V. wrote the manuscript, and all the authors contributed to the final version.

Data accessibility

GenBank Accessions nos. for voucher specimens are KT876876–KT876902; KT876904–KT876915; original data sets have been uploaded as fastq files in Dryad doi:10.5061/dryad.jr6r5; all supplementary details, tables and figures cited in the main text have been uploaded as online Supporting Information.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Mitochondrial metagenomics of voucher specimens.

Appendix S2 Global taxonomic compostion of the four metagenomes.

Table S3 Detection of different DNA markers (mitochondria, rRNAs, plastids, bacterial symbiont genomes).

Table S4 rRNA contigs assembled in the libraries.

© 2016 John Wiley & Sons Ltd