



**HAL**  
open science

## Lessons from genome skimming of arthropod-preserving ethanol

Benjamin Linard, P. Arribas, C. Andújar, A. Crampton-Platt, A. P. Vogler

► **To cite this version:**

Benjamin Linard, P. Arribas, C. Andújar, A. Crampton-Platt, A. P. Vogler. Lessons from genome skimming of arthropod-preserving ethanol. *Molecular Ecology Resources*, 2016, 16 (6), pp.1365-1377. 10.1111/1755-0998.12539 . hal-01636888

**HAL Id: hal-01636888**

**<https://hal.science/hal-01636888>**

Submitted on 17 Jan 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



24 **Running title**

25 Metagenome skimming of preservative ethanol

26

27 **Abstract**

28 Field-collected specimens of invertebrates are regularly killed and preserved in ethanol, prior  
29 to DNA extraction from the specimens, while the ethanol fraction is usually discarded.  
30 However, DNA may be released from the specimens into the ethanol, which can potentially  
31 be exploited to study species diversity in the sample without the need for DNA extraction  
32 from tissue. We used shallow shotgun sequencing of the total DNA to characterize the  
33 preservative ethanol from two pools of insects (from a freshwater and terrestrial habitat) to  
34 evaluate the efficiency of DNA transfer from the specimens to the ethanol. In parallel, the  
35 specimens themselves were subjected to bulk DNA extraction and shotgun sequencing,  
36 followed by assembly of mitochondrial genomes for 39 of 40 species in the two pools.  
37 Shotgun sequencing from the ethanol fraction and read-matching to the mitogenomes detected  
38 ~40% of the arthropod species in the ethanol, confirming the transfer of DNA whose quantity  
39 was correlated to the biomass of specimens. The comparison of diversity profiles of  
40 microbiota in specimen and ethanol samples showed that ‘closed association’ (internal tissue)  
41 bacterial species tend to be more abundant in DNA extracted from the specimens, while ‘open  
42 association’ symbionts were enriched in the preservative fluid. The vomiting reflex of many  
43 insects also ensures that gut content is released into the ethanol, which provides easy access to  
44 DNA from prey items. Shotgun sequencing of DNA from preservative ethanol provides novel  
45 opportunities for characterising the functional or ecological components of an ecosystem and  
46 their trophic interactions.

47

## 48 **Introduction**

49 The exploration of biodiversity using high-throughput sequencing (HTS) opens a path to new  
50 questions and novel empirical approaches. Although initially focusing on microbial diversity  
51 (Sogin *et al.* 2011), more recent HTS studies have tackled the characterisation of complex  
52 communities of macroscopic organisms (e.g. Fonseca *et al.* 2010; Ji *et al.* 2013; Andújar *et al.*  
53 2015). The high sensitivity of these methods also permits the study of DNA isolated directly  
54 from the environment (eDNA), such as soil (e.g. Andersen *et al.* 2012) and water (e.g. Jerde  
55 *et al.* 2011; Thomsen *et al.* 2012), or ingested DNA from the gut of predators (Paula *et al.*  
56 2014) or blood-sucking invertebrates (iDNA) (e.g. Schnell *et al.* 2012). Most studies have  
57 used PCR amplification for targeting particular gene regions and taxonomic groups  
58 (metabarcoding), and result in a set of sequences used for profiling the species mixture (Ji *et*  
59 *al.* 2013). As an alternative to metabarcoding, the DNA of such mixtures can also be  
60 characterised by metagenomic shotgun sequencing, in a procedure commonly referred to as  
61 ‘genome skimming’ (GS) (Straub *et al.* 2012) and its extension to metagenomes  
62 (‘metagenome skimming’, MGS) (Linard *et al.* 2015). Shallow sequencing of the total DNA  
63 and subsequent assembly of reads with genome assemblers preferentially extracts the high-  
64 copy number fraction of a sample including the mitochondrial genomes (Gillett *et al.* 2014;  
65 Andújar *et al.* 2015; Crampton-Platt *et al.* 2015; Tang *et al.* 2015). In addition, MGS can  
66 provide useful information about the species’ nuclear genomes and concomitant biodiversity  
67 such as bacterial symbionts or gut content (e.g. Paula *et al.* 2014; Linard *et al.* 2015).

68 Assemblages of invertebrates, which may be a primary target of such HTS efforts, are  
69 frequently collected into ethanol as preservative in the field until DNA extraction is  
70 performed at some later point. Frequently, multiple conspecific or heterospecific individuals

71 and even complete communities are stored together in a single container, under the  
72 assumption that cross-contamination is too low to be detectable in the Sanger sequencing of  
73 the individual specimens. However, reports of PCR amplification of arthropod genes from  
74 ethanol and even from alcoholic beverages indicate that traces of DNA are transferred from  
75 the specimen to the preservative (e.g. Shokralla *et al.* 2010; Hajibabaei *et al.* 2012), and with  
76 the much greater sensitivity of single-molecule sequencing, the question about the magnitude  
77 of cross-contamination takes on a new significance. In addition, detecting low concentration  
78 DNAs in the preservative opens exciting new opportunities for the study of bulk biodiversity  
79 samples, as extractions directly from the ethanol may avoid the need for tissue preparations  
80 and the resulting damage to specimens caused by standard methods. This would be  
81 particularly useful for the sequencing of spirit-preserved collections in the world's natural  
82 history museums.

83 In a recent metabarcoding study of benthic arthropods, the set of species obtained directly  
84 from the specimen mixture were reported to be detectable also in the ethanol in which these  
85 specimens had been stored (Hajibabaei *et al.*, 2012). However, these PCR-based studies did  
86 not provide a quantitative measure of the amount of transferred DNA. The great sequencing  
87 depth achievable with Illumina sequencing now permits a more direct approach to address the  
88 question about DNA transfer to the ethanol with PCR-free methods by shotgun sequencing of  
89 DNA from the preservative ethanol. This approach could be a straightforward, non-  
90 destructive way to study bulk-collected arthropods. In addition, the non-targeted sequencing  
91 of total DNA could also be used to explore specific fractions of the associated biodiversity  
92 that are released into the preservative, e.g. from the gut or attached to the exoskeleton, which  
93 may be different in composition from the directly sequenced specimen. Therefore, shallow

94 metagenomic sequencing of preservative ethanol could be used as an alternative tool to study  
95 species diversity and biotic associations.

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

110

## 111 **Materials and Methods**

### 112 **Specimen collection**

113 Two arthropod pools were generated with specimens collected from terrestrial and aquatic  
114 environments in Richmond Park, Surrey, UK (coordinates: 51.456083, -0.264840). Aquatic  
115 arthropods were collected along the edge of a pond using a 5 mm mesh. Live specimens were

116 transferred to a 100 ml sterile vial containing 80 ml of 100% (pure) ethanol to generate a  
117 pooled 'aquatic' sample (Figure 1A). A 'terrestrial' sample was obtained by hand collection of  
118 beetles under stones and logs in the area surrounding the pond. Both were conserved for less  
119 than a day at ambient temperature and maintained at -18°C for two weeks before DNA  
120 extraction was performed. The specimens occupied up to half of the volume of the collecting  
121 vial, reducing the final concentration of the ethanol to an unknown degree.

### 122 **Mitochondrial metagenomics of voucher specimens**

123 Specimens from each pool (*vouchers*) were individually removed from the ethanol using  
124 sterilised forceps, identified to genus level, grouped by morphospecies, and their body length  
125 measured (Figure 1B). Individual non-destructive DNA extraction was performed on up to  
126 four specimens of each morphospecies using the DNeasy Blood & Tissue Spin-Column Kit  
127 (Qiagen). The 5' half of the *cox1* gene (barcode fragment) was PCR amplified using the *FoldF*  
128 and *FoldR* primers (see Suppl. File S1 for details) and the PCR products were Sanger  
129 sequenced with ABI technology. Morphological identifications were validated by BLAST  
130 searches against the NCBI and BOLD databases (accessed on 29-04-2015). DNA  
131 concentrations of specimen extractions were estimated using the Qubit dsDNA HS Assay Kit  
132 (Invitrogen) and equimolar pooled aliquots were used to prepare two specimen pools:  
133 *Terrestrial Vouchers* (TV) and *Aquatic Vouchers* (AV). Two Illumina TruSeq DNA PCR-free  
134 libraries were prepared and sequenced on an Illumina MiSeq sequencer (2 x 250 bp paired-  
135 end reads).

136 Raw paired reads were trimmed to remove residual library adaptors with Trimmomatic v0.32  
137 (Bolger *et al.* 2014), and Prinseq v0.20.4 (Schmieder & Edwards 2011) was used for filtering  
138 low-quality reads. Filtered reads from each pool were then assembled using four different

139 assemblers; Celera Assembler v7.0 (Myers 2000), IDBA-UD v1.1.1 (Peng *et al.* 2012),  
140 Newbler v2.7 (Miller *et al.* 2010) and Ray-meta v1.6.5 (Boisvert *et al.* 2012). Contigs with  
141 regions of high similarity produced by the different assemblers were merged with the '*De*  
142 *Novo Assembly*' function of Geneious v7.1.8 (minimum overlap = 500 bp; minimum overlap  
143 identity = 99%). The resulting mitogenomes were first annotated with the MITOS server  
144 (Bernt *et al.* 2013), then manually curated to validate all protein-coding, rRNA and tRNA  
145 genes. Finally, mitogenomes were matched with the corresponding Sanger *coxI* sequences for  
146 species assignment. For further details on the mitochondrial metagenomics pipeline see  
147 Crampton-Platt *et al.* (2015) and Suppl. File S1.

#### 148 **Metagenomics of voucher specimens and preservative ethanol**

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

160 *Voucher species recovery from the preservative ethanol*



161 Species recovery from the preservative ethanol was assessed by matching the filtered TE and  
162 AE reads against the voucher sequences using BLAST ( $\geq 97\%$  similarity over  $\geq 150$  bp).  
163 Sanger sequences, full-length assembled mitogenomes, and the protein-coding genes only (i.e.  
164 excluding the less variable rRNA genes) were used as references to check for differences in  
165 species recovery depending on the voucher information used. The biomass of each species in  
166 the pools was estimated using specimen length as a proxy for body size, multiplied by number  
167 of specimens, and was subsequently correlated with the number of matching reads from the  
168 *ethanol* libraries.

#### 169 *Phylogenetic profile of the vouchers and the preservative ethanol*

170 The diversity of concomitant DNA (reads presumed not to be derived from the genomes of  
171 voucher specimens) was estimated for each library (Figure 1C) by (i) a general taxonomic  
172 characterisation of the paired reads and (ii) a more precise assignment of the reads to  
173 mitochondria, plastids, nuclear rRNAs and putative bacterial symbionts. The general  
174 taxonomic characterisation is based on a custom database combining the whole content of the  
175 preformatted NCBI *nt* (nucleotides) database and all coleopteran assemblies currently  
176 available in the NCBI *wgs* database (Suppl. File S1 for the reason motivating this choice).  
177 Each library was aligned to this custom database with megaBLAST from the BLAST+  
178 package (Camacho *et al.* 2009), retaining only hits with a maximum E-value of  $1e-15$ .  
179 BLAST outputs were then analysed with MEGAN 5.10.3 (Huson *et al.* 2007). The MEGAN  
180 LCA (Lowest Common Ancestor) clustering was set to consider paired reads as belonging to  
181 the same entity and only the top 20% of BLAST hits were considered for taxonomic  
182 assignments, with all other MEGAN clustering parameters kept at default values. Pie charts

183 describing the taxonomic content of the *voucher* and *ethanol* libraries were also generated  
184 with MEGAN.

185 Assignment of reads to four specific categories of DNA markers was based on read matches  
186 to four custom reference databases, including (i) “Mitochondria” containing all complete and  
187 partial mitochondrial genomes (minimum 10 kb) from the NCBI *nt* database (downloaded on  
188 05-05-2015); (ii) “Plastids” obtained by retrieving all complete and fragmented plastid  
189 genomes (minimum 10 kb) from the NCBI Nucleotide database (downloaded on 04-05-2015);  
190 (iii) “Symbionts” based on all complete genomes available from NCBI for a panel of bacterial  
191 genera known for their symbiotic interactions in different arthropod lineages, including 27  
192 bacterial genera reported in Russel et al. (2012) (retrieved from the NCBI Genome database  
193 on 08-07-2014; details in Suppl. File S1); (iv) “Nuclear rRNAs” corresponding to the whole  
194 content of the SILVA database (Quast *et al.* 2013) (release 119, containing manually curated  
195 18S and 28S rRNAs for 2,100,000 bacteria, 49,000 archaea, 95,000 eukaryotes and 44,000  
196 unclassified cultured organisms). Reads of all libraries were aligned to these databases with  
197 megaBLAST and the taxonomic classification of the BLAST best hit was assigned based on  
198 stringent similarity thresholds (Suppl. File S1). Mitochondrial and plastid reads were then  
199 grouped according to high taxonomic levels (Arthropods, Plants, Fungi, etc.), while bacterial  
200 symbionts and rRNA reads were assigned to genera when more than 99% similar to a  
201 reference for >90% of the read. Only taxa supported by more than 5 matching reads in one of  
202 the libraries were considered for further analyses.

203 The proportion of reads assigned to the above four classes of DNA markers in different taxa  
204 were compared between the *vouchers* (AV, TV) and the *ethanol* (AE, TE) libraries. For a  
205 single library, a marker proportion is reported as the ratio of base pairs assigned to a particular

206 taxon over the total number of base pairs sequenced in the library. The percentage difference  
 207 (increase or decrease) of this proportion in the ethanol compared to the voucher libraries was  
 208 calculated. Formally, in a library  $L$  of size  $S$  (bp) we define a pair  $\{C, M\}$  representing a clade  
 209  $C$  and a DNA marker  $M$ . In  $L$ , the number of bp  $n$  associated to  $M$  and identified as belonging  
 210 to  $C$  is noted  $n_{\{C,M\}}^L$  and is then converted to a library proportion  $P_{\{C,M\}}^L$  with the formula:

$$P_{\{C,M\}}^L = \frac{n_{\{C,M\}}^L}{S^L}$$

211  
 212 The percentage change (% change) observed for a pair  $\{C,M\}$  in a library  $L_2$  compared to a  
 213 library  $L_1$ , as well as the magnitude of change corresponding to this increase (when positive)  
 214 or decrease (when negative) is then defined as:

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

215  
 216 Typically,  $L_2$  will correspond to an *ethanol* library (E) that is compared to  $L_1$  constituting a  
 217 *voucher* library (V) and a pair of clade and marker could be for instance {Bacterial symbiont,  
 218 rRNAs}. Then, the differential recovery obtained from the ethanol is reported as the order of  
 219 magnitude ( $\log_{10}$ ) of the difference  $\Delta F_{EV}$  in nucleotide counts between both libraries, i.e.

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

220  
 221 For instance, for the pair {Bacterial symbiont, rRNAs} a  $\Delta F_{EV} = 2$  indicates a recovery of  
 222 symbionts rRNA base pairs 100 times higher in the *ethanol* (preservative) compared to the  
 223 *voucher* (the specimen itself).

224

## 225 **Results**

### 226 **Assembly of mitogenomes from voucher specimens**

227 A total of 126 and 49 specimens were collected respectively in the aquatic and terrestrial  
228 habitats, which in total represented 38 morphospecies from the order Coleoptera and one  
229 morphospecies each of Trichoptera and Megaloptera encountered as larval stages in the  
230 freshwater pool. Representatives of all morphospecies were selected as vouchers, and  
231 depending on body size and where possible, up to four specimens were subjected to DNA  
232 extractions (to standardize the amount of DNA for improved assembly), for a total of 72  
233 specimens (see Table 1). Sanger sequencing generated successful *coxI* barcodes for 37 of the  
234 40 morphospecies (Table 1). BLAST matches of these voucher *coxI* sequences against the  
235 NCBI and BOLD databases showed good agreement with the morphospecies identifications  
236 (Table 1). The voucher DNA extracts were pooled in equal concentrations to generate two  
237 mixtures, one terrestrial (TV) and one aquatic (AV). Illumina MiSeq sequencing on these  
238 pools produced, respectively, 10,782,446 and 26,867,180 paired reads after quality control  
239 and resulted in successful assembly of complete or nearly complete mitochondrial genomes  
240 for 39 of the 40 morphospecies (Table 1).

### 241 **Metagenomics of voucher specimens and preservative ethanol**

#### 242 *Voucher species recovery from the preservative ethanol*

243 The TE and AE libraries built from the preservative ethanol produced a total of 1,960,740 and  
244 1,772,094 paired reads, respectively. Matching these reads against the voucher *coxI*  
245 sequences recovered only 4 species, while using the full-length and protein-coding genes of  
246 the assembled mitogenomes recovered 15 and 13 species. The species with highest recovery

247 were those with high biomass in the samples, including the larval specimens of *Sialis sp.*  
248 (Neuroptera) and *Dorcus sp.* (Coleoptera:Lucanidae) (see Table 1), and a strong correlation  
249 was found between the log transformed number of reads in the preservative ethanol and the  
250 estimated biomass of each species (Pearson R = 0.88, p-value = 0.0001; Figure 2).

### 251 *Phylogenetic profile of the vouchers and the preservative ethanol*

252 The general taxonomic characterisation of the paired reads showed that in all libraries a large  
253 proportion of reads has no BLAST hits to our custom reference databases, with 95.3, 95.5,  
254 93.0 and 95.2% of reads unmatched in AV, TV, AE and TE, respectively. The inclusion of  
255 coleopteran genome assemblies (from NCBI *wgs* data) in the reference database contributed  
256 significantly to the MEGAN identification of arthropod nuclear DNA (compared to using  
257 NCBI Nucleotide reference set alone; see Suppl. File S2). This was particularly striking for  
258 the aquatic pool, for which the number of identified coleopteran reads increased by a factor  
259 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.

260 Identified reads showed different profiles in the voucher and ethanol libraries, but also  
261 between the two habitats (Figure 3). In the *voucher* libraries the great majority of these reads  
262 were apparently derived from the target specimens, with 78.6 and 77.4% identified as  
263 arthropod reads in AV and TV. This proportion was reduced in the *ethanol* libraries to 17.2  
264 and 7.1% in AE and TE. Other DNAs were present in low proportions in the vouchers but  
265 dominant in the preservative ethanol. In both *voucher* libraries, Proteobacteria were the 2<sup>nd</sup>  
266 most dominant clade. In AV, Proteobacteria are followed by Nematoda, Platyhelminthes and  
267 Chordata reads in decreasing proportions, with more than half of the Chordata reads identified  
268 as sequences of *Cyprinus carpio* (common Eurasian carp). Within Platyhelminthes, 10,158  
269 reads were assigned at the species level to the tapeworm *Hymenolepis diminuta*. No species-

270 level identifications were obtained for Nematoda, which produced scattered matches to  
271 numerous sub-taxa. TV showed a similar profile with a dominance of Proteobacteria,  
272 followed by a more diverse pattern of various bacterial phyla.

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

279 Further analyses allowed the assignment of the reads to three main groups, including (i)  
280 arthropods, (ii) taxa potentially associated to the gut or the environment, and (iii) bacterial  
281 endosymbionts. Their relative proportion was compared in the *voucher* and *ethanol* libraries  
282 (Figure 4, Suppl. Table S3). Generally, DNA reads were recovered, in decreasing order of  
283 abundance, from plastids, mitochondria and rRNA genes in eukaryotes, and from complete  
284 genomes and rRNAs in bacterial symbionts, reflecting that longer markers produced more  
285 read matches. In agreement with Figure 3, the proportion of Arthropoda reads in the *ethanol*  
286 was much lower than in the *vouchers* for both habitats. On average, a two-orders of  
287 magnitude ( $F=2.0$ ) loss was observed for both the mitochondrial and the rRNA sequences  
288 (Figure 4A). In contrast, read numbers for some taxa potentially associated with the  
289 environment and gut content (Figure 4B) were increased in the *ethanol* by between 2.2 (Fungi  
290 rRNA) to 4.6 (Annelida rRNA) orders of magnitude. Following Douglas et al. (2015), the  
291 symbiont species were divided into those with “closed associations” representing strict  
292 bacterial symbionts confined to bacteriocytes or specific host tissues, and those in “open

293 associations” representing bacterial infections, loose symbiotic interactions or commensals of  
294 the gut. All genera in closed associations (*Wolbachia*, *Rickettsia*, *Regiella*) showed a lower  
295 recovery from the *ethanol* compared to the *vouchers*, and *Wolbachia* and *Rickettsia*,  
296 respectively, were absent altogether in TE and AE, despite their strong signal in the *vouchers*  
297 (Figure 4C). On the other hand, symbiont genera with open associations showed more  
298 complex patterns, but in general recovery was higher or at least at similar levels in the *ethanol*  
299 than in the *vouchers*. Interestingly, in both TV and TE we noticed the presence of rRNA  
300 genes from endosymbionts typically associated with Collembola, possibly providing indirect  
301 evidence for predation on arthropod microfauna in some of the voucher specimens of the  
302 terrestrial pool (Figure 4C).

303

## 304 **Discussion**

### 305 **Species recovery and shotgun metagenomic sequencing from preservative ethanol**

306 Earlier PCR-based studies have demonstrated that specimen DNA can be obtained from the  
307 preservative ethanol (e.g. Shokralla *et al.* 2010; Hajibabaei *et al.* 2012), while here we  
308 established the power of direct shotgun sequencing, for a broader characterisation of the  
309 sampled specimens. PCR-based approaches are effective for detection of low DNA  
310 concentration templates, and thus have been successful for generating fairly complete species  
311 inventories from the ethanol fraction (Hajibabaei *et al.* 2012). We show that the number of  
312 DNA reads pertaining to the specimens themselves is rather low and, at the selected  
313 sequencing depth, less than half of species present in the samples could be identified from the  
314 reads, despite the availability of complete reference mitogenomes. If it is the aim of a study to

315 detect all species in the sample, PCR amplification may be the more efficient approach, but  
316 with the proviso that the specific primers used in the assay limit the outcome of the detected  
317 taxa (only *coxI* was used in previous studies). Alternatively, a combination of primer sets  
318 (Hajibabaei *et al.* 2012) can be used but holds the risk of cross-sample contamination, in  
319 particular if samples differ greatly in the concentration of DNA. In addition, the PCR  
320 approach may not be universally successful. In our attempts to replicate the *coxI* results on  
321 the ethanol samples generated here, we experienced a complete failure of amplification  
322 despite the use of various primers and PCR protocols (data not shown). The DNA  
323 concentration and level of preservation were sufficient for metagenomic libraries, which  
324 generally requires much more DNA template than the PCR, ruling out issues affecting the  
325 quality or quantity of the template for PCR failure. Instead, PCR inhibitors from the  
326 environment or the gut may be enriched in the ethanol fraction, which apparently affects the  
327 PCR, but less so the library construction and direct sequencing of the DNA.

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



339 The availability of reference sequences was a key requirement for the shotgun approach. We  
340 generated an almost-complete reference set of mitogenomes following an established protocol  
341 (Crampton-Platt *et al.* 2015, 2016). At the read depth used here (approximately 1% of a  
342 MiSeq flow cell per species) this procedure was highly efficient and even exceeded the  
343 species identification rate of *cox1* PCR-based Sanger sequencing of the same specimens. In  
344 addition, the *ethanol* libraries produced many matches to arthropod nuclear DNA, including  
345 rRNA genes that could be identified against external databases (Figure 4A). Although  
346 complementing mitochondrial references with rRNA markers would greatly increase the  
347 sensitivity of species recovery, the assembly of rRNA genes remains challenging. In our tests,  
348 no unequivocal contigs were produced in both TV and AV, despite the use of four different  
349 assemblers (Suppl. Table S4). While present in high copy number in metazoan genomes,  
350 alternating highly conserved and rapidly evolving expansion segments in the primary  
351 sequence of rRNA genes (Stage & Eickbush 2007) currently prevent the assembly from short  
352 sequence reads.

### 353 **Exploration of concomitant biodiversity from the preservative ethanol**

354 The *ethanol* libraries may be considered as complex ‘environmental DNA’ (eDNA) mixtures  
355 that include the DNA released from the focal specimens, together with organisms associated  
356 with these specimens and potentially unconnected organisms carried over from the wider  
357 ecosystem (Bohmann *et al.* 2014). Bacteria are expected to have a high chance of recovery in  
358 the DNA reads, as they are present in high copy numbers and they are detected by read  
359 matching against full genomes. Some bacterial genera detected in the ethanol are known to be  
360 associated to specific habitats (e.g. *Acinetobacter*, *Hydrogenophaga*; Figure 4B). These were  
361 present in small proportions (Figure 3), as would be expected in specimens collected

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

379 Other “closed association” bacterial endosymbionts show the reverse pattern, i.e. a higher  
380 DNA proportion in the vouchers than in the preservative ethanol. These species reside in the  
381 bacteriocytes, specialized intracellular compartments that are not expected to be released into  
382 the preservative medium. Specifically, *Wolbachia*, *Regiella* and *Rickettsia* are present in most  
383 arthropod communities (Werren *et al.* 2008) and in our samples are easily detectable in the  
384 voucher libraries but are poorly, if at all, recovered from the ethanol (Figure 4C). By contrast,  
385 several bacterial genera implicated in “open” symbiotic associations as commensals outside

386 of the bacteriocytes (Moran *et al.* 2005) show more mixed patterns. This category of bacteria  
387 appears to be the main candidate if one intends to use the preservative ethanol for the study of  
388 insect symbiont communities. Finally, some eukaryotic species relevant to insect biology  
389 were also detected (Figure 4). The Viridiplantae and Stramenophiles were greatly enriched in  
390 the ethanol (Figure 4) and may represent ingested food items. Potential infectious agents, such  
391 as the entomopathogenic fungus *Metharizium* (Jackson & Jaronski 2009) represented as much  
392 as 75% of fungal reads in TE. In contrast, the fungal genus *Hymenolepis* known to have  
393 parasitic life cycles using insects as intermediary hosts (Shostak 2014) is strongly detected in  
394 AV (10,160 reads identified to genus level) and its absence in AE suggests an association  
395 with internal tissues but not the gut content.

#### 396 **The value of the preservative ethanol**

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

404 For bulk samples the interactions cannot be ascribed to any particular species in the mixture,  
405 but the information is still highly valuable to characterise the functional or ecological  
406 components of an ecosystem *in toto*, for example through the parallel study of macro- and  
407 microbiomes of bulk samples. For higher precision, the methodology can be modified to  
408 include only members of a single species or possibly individually preserved specimens,

409 allowing comparisons among co-distributed species for analyses of resource segregation or  
410 the turnover in feeding source for a given species or assemblage among different sites.  
411 Additionally, the regurgitation of gut content into the ethanol provides a procedure for non-  
412 invasive DNA isolation for identification of food items, and it overcomes the problem that the  
413 degraded DNA of the gut content makes up only a small proportion of sequence reads  
414 compared to the well-preserved gut tissue that cannot be removed even with careful  
415 dissections (e.g. Paula *et al.* 2014). The greatest value of these techniques lies in the  
416 possibility for making comparison of numerous samples, each of them surveyed for multiple  
417 types of trophic interactions, given a different ecological context in which the target taxa are  
418 found. The high cost of shotgun sequencing relative to PCR-based metabarcoding may be a  
419 deterrent for such studies, but due to the emergence of cheaper methods for library  
420 construction (e.g. Baym *et al.* 2015) and the limited amount of sequencing required (e.g. 5%  
421 of MiSeq per sample in the current study), these costs are not prohibitive. Thus, the use of the  
422 preservative ethanol extends the metasytematic approach to biodiversity assessment and  
423 environmental monitoring, for more effective analysis and management of complex  
424 ecosystems (Gibson *et al.* 2014). The biomass-dependence of shotgun sequencing is another  
425 strength of this approach, to provide abundance estimates for ecological studies, while also  
426 recovering rare components without PCR biases. Increased sequencing depth and/or biomass  
427 pre-processing of the samples could be useful strategies when recovering low biomass entities  
428 is required. At the same time, the extension of reference databases, including complete  
429 mitochondrial genomes or nuclear genomes, will also increase the reliability of these  
430 approaches, reducing their dependency on the completeness of existing public databases.  
431 Beyond the study of freshly collected samples, the significance of bulk sampling and  
432 preservative sequencing may arise from the molecular analysis of historical spirit collections.

433 Museum collections provide enormous resources as a base-line against which modern  
434 observations can be compared, helping us to build predictive models in a world increasingly  
435 influenced by human activities (Suarez & Tsutsui 2004). A holistic approach to the study of  
436 preservative ethanol (specimen + eDNA) should reconsider specimen collection and storage  
437 practices. A widespread practice to obtain 'cleaner' samples from field collections is the  
438 replacement of the original ethanol fraction, which is usually discarded, but this procedure  
439 loses valuable information and efforts should be made to store this initial preservative (as  
440 volume can easily be reduced through evaporation). Ethanol should also be carefully  
441 considered in the management and maintenance of these collections, such as following  
442 protocols based on a "topping-up" of the ethanol (e.g. Notton 2010) instead of replacement.

443 Long-term microbiota characterisation appears to be a potential outcome from insect spirit  
444 collections. The ability to quantify the microbiotas in insect specimen vs. ethanol fractions  
445 can establish their relationships with the 'host' specimens, while the co-existence of similar  
446 organisms within samples from different ecosystems may uncover the pathogenic or  
447 ecological role played by the insect microbiome (Mira *et al.* 2010). Similarly, organisms  
448 attached to the surface of specimens, such as pollen in the leg baskets of bees or fungi  
449 contained in the mycangia of wood-boring beetles, may be present in the preservative  
450 medium. Such molecular information can complement the information associated to  
451 collection records making the ethanol metagenome itself a record from which more  
452 associations may be identified in the future when more DNA reads will be identified against  
453 the growing genome reference set. Further studies on the dynamics of DNA transfer from  
454 specimens to ethanol under different conditions and how this DNA degrades through time are  
455 needed to uncover the full potential of the preserving ethanol into which specimens are  
456 collected. But it appears that preservative ethanol is an unexpected source of molecular

457 knowledge: it will contain both the specimen and concomitant biodiversity and can provide  
458 valuable biological information when subjected to shallow metagenomic sequencing.

459

## 460 **Acknowledgements**

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

## 470 **Data Accessibility**

471 GenBank Accessions numbers for voucher specimens are KT876876-KT876902; KT876904-  
472 KT876915; original datasets have been uploaded as fastq files in Dryad doi:  
473 doi:10.5061/dryad.jr6r5; all supplementary details, tables and figures cited in the main text  
474 have been uploaded as online Supporting Information.

## 475 **Author Contributions**

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

479

## 480 **References**

- 481 Andersen K, Bird KL, Rasmussen M *et al.* (2012) Meta-barcoding of “dirt” DNA from soil reflects  
482 vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–79.
- 483 Andújar C, Arribas P, Ruzicka F *et al.* (2015) Phylogenetic community ecology of soil biodiversity  
484 using mitochondrial metagenomics. *Molecular Ecology*, **24**, 3603–3617.
- 485 Baym M, Kryazhimskiy S, Lieberman TD *et al.* (2015) Inexpensive multiplexed library preparation  
486 for megabase-sized genomes. *PLoS ONE*, **10**, 1–15.
- 487 Bernt M, Donath A, Jühling F *et al.* (2013) MITOS: Improved de novo metazoan mitochondrial  
488 genome annotation. *Molecular Phylogenetics and Evolution*, **69**, 313–319.
- 489 Bohmann K, Evans A, Gilbert MTP *et al.* (2014) Environmental DNA for wildlife biology and  
490 biodiversity monitoring. *Trends in Ecology & Evolution*, **29**, 358–67.
- 491 Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray Meta: scalable de novo  
492 metagenome assembly and profiling. *Genome Biology*, **13**, R122.
- 493 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data.  
494 *Bioinformatics*, **30**, 2114–2120.
- 495 Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC*  
496 *Bioinformatics*, **10**, 421.
- 497 Carbajal-Rodríguez I, Stöveken N, Satola B, Wübbeler JH, Steinbüchel A (2011) Aerobic degradation  
498 of mercaptosuccinate by the gram-negative bacterium *variovorax paradoxus* strain B4. *Journal of*  
499 *Bacteriology*, **193**, 527–539.
- 500 Carrino-Kyker SR, Swanson AK (2008) Temporal and spatial patterns of eukaryotic and bacterial  
501 communities found in vernal pools. *Applied and Environmental Microbiology*, **74**, 2554–2557.
- 502 Caspers H (1986) Aquatic Oligochaeta. Proceedings of the Second International Symposium on  
503 Aquatic Oligochaete Biology, held in Pallanza, Italy, September 1982. *Internationale Revue der*  
504 *gesamten Hydrobiologie und Hydrographie*, **71**, 583–583.
- 505 Caspi-Fluger A, Inbar M, Mozes-Daube N *et al.* (2011) Rickettsia “in” and “out”: Two different  
506 localization patterns of a bacterial symbiont in the same insect species. *PLoS ONE*, **6**.
- 507 Cordaux R, Paces-Fessy M, Raimond M *et al.* (2007) Molecular characterization and evolution of  
508 arthropod-pathogenic Rickettsiella bacteria. *Applied and Environmental Microbiology*, **73**, 5045–  
509 5047.
- 510 Crampton-Platt A, Timmermans MJTN, Gimmel ML *et al.* (2015) Soup to Tree: The Phylogeny of  
511 Beetles Inferred by Mitochondrial Metagenomics of a Bornean Rainforest Sample. *Molecular*  
512 *Biology and Evolution*, **32**, 2302–2316.
- 513 Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes  
514 out of the bottle. *GigaScience*, **5**, 15.



- 515 Douglas AE (2015) Multiorganismal Insects: Diversity and Function of Resident Microorganisms.  
516 *Annual Review of Entomology*, **60**, 17–34.
- 517 Envall I, Källersjö M, Erséus C (2006) Molecular evidence for the non-monophyletic status of  
518 Naidinae (Annelida, Clitellata, Tubificidae). *Molecular Phylogenetics and Evolution*, **40**, 570–  
519 84.
- 520 Fonseca VG, Carvalho GR, Sung W *et al.* (2010) Second-generation environmental sequencing  
521 unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.
- 522 Gasparich GE, Whitcomb RF, Dodge D *et al.* (2004) The genus Spiroplasma and its non-helical  
523 descendants: phylogenetic classification, correlation with phenotype and roots of the  
524 Mycoplasma mycoides clade. *International Journal of Systematic and Evolutionary*  
525 *Microbiology*, **54**, 893–918.
- 526 Gibson J, Shokralla S, Porter TM *et al.* (2014) Simultaneous assessment of the macrobiome and  
527 microbiome in a bulk sample of tropical arthropods through DNA metasytematics. *Proceedings*  
528 *of the National Academy of Sciences of the United States of America*, **111**, 8007–12.
- 529 Gillett CPDT, Crampton-Platt A, Timmermans MJTN *et al.* (2014) Bulk de novo mitogenome  
530 assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera:  
531 Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.
- 532 Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, Baselga A, Vogler AP (2015)  
533 Validating the power of mitochondrial metagenomics for community ecology and phylogenetics  
534 of complex assemblages. *Methods in Ecology and Evolution*, **6**, 883–894.
- 535 Grimont F, Grimont PD (2006) The Genus Serratia. In: *The Prokaryotes SE - 11* (eds Dworkin M,  
536 Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E), pp. 219–244. Springer New York.
- 537 Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a  
538 freshwater benthic macroinvertebrate community through non-destructive environmental  
539 barcoding of DNA from preservative ethanol. *BMC Ecology*, **12**, 28.
- 540 Haselkorn TS, Markow TA, Moran NA (2009) Multiple introductions of the Spiroplasma bacterial  
541 endosymbiont into Drosophila. *Molecular Ecology*, **18**, 1294–305.
- 542 Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome*  
543 *Research*, **17**, 377–86.
- 544 Jackson MA, Jaronski ST (2009) Production of microsclerotia of the fungal entomopathogen  
545 Metarhizium anisopliae and their potential for use as a biocontrol agent for soil-inhabiting  
546 insects. *Mycological Research*, **113**, 842–850.
- 547 Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) “Sight-unseen” detection of rare aquatic  
548 species using environmental DNA. *Conservation Letters*, **4**, 150–157.
- 549 Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity  
550 via metabarcoding. *Ecology Letters*.
- 551 Koga R, Meng X-Y, Tsuchida T, Fukatsu T (2012) Cellular mechanism for selective vertical

552 transmission of an obligate insect symbiont at the bacteriocyte-embryo interface. *Proceedings of*  
553 *the National Academy of Sciences of the United States of America*, **109**, E1230–7.

554 L. Dijkshoorn AN (2008) The diversity of the genus *Acinetobacter*. In: *Acinetobacter Molecular*  
555 *Microbiology*, p. 348. Horizon Scientific Press.

556 Linard B, Crampton-Platt A, Timmermans MJTN, Vogler AP (2015) Metagenome skimming of insect  
557 specimen pools: potential for comparative genomics. *Genome Biology and Evolution*, **7**, 1474–  
558 1489.

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

561 Mira A, Martín-Cuadrado AB, D’Auria G, Rodríguez-Valera F (2010) The bacterial pan-genome: a  
562 new paradigm in microbiology. *International microbiology : the official journal of the Spanish*  
563 *Society for Microbiology*, **13**, 45–57.

564 Morales-Jiménez J, Zúñiga G, Villa-Tanaca L, Hernández-Rodríguez C (2009) Bacterial community  
565 and nitrogen fixation in the red turpentine beetle, *Dendroctonus valens* LeConte (Coleoptera:  
566 Curculionidae: Scolytinae). *Microbial Ecology*, **58**, 879–91.

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

570 Myers EW (2000) A Whole-Genome Assembly of *Drosophila*. *Science*, **287**, 2196–2204.

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

573 Paula DP, Linard B, Andow D *et al.* (2014) Detection and decay rates of prey and prey symbionts in  
574 the gut of a predator through metagenomics. *Molecular Ecology Resources*.

575 Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and  
576 metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, **28**,  
577 1420–8.

578 Quast C, Pruesse E, Yilmaz P *et al.* (2013) The SILVA ribosomal RNA gene database project:  
579 improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–6.

580 Russell JA, Funaro CF, Giraldo YM *et al.* (2012) A Veritable Menagerie of Heritable Bacteria from  
581 Ants, Butterflies, and Beyond: Broad Molecular Surveys and a Systematic Review. *PLoS ONE*,  
582 **7**.

583 Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets.  
584 *Bioinformatics*, 3–5.

585 Schnell IB, Thomsen PF, Wilkinson N *et al.* (2012) Screening mammal biodiversity using DNA from  
586 leeches. *Current Biology : CB*, **22**, R262–3.

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

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

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

594 Sogin ML, Morrison HG, Huber JA *et al.* (2011) Microbial Diversity in the Deep Sea and the  
595 Underexplored “Rare Biosphere.” *Handbook of Molecular Microbial Ecology II: Metagenomics*  
596 *in Different Habitats*, 243–252.

597 Stage DE, Eickbush TH (2007) Sequence variation within the rRNA gene loci of 12 Drosophila  
598 species. *Genome Research*, **17**, 1888–97.

599 Straub SCK, Parks M, Weitemier K *et al.* (2012) Navigating the tip of the genomic iceberg: Next-  
600 generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–64.

601 Suarez A V., Tsutsui ND (2004) The Value of Museum Collections for Research and Society.  
602 *BioScience*, **54**, 66.

603 Tang M, Hardman CJ, Ji Y *et al.* (2015) High-throughput monitoring of wild bee diversity and  
604 abundance via mitogenomics (M Gilbert, Ed.). *Methods in Ecology and Evolution*, doi:  
605 10.1111/2041–210X.12416.

606 Thomsen PF, Kielgast J, Iversen LL *et al.* (2012) Monitoring endangered freshwater biodiversity using  
607 environmental DNA. *Molecular Ecology*, **21**, 2565–73.

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

611 Tsuchida T, Koga R, Fujiwara A, Fukatsu T (2014) Phenotypic Effect of “Candidatus Rickettsiella  
612 viridis,” a Facultative Symbiont of the Pea Aphid (*Acyrtosiphon pisum*), and Its Interaction  
613 with a Coexisting Symbiont. *Applied and Environmental Microbiology*, **80**, 525–533.

614 Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology.  
615 *Nature reviews. Microbiology*, **6**, 741–51.

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

618 Yoshikawa H, Wu Z, Howe J *et al.* (2007) Ultrastructural and phylogenetic studies on Blastocystis  
619 isolates from cockroaches. *The Journal of Eukaryotic Microbiology*, **54**, 33–7.

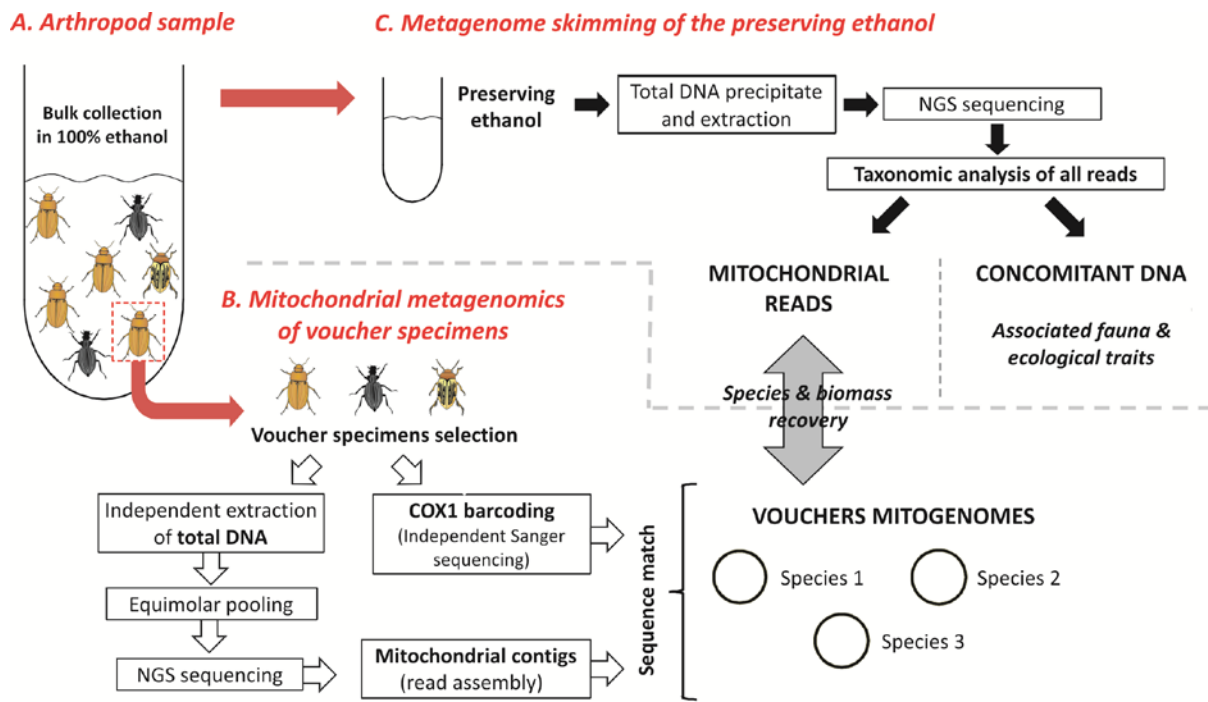
620 Yun J-H, Roh SW, Whon TW *et al.* (2014) Insect gut bacterial diversity determined by environmental  
621 habitat, diet, developmental stage, and phylogeny of host. *Applied and Environmental*  
622 *Microbiology*, **80**, 5254–64.

623

**Table 1.** Dataset 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.

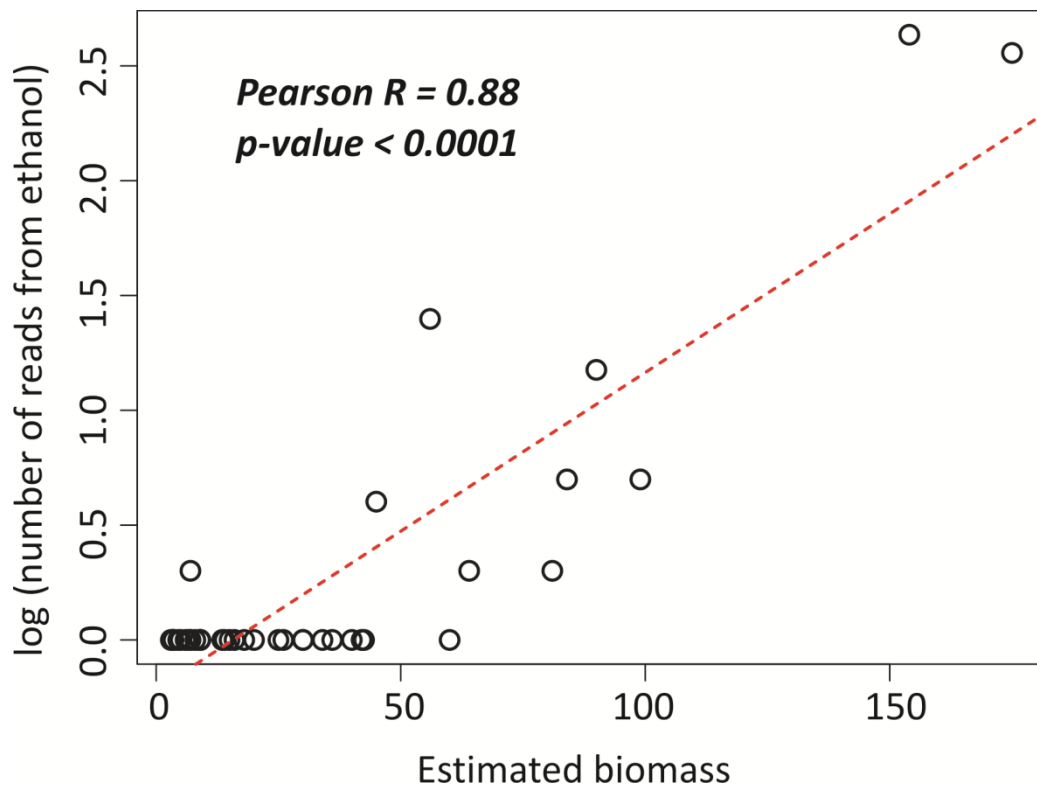
Species	Community	Stage	Total specimens	Specimens used as vouchers	Total estimated biomass	cox1_Sanger	mitogenome	ethanol reads matching cox1	ethanol reads matching complete mitogenomes	ethanol reads matching protein-coding mito-genes
<i>Acilius sulcatus</i> BMNH1425211	Aquatic	adult	2	1	36	X	X	0	0	0
<i>Berosus affinis</i> BMNH1425169	Aquatic	adult	3	2	13.5	X	X	0	0	0
<i>Colymbetes fuscus</i> BMNH1425212	Aquatic	adult	5	2	90	X	X	0	15	15
<i>Dryops luridus</i> BMNH1425163	Aquatic	adult	4	3	20	X	X	0	2	1
<i>Haliphus immaculatus</i> BMNH1425121	Aquatic	adult	3	2	9	X	X	0	0	0
<i>Haliphus lineatocollis</i> BMNH1425118	Aquatic	adult	5	3	15	X	X	0	2	0
<i>Helochaeres</i> sp. BMNH1425100	Aquatic	adult	10	4	60	X	X	0	0	0
<i>Hydrochus</i> sp. BMNH1425167	Aquatic	adult	2	2	6	X	X	0	0	0
<i>Hydroporus planus</i> BMNH1425115	Aquatic	adult	1	2	4.5	X	X	0	0	0
<i>Hydroporus discretus</i> BMNH1425116	Aquatic	adult	2	2	8	X	X	0	0	0
<i>Hydroporus gyllenhalii</i> BMNH1425127	Aquatic	adult	2	2	7	X	X	0	2	0
<i>Hydroporus obscurus</i> BMNH1425129	Aquatic	adult	1	2	3.5	X	X	0	0	0
<i>Hydroporus erythrocephalus</i> BMNH1425131	Aquatic	adult	27	3	81	X	X	0	2	2
<i>Hydropsyche pellucidulla</i> BMNH1425186	Aquatic	larva	4	2	56	X	X	2	55	25
<i>Hygrobia hermanni</i> BMNH1425190	Aquatic	adult	3	1	30	X	X	0	0	0
<i>Hygrotus inaequalis</i> BMNH1425126	Aquatic	adult	1	1	3	X	X	0	1	1
<i>Hygrotus impressopunctatus</i> BMNH1425158	Aquatic	adult	5	3	25	X	X	0	0	0
<i>Hygrotus confluens</i> BMNH1425172	Aquatic	adult	1	1	3.5	X	X	0	0	0
<i>Liopterus haemorrhoidalis</i> BMNH1425193	Aquatic	adult	6	2	42	X	X	0	0	0
<i>Noterus clavicornis</i> BMNH1425090	Aquatic	adult	22	3	99	X	X	0	9	5
<i>Sialis lutaria</i> BMNH1425199	Aquatic	larva	11	2	154	NO	X	24	476	432
<i>Abax parallelepipedus</i> BMNH1425236	Terrestrial	adult	2	1	40	X	X	0	0	0
<i>Agriotes obscurus</i> BMNH1425233	Terrestrial	larva	2	1	30	X	X	0	0	0
<i>Anisosticta novemdecimpunctata</i> BMNH1425231	Terrestrial	adult	1	1	3.5	NO	X	0	0	0
<i>Athous haemorrhoidalis</i> BMNH1425235	Terrestrial	larva	1	1	9	X	X	0	1	1
<i>Atrecus affinis</i> sp. BMNH1425232	Terrestrial	adult	1	1	7	X	X	0	2	2
<i>Calathus melanocephalus</i> BMNH1425227	Terrestrial	adult	1	1	7	NO	X	0	0	0
<i>Cyphon variabilis</i> BMNH1425225	Terrestrial	adult	2	2	9	X	X	0	0	0
<i>Dorcus parallelepipedus</i> BMNH1425260	Terrestrial	larva	7	1	175	X	X	17	478	360
<i>Melanotus villosus</i> BMNH1425245	Terrestrial	larva	8	4	45	X	X	0	6	4
<i>Nalassus laevioctostriatus</i> BMNH1425217	Terrestrial	adult	5	2	42.5	X	X	0	0	0
<i>Nebria brevicollis</i> BMNH1425256	Terrestrial	adult	1	1	14	X	X	0	0	0
<i>Ocyopus olens</i> BMNH1425259	Terrestrial	larva	1	1	16	X	X	0	0	0
<i>Pterostichus niger</i> BMNH1425241	Terrestrial	adult	4	1	84	X	X	0	12	5
<i>Pterostichus madidus</i> BMNH1425238	Terrestrial	adult	4	2	64	X	X	0	2	2
<i>Stenus clavicornis</i> BMNH1425222	Terrestrial	adult	3	2	18	X	X	0	0	0
<i>Stenus boops</i> BMNH1425230	Terrestrial	larva	1	1	5	X	X	0	0	0
<i>Stomis pumicatus</i> BMNH1425229	Terrestrial	adult	1	1	6.5	X	X	0	0	0
<i>Tasgius</i> sp. BMNH1425251	Terrestrial	adult	2	1	34	X	NO	7	0	0
<i>Uloma</i> sp. BMNH1425257	Terrestrial	larva	2	2	26	X	X	0	0	0

**Figure 1**



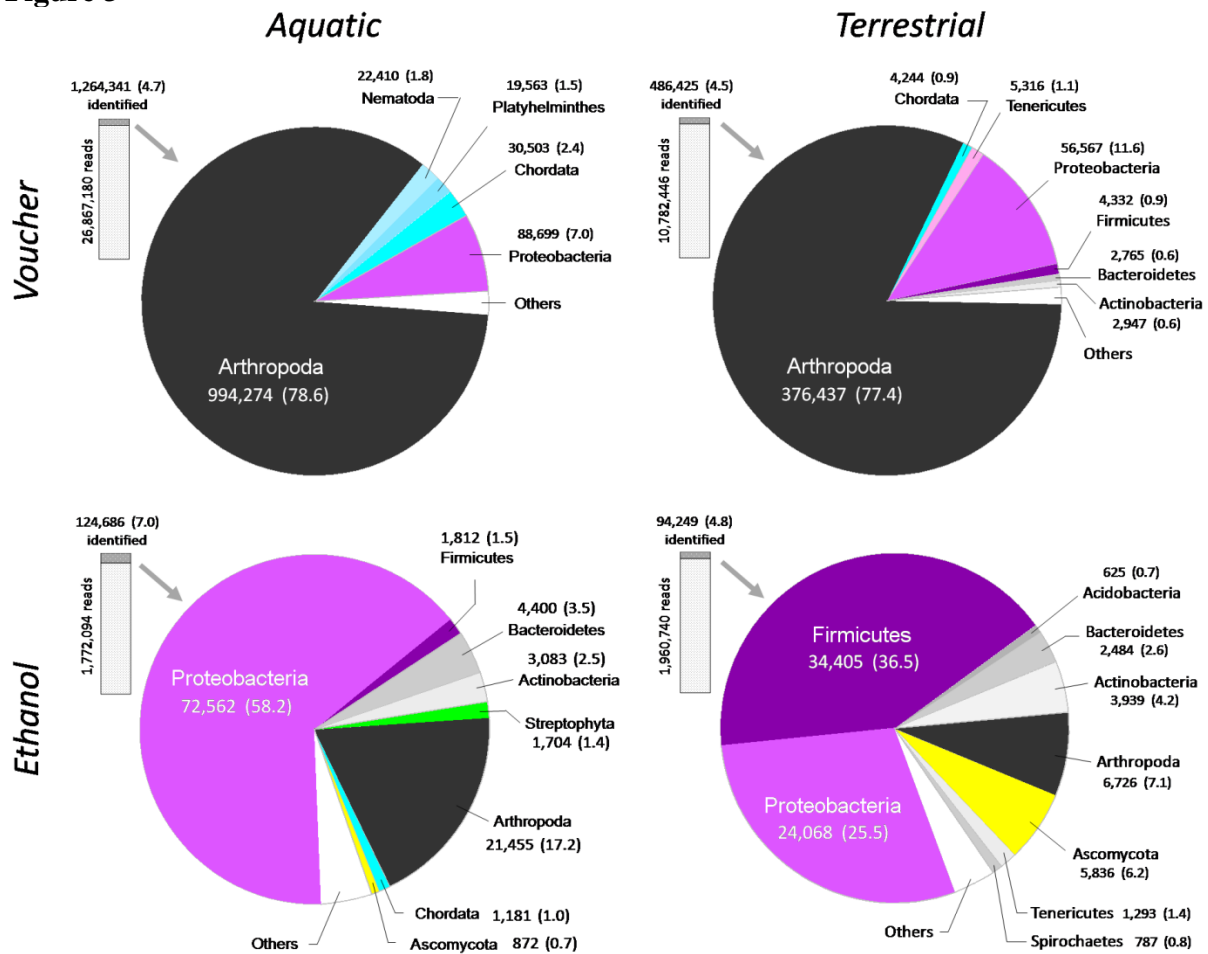
**Figure 1** Schematic representation of the experimental design and bioinformatics pipeline followed in this study.

**Figure 2**



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

**Figure 3**



**Figure 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.

**Figure 4**

	Clade	Marker	Aquatic		$\Delta F_{E/V}$ (log)	Terrestrial		$\Delta F_{E/V}$ (log)	Comments	
			V	E		V	E			
<b>A.</b>	<b>Arthropoda</b>	Mito			1.9			2.0		
		rRNA			2.0			2.0		
<b>B. Environment and gut content</b>	Eukaryota	Annelides			4.6				>99% similar to Enchytraeidae and Naididae, found in benthic and wet soil habitats <sup>a,b</sup>	
		Fungi			---			1.8	In TE, 75% of mito. reads are >99% similar to Metarhizium, an entomopathogen genera <sup>c</sup>	
		Viridiplantae	Plastid			3.7			3.2	
			Mito			3.5			3.0	
		rRNA			4.6			---		
	Stramenophiles	Plastid			3.3			3.2		
		Mito			2.3			2.9		
		Blastocystis			4.1			3.2	Insect gastrointestinal tracts habitat <sup>d</sup>	
	Bacteria	Acinetobacter						1.8	Soil mineralization and found in beetle guts <sup>e,f</sup>	
		Hydrogenophaga			2.7				Oxygenates-rich water habitats <sup>g</sup>	
Variovorax				2.8				Soil and water habitats <sup>h,i</sup>		
<b>C. Bacterial symbionts</b>	Closed association	Wolbachia			1.9			>2.0	Intracellular facultative endosymbiont, Widespread in arthropods <sup>j</sup>	
		rRNA			>2.0			>2.0		
		Regiella			1.8			1.9	Facultative symb. associated to bacteriocytes <sup>k</sup>	
	Rickettsia			>2.0				"Scattered" association to bacteriocytes <sup>l</sup>		
	Open association	Collembola endosym.						1.7	Coxiellaceae symbiont (unpublished, gi:13507245)	
		Rickettsiella	Genomes			3.6			1.9	Intracellular pathogens of arthropods <sup>m</sup> interacting with coexisting endosymbionts <sup>n</sup>
			rRNA						2.0	
Serratia		Genomes			2.5			1.1	Genera found ubiquitously in water, soil and insect guts habitats <sup>o</sup> Some species are facultative symbionts playing a role in bacteriocyte/embryo transmission <sup>p</sup>	
rRNA							2.4			
Spiraplasma	Genomes						1.5	Found in plants/insect guts <sup>q</sup> , heritable symbiont in some insect species <sup>r</sup>		

**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 Figure 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<sub>10</sub> 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: a. Caspers (1986) b. Envall *et al.* (2006) c. Jackson & Jaronski (2009) d. Yoshikawa *et al.* (2007) e. Morales-Jiménez *et al.* (2009) f. L. Dijkshoorn (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).