Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 1142–1150 doi: 10.1111/2041-210X.12120

SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons

Shanlin Liu[1], Yiyuan Li[1], Jianliang Lu[1], Xu Su[1], Min Tang[1], Rui Zhang[1], Lili Zhou[1], Chengran Zhou[1,2], Qing Yang[1], Yinqiu Ji[3], Douglas W. Yu[3,4] and Xin Zhou[1]*

[1]*China National GeneBank, BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen, Guangdong Province 518083, China;* [2]*College of Life Sciences, Sichuan University, Chengdu, Sichuan Province 610000, China;* [3]*State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China; and* [4]*School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR4 7TJ, UK*

Summary

1. Metabarcoding of mixed arthropod samples for biodiversity assessment has mostly been carried out on the 454 GS FLX sequencer (Roche, Branford, Connecticut, USA), due to its ability to produce long reads (≥400 bp) that are believed to allow higher taxonomic resolution. The Illumina sequencing platforms, with their much higher throughputs, could potentially reduce sequencing costs and improve sequence quality, but the associated shorter read length (typically < 150 bp) has deterred their usage in next-generation-sequencing (NGS)-based analyses of eukaryotic biodiversity, which often utilize standard barcode markers (e.g. *COI*, *rbcL*, *matK*, *ITS*) that are hundreds of nucleotides long.

2. We present a new Illumina-based pipeline to recover full-length COI barcodes from mixed arthropod samples. Our new assembly program, *SOAPBarcode*, a variant of the genome assembly program *SOAPdenovo*, uses paired-end reads of the standard COI barcode region as anchors to extract the correct pathways (sequences) out of otherwise chaotic '*de Bruijn* graphs', which are caused by the presence of large numbers of COI homologs of high sequence similarity.

3. Two bulk insect samples of known species composition have been analysed in a recently published 454 metabarcoding study (Yu *et al.* 2012) and are re-analysed by our analysis pipeline. Compared to the results of Roche 454 (*c.* 400-bp reads), our pipeline recovered full-length COI barcodes (658 bp) and 17–31% more species-level operational taxonomic units (OTUs) from bulk insect samples, with fewer untraceable (novel) OTUs. On the other hand, our PCR-based pipeline also revealed higher rates of contamination across samples, due to the Illumina's increased sequencing depth. On balance, the assembled full-length barcodes and increased OTU recovery rates resulted in more resolved taxonomic assignments and more accurate beta diversity estimation.

4. The HiSeq 2000 and the *SOAPBarcode* pipeline together can achieve more accurate biodiversity assessment at a much reduced sequencing cost in metabarcoding analyses. However, greater precaution is needed to prevent cross-sample contamination during field preparation and laboratory operation because of greater ability to detect non-target DNA amplicons present in low-copy numbers.

Key-words: high-throughput sequencing, metabarcoding, next-generation-sequencing, operational taxonomic units, phylogenetic diversity, species richness, standard barcode

Introduction

Recently, DNA *metabarcoding* (Taberlet *et al.* 2012a), which takes advantage of high-throughput capacity of next-generation-sequencing (NGS) platforms, has emerged for characterizing the biodiversity of large volumes of eukaryote samples that consist of upwards of hundreds of thousands of individuals and thousands of species. Metabarcoding has been tested for large-scale surveys of metazoan, plant and fungal biodiversity from mass samples and on environmental extracellular DNA collected from water and soil (Taberlet *et al.* 2012b). Most studies have used universal PCR primers to mass amplify a taxonomically informative gene from collections of organisms or from environmental DNA, for example *COI* for insects (Hajibabaei *et al.* 2011; Yu *et al.* 2012) and short hypervariable regions of non-standard-barcode markers such as *18S SSU rRNA*, *trnL* and *ITS* (Creer *et al.* 2010; Yoccoz *et al.* 2012). Zhou *et al.* (2013) have also shown how to use *de novo* assembly to extract large portions of mitochondrial genomes from non-PCR-amplified genomic DNA

*Correspondence author. E-mail: xinzhou@genomics.cn

samples, which avoids the taxonomic bias introduced by PCR. There are now too many metabarcoding papers to list here, but some entries to the literature are Creer *et al*. 2010; Bik *et al*. 2012; Yu *et al*. 2012 and a special issue in *Molecular Ecology* (Taberlet *et al*. 2012a). The metabarcoding of mixed environmental samples opens the way to what Baird & Hajibabaei (2012) call 'BIOMONITORING 2.0'.

Compared to short molecular markers, such as fragments from 12S or 16S genes, the use of full-length DNA barcodes in NGS biodiversity has an obvious advantage in terms of leveraging existing barcode reference data bases, such as the Barcode of Life Data Systems (BOLD, www.boldsystems.org), which currently holds >2 million barcodes for animals, plants, fungi and other groups (accessed on March 20, 2013). With the aim of taking advantage of the sequence information in BOLD, most NGS-based biodiversity studies to date have been carried out using Roche 454 GS FLX technology, due to its ability to produce long sequence reads (≥400 bp). However, while longer barcodes deliver greater taxonomic resolution, the 454 only generates *c*. 700 000 reads per run. In contrast, the Illumina HiSeq 2000 can produce up to 3 billion reads per run for a similar cost (Quail *et al*. 2012). This much greater sequencing coverage, which improves the likelihood of detection of low-biomass species and reduces costs (Zhou *et al*. 2013), is balanced against shorter maximum read lengths of 150 bp [or 2 × 150 paired-end (PE)] and consequent lost taxonomic information.

To test whether the HiSeq 2000 platform can be used to produce COI barcodes, which are 658-bp long, we PCR-amplify, sequence, assemble and analyse two bulk arthropod samples of known composition that had been previously sequenced using the Roche 454 sequencing platform by Yu *et al*. (2012). We show that, enabled by a new sequencing pipeline and software package, SOAPBarcode, the short Illumina reads can be assembled into full-length COI barcode sequences, thereby increasing taxonomic resolution per barcode, reducing drop-out rates and rendering beta diversity characterization more accurate, relative to 454-generated data sets.

Materials and methods

BULK ARTHROPOD SAMPLES

In Yu *et al*. (2012), bulk arthropod samples of three sites in Yunnan Province, China – Xishuangbanna (XSBN), Kunming (KMG) and Honghe (HONGHE) – were sampled using Malaise traps. A total of 318, 795 and 316 arthropod individuals were hand-picked from HONGHE, XSBN and KMG, respectively (Yu *et al*. 2012). All specimens (mostly insects) were individually Sanger-sequenced first, and then, the genomic DNA of each individual was pooled within each sample for subsequent bulk PCR and 454 sequencing (Table 1). Yu *et al*. (2012) found that, overall, 76% of the input species could be recovered after 454-sequencing and informatics analysis.

PCR AMPLIFICATION AND SEQUENCING

One to 3 μL of DNA extracted from individual specimens collected from XSBN and KMG, with a concentration of 30–50 ng/μL, was mixed together to recreate each bulk sample from Yu *et al*. (2012). Because the quantity of HONGHE DNA was insufficient for Illumina library construction, this sample was not analysed. Using the same primers and PCR conditions described by Yu *et al*. (2012), six independent PCR were carried out for each bulk sample to obtain enough DNA product and to minimize random amplification bias. Each PCR contained 1 μL DNA template, 2·5 μL 10*EX-taq buffer, 0·2 mM dNTPs, 0·4 μM of each primer and 1·5 U EX-Taq DNA polymerase (TaKaRa Biosystems, Otsu, Japan), with a touchdown thermocycling profile of 95 °C for 2 min; 11 cycles of 95 °C for 15 s; 51 °C for 30 s; 72 °C for 3 min, decreasing the annealing temperature by 1 degree every cycle; then 21 cycles of 95 °C for 15 s, 41 °C for 30 s, 72 °C for 3 min and a final extension of 72 °C for 10 min. The amplification products were purified using QIAquick PCR purification kit (Qiagen, Venlo, the Netherlands).

The six amplification products per bulk sample were pooled and then evenly divided into two parts to construct two Illumina insert-size libraries: (i) a full-length library (insert size = 658 bp) and (ii) a shotgun library (insert size = 200 bp). For convenience, we call the two sequence data sets: full-length library sequences (FLS) and shotgun library sequences (SLS). Both libraries were sequenced using the 150-bp PE protocol on an Illumina HiSeq 2000 according to the manufacturer's instructions. A base-calling pipeline (Sequencing Control Software, SCS; Illumina, San Diego, California, USA) was used to process the raw fluorescent images and to call sequences.

Our rationale for building two libraries is as follows. The first (FLS) library is the Illumina equivalent of the library that is normally sent to be sequenced on the 454 machine. The Illumina HiSeq 2000 can only sequence, at most, 150 bp at each end, leaving *c*. 400-bp unsequenced gap in the middle with primers being trimmed, which is a much inferior result compared to the 454's ≥400-bp reads.

The purpose of the second (SLS) library is to bridge the gap. The 707-bp-long (658 bp + two primers) PCR products are broken randomly into 200-bp-long 'shotgun' sequences, which are short enough to be sequenced in full on an Illumina (by overlapping the two 150-bp sequences from each end). Naturally, these shotgun reads include many that partially overlap both the gaps and the FLS sequences, plus many

Table 1. Sample information for XSBN, KMG and HONGHE

Database		XSBN	KMG	HONGHE
Specimens	Number of individuals	795	316	318
Sanger sequences	Unique haplotypes	292	184	197
	USEARCH-clustered operational taxonomic units (OTUs) (98% similarity)[1]	230	153	167
454 sequencing	CROP-clustered OTUs (≥97% similarity)[2]	183 (148)[3]	174 (118)	192 (122)

[1]USEARCH from (Edgar 2010), version 5.1.221.
[2]CROP from Hao, Jiang & Chen (2011).
[3]OTUs recovered using 454 sequencing and informatics pipeline by Yu *et al*. (2012) (number of true positives).

that overlap the middle portion of the gap. This allows special genome assembly software to use the (SLS) shotgun reads to step from one sequenced end of a (FLS) gapped 658-bp read to the other end. The (scientific) magic of genome assembly software is that it can unmix everything, choosing the correct shotgun reads out of the SLS library to bridge the gap of each read in the FLS library. A schematic of the Illumina pipeline is shown in Fig. 1.

ANALYSIS

SOAPBarcode

The SOAPBarcode pipeline uses two main steps to transform raw Illumina reads into full-length COI barcode sequences. All relevant scripts are available at https://sourceforge.net/projects/metabarcoding/. The user's manual is in Data S1 (Supporting Information).

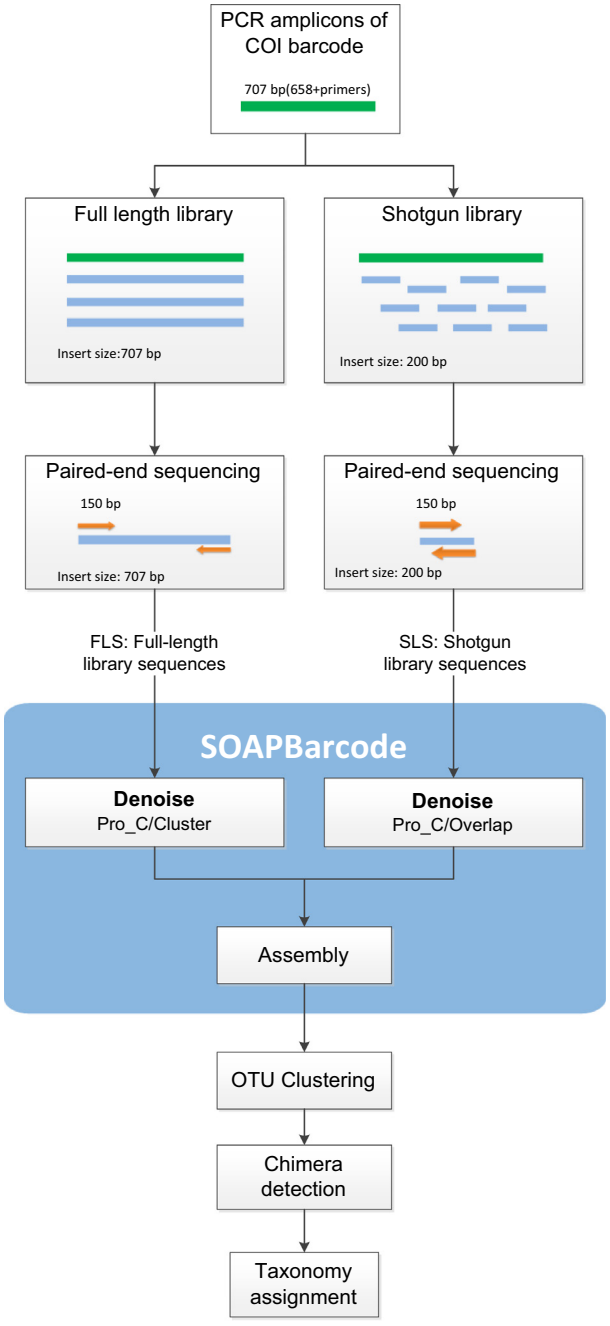

Fig. 1. Schematic of the PCR-based Illumina metabarcoding pipeline.

Denoising. *Full-length library sequences:* First, the primer sequences were removed from both the 5′ and 3′ ends. After the removal of primer sequences, all Illumina reads had a fixed length of 127 bp on the 5′ end of the full-length barcode and 124 bp on the 3′ end. Then, these PE reads were connected and checked for duplicates. Singletons were excluded, on the assumption that these were caused by PCR or sequencing errors. After that, because the animal barcode (COI) is a protein-coding gene, and thus, conserved, we developed a *perl* program, Pro_C, to exclude all reads that did not match the hydrophily (see Terminology in Data S3, Supporting Information) profile of a consensus arthropod COI amino acid sequence, on the assumption that non-matching reads are not COI amplicons or heavily error-ridden (see the following 'Pro_C workflow' section). Finally, USEARCH (Edgar 2010) was applied to filter chimeras (see Terminology in Data S3, Supporting Information), and the remaining sequences were then clustered at 98% to improve computational efficiency for SOAPBarcode assembly.

Shotgun library sequences: The software package COPE (Liu *et al.* 2012) was used to compare the forward and reverse reads, discarding sequences with overlap similarity <95%, which indicates sequencing errors. Then, the contigs were checked for duplicates and amino acid conservation using Pro_C.

Pro_C workflow: First, Pro_C discarded reads that could not be translated to amino acids and those containing stop codons. Secondly, Pro_C checked the conservativeness of hydrophily for the remaining Illumina reads against a comprehensive COI sequence reference library. COI sequences that fell out of the hydrophily ranges were removed from subsequent analyses. We built the reference library by downloading all full-length arthropod COI sequences from GenBank and filtering out non-iBOL compliant reads (e.g. sequences with >1% Ns or those obtained via single-direction sequencing), resulting in a total of 302 476 COI sequences. Taxonomic information and the corresponding number of sequences included in this COI reference are summarized in Table S1 (Supporting Information). We then translated the nucleotide sequences into amino acids and calculated hydrophily values using a 3-bp sliding window (see Fig. S1, Supporting Information). By examining amino acid hydrophily properties, Pro_C can remove errors caused by PCR, sequencing and some nuclear copies of COI (numts).

De novo assembly

The filtered reads were assembled using a modified algorithm of the widely used SOAPdenovo genome assembly software (Li *et al.* 2008), which adopts the *de Bruijn* graph data structure (see Terminology in Data S3, Supporting Information). Different from typical metagenomic data – which include shotgun reads from multiple regions of mixed genomes, the NGS reads derived from mixed COI amplicons by definition contain a large number of homologous fragments that can possess high levels of sequence similarity among taxa. When the *de Bruijn* graph was constructed using existing genome assembly software, these homogeneous COI fragments were merged into one gargantuan and reticulated graph, leaving it impossible to construct scaffolds corresponding to individual COI sequences.

We thus devised a new assembly strategy, SOAPBarcode, which uses the PE reads in the FLS amplicons as anchor points for the 5′ and 3′ terminuses of the COI sequence and builds scaffolds upon this sequence framework.

The following description of the assembly algorithm unavoidably employs terms from genome assembly informatics. For background,

interested readers are referred to relevant publications (e. g. Pevzner, Tang & Waterman 2001; Li *et al.* 2008).

Assembly strategy: In brief, the assembly strategy is to connect the beginnings and ends of the FLS reads with the kmer set (see Terminology in Data S3, Supporting Information) from the SLS library (Fig. 2). The 5′ ends of the FLS reads are defined as the start point, and the 3′ ends of the FLS reads are defined as the end point. Then, for each paired FLS read, the kmer set from *de Bruijn* graph is walked step by step from the start point to the end point to find potential assembly paths. Several assembly strategies are used to aid the search for the correct paths: (i) removing kmers whose abundance is <10% of the average kmer abundance before path bifurcation; (ii) if there is more than one *out degree* remaining after step one, the common reads between different out degrees and the kmer located just before the last bifurcation are counted, and the out degrees with common reads fewer than 10% of the average abundance of kmers before this bifurcation are removed; (iii) paths expanding beyond the pre-set length (i.e. 660 bp for our data, slightly longer than the COI standard barcode region) without end point are removed.

Removing errors via abundance estimation: The Burrows-Wheeler Aligner (Li & Durbin 2009) was applied to align shotgun reads to the full-length barcode assemblies at 100% match. The abundance of every nucleotide base was calculated by the number of successfully aligned shotgun reads. Regional abundance was calculated as the average abundance of all bases in that region. Assembled barcodes are considered errors and abandoned if any of its base-sites has an abundance value of 0 or if there is 10^3 times difference in regional abundances between the 100-bp region on the 5′ and that of the 3′.

Operational taxonomic units processing, chimera detection and taxonomy assignment

Assembled results were sorted by abundance and then clustered into operational taxonomic units (OTUs) using the *uclust* function in USEARCH (Edgar 2010) at a 98% similarity threshold. Chimeras were filtered out by applying *de novo* chimera detection using UCHIME (Edgar *et al.* 2011). All output full-length COI sequences were compared against the BARCODE OF LIFE DATA SYSTEMS version 3.0 (Ratnasingham & Hebert 2007) for taxonomic assignment, and non-arthropod OTUs (e.g. bacteria) were considered as non-target taxa and removed.

Ecological relatedness

To quantify the extent to which we successfully recovered the input sequences, we conducted a BLAST search for each OTU against the Sanger references making up the KMG and XSBN samples, respectively, using a similarity threshold of 98% of full-length alignment (Table 1, Yu *et al.* 2012). To detect potential cross-sample contamination, we also BLASTed OTUs against the Sanger sequences of other sampling sites (e.g. the KMG OTUs were BLASTed against HONGE and XSBN). Beta diversities were estimated using UniFrac (Lozupone *et al.* 2010) with a Neighbour-Joining tree (Data S2, Supporting Information) generated by PHYLIP (Felsenstein 2002). The dissimilarity matrix was visualized using principal coordinate analysis (PCoA, see Terminology in Data S3, Supporting Information).

Results

SOAPBarcode and OTU Processing

After adapter removal, we obtained *c.* 1·28 Gbp FLS raw data and 2·00 Gbp SLS raw data for XSBN, 1·87 Gbp FLS raw data and 1·84 Gbp SLS raw data for KMG (Table 2). A total of 1·10 Gbp and 1·33 Gbp clean data were generated after denoising the KMG and XSBN samples, respectively, including both insert-size libraries.

Using SOAPBarcode and the FLS and SLS libraries, we assembled 231 and 229 scaffolds for XSBN and KMG, respectively. OTU clustering at 98% threshold generated 204 and

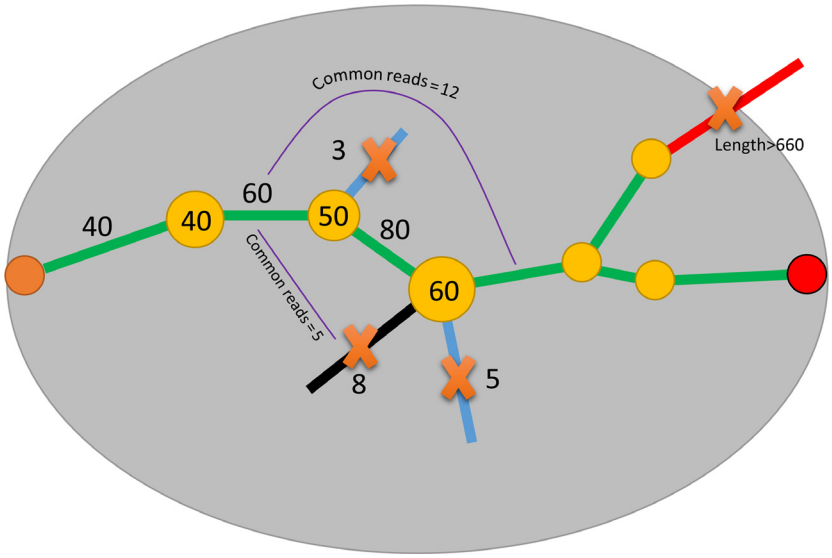

Fig. 2. Schematic pipeline of the assembly strategy of *SOAPBarcode*. Orange and red circles represent the start and end points, respectively, which are identified using the paired-end reads from the full-length barcode insert library. The green line represents accepted paths. The number in the yellow circle represents the average kmer abundance before this bifurcation. The number above the line represents its abundance. The blue line represents a path that has been abandoned because of low abundance (<10%). The red line represents a path that has been abandoned due to unacceptable length. The thin purple line represents the common reads shared between the targeted kmer and the kmer right before the last bifurcation. The black line represents a path that has been removed because its common reads support is <10%.

Table 2. Bioinformatics procedures of the PCR-based Illumina pipeline and corresponding data set sizes

	XSBN		KMG		
Procedures	FLS (full-length)	SLS (shotgun)	FLS (full-length)	SLS (shotgun)	Program[1]
Reads (raw data)	4 258 353 (1·28 Gbp)	6 666 667 (2·00 Gbp)	6 246 714 (1·87 Gbp)	6 123 509 (1·84 Gbp)	
Denoising (% of raw reads)	H: 977 738 (22·96)[2] P: 616 598 (14·48) S: 477 051 (11·20) P: 470 075 (11·04) C: 398 194 (9·35)	O: 5 918 088 (88·77) P: 4 030 776 (60·46)	H: 1 582 200 (25·33) P: 829 501 (13·28) S: 664 241 (10·63) P: 632 113 (10·12) C: 608 492 (9·74)	O: 5 942 940 (97·05) P: 3 795 119 (62·52)	SOAPBarcode
Full-length COI barcode assemblies	231		229		SOAPBarcode
OTU clustering[3]	204		207		USEARCH
Chimera detection	203		207		USEARCH
Removal of non-target COI	203		199		BOLD v3

SLS, shotgun library sequences; FLS, full-length library sequences; SLS, shotgun library sequences; OTU, Operational taxonomic units.
[1]All newly developed programs have been uploaded to: https://sourceforge.net/projects/metabarcoding/.
[2]Presented as analysis procedure: read numbers (proportion of raw reads after filtering). H: high-quality reads without low-quality base pair; P: primer match at 100% identity; S: singleton reads removal; E: protein hydrophily check; C: chimera removal; O: overlapped paired-end reads of SLS.
[3]OTUs were clustered using a similarity threshold of 98%.

207 OTUs for XSBN and KMG, respectively, of which only one sequence found in XSBN was identified as a chimera and deleted. Eight remaining OTU clusters were identified as bacteria and other non-arthropods (i.e. an earthworm) and deleted (Table 2). Finally, we obtained 203 and 199 OTUs for XSBN and KMG, respectively. The sequences and taxonomic assignments of these OTUs are summarized in Table S2 (Supporting Information).

OPERATIONAL TAXONOMIC UNITS RECOVERY AND TAXONOMIC ASSIGNMENTS

A total of 176 OTUs were recovered out of the 230 OTUs present in the Sanger XSBN data set (76·5%), while 149 out of 153 input OTUs were recovered for KMG (97·4 %). The overall recovery rate for true positives for both samples was thus 84·9%, which is higher than the 67·4% recovery rate for the KMG and XSBN samples using the 454 sequencer (Yu *et al.* 2012) (Table 3). We also calculated recovery rates for nine arthropod orders (Table 3). Interestingly, Hymenoptera, which was the order with the lowest recovery in the 454 study, at 45·6% for two samples, was recovered by the Illumina pipeline at a more satisfactory rate of 68·4%. This increase suggests that low Hymenoptera recovery in the 454 study was not only due to outright PCR failure but also due to some combination of too-low-sequencing coverage and, possibly, over-clustering in the 454 pipeline (Yu *et al.* 2012).

There were also fewer false-negative (dropout) and untraceable (novel) OTUs generated by our Illumina pipeline compared to the 454 pipeline (Fig. 3). However, cross-sample contamination was also observed: nine OTUs found in KMG came from XSBN, and 20 OTUS found in XSBN came from KMG (11) and HONGHE (9), respectively. Recall that HONGHE was PCR-amplified but not sequenced due to low DNA quality. Finally, as in the 454 data set, both Illumina samples contained 'novel OTUs' that were not present in any

Table 3. Comparison of taxon recovery rates between 454 method and Illumina SOAPBarcode method, subdivided by major taxa. The 454 rates reported here represent only those from two bulk samples, KMG and XSBN, and thus are subsets of those reported in Yu *et al.* (2012)

		Roche 454		Illumina SOAPBarcode	
Taxa (order)	Sanger Reference OTUs	Recovered OTUs	Recovery rate, %	Recovered OTUs	Recovery rate, %
Araneae	4	2	50	4	100
Blattaria	2	2	100	2	100
Coleoptera	24	17	70·8	21	87·5
Diptera	87	66	75·9	79	90·8
Ephemeroptera	1	1	100	1	100
Hemiptera	29	25	86·2	25	86·2
Hymenoptera	79	36	45·6	54	68·4
Lepidoptera	152	105	69·1	135	88·3
Psocoptera	5	4	80	4	80
Total	383	258	67·4	325	84·9

OTU, operational taxonomic units.

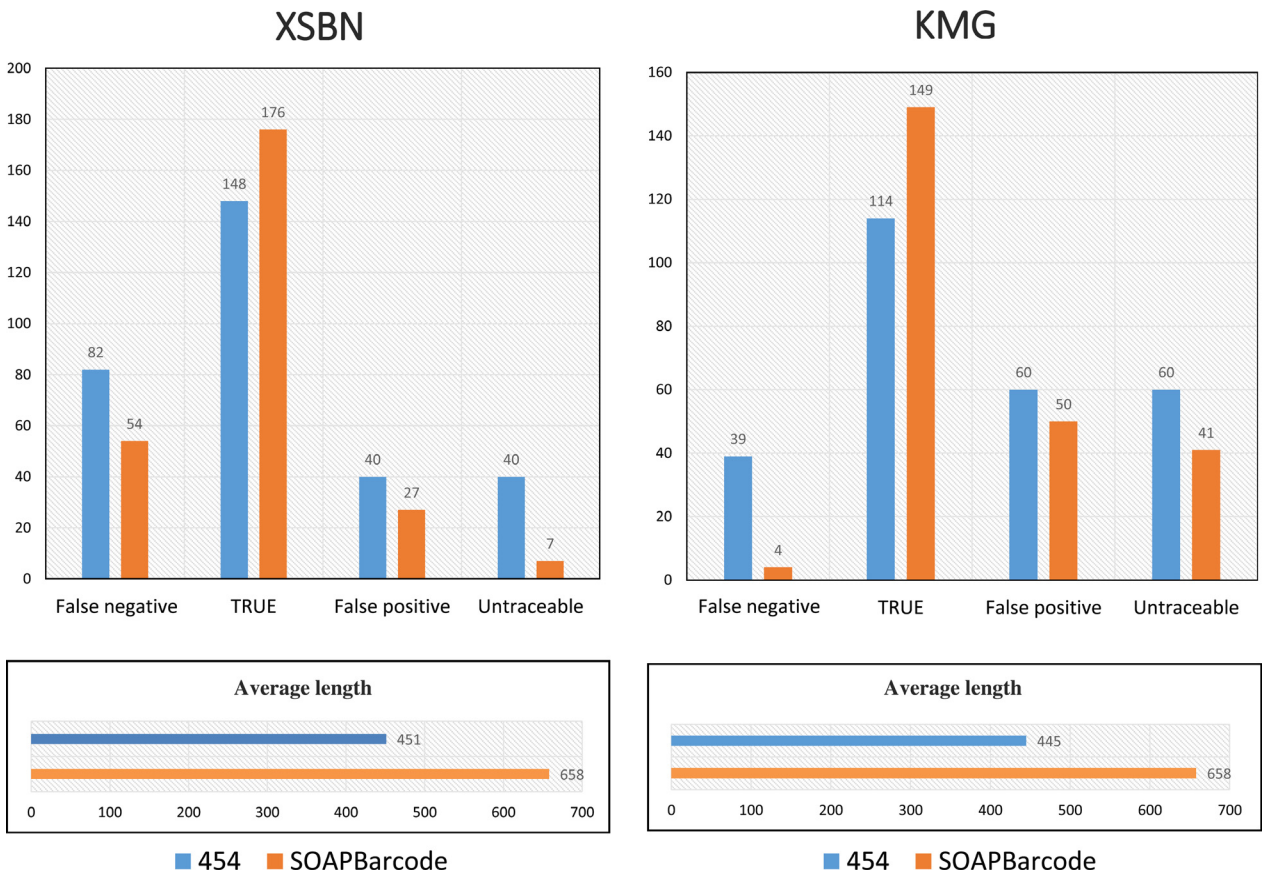

Fig. 3. OTU recoveries by Illumina and 454 approaches relative to the Sanger input data set. False negatives – Sanger operational taxonomic units (OTUs) absent in the next-generation sequencer (NGS) data set (=dropouts); True positives – NGS OTUs present in the Sanger input data base (=successful amplification and recovery); False positives – NGS OTUs absent in the corresponding Sanger reference data base (including cross-sample contamination and untraceable OTUs); Untraceables – NGS OTUs absent in all Sanger reference data bases, including those of other sampling sites (=novel OTUs).

of the Sanger data sets. These results are compared with the 454 results in Fig. 3.

Importantly, our Illumina pipeline generated full-length COI barcodes (658 bp), with an average length >200-bp longer than the 454 OTUs, producing more resolved taxonomic assignments. As shown by Yu *et al.* (2012), full-length COI Sanger barcodes result in almost twice the taxonomic assignment success, using SAP (Munch *et al.* 2008), at family, genus and species levels (*c.* 35–36%), relative to the shorter 454 sequences (*c.* 16–17%) (Table S3, Supporting Information).

We compared taxonomic assignment resolution between our assembled full-length COI barcodes and the same barcodes truncated at 150 bp from the 5′ end to represent mini-barcodes. One hundred and fifty base pair is also the HiSeq 2000's single-end read length. The BARCODE OF LIFE DATA SYSTEMS version 3.0 (Ratnasingham & Hebert 2007) data base was used to assign taxonomy, and we counted the proportion of barcodes that found a match at ≥97% similarity, which is taken as a match to genus level (following Ratnasingham & Hebert 2007). Ninety-seven of 409 (26%) of taxonomic assignments were inconsistent between the full-length and the mini-barcodes, and, perhaps surprisingly, it was the *mini-barcodes* that returned more successful assignments at ≥97% similarity (mini: 193/409, 47%; full-length: 112/409, 27%) (Table S4, Supporting Information). We hypothesized that the greater proportion of positive assignments of the mini-barcodes were false positives because of the following: (i) the accuracy of the full-length OTUs had already been verified against the reference Sanger OTUs; (ii) many of these OTUs are not expected to have BOLD match due to the fact that this area of the world has not yet been well-covered by barcode campaigns. To test this hypothesis, we scored the geographic distributions of the *assigned* species. We divided the assigned species into those with distributions in Asia ('Asia-wide') and those with distributions outside Asia ('Non-Asian') (e.g. Africa, Europe, America, Australia, …), using BOLD-provided biogeographic information for the top-hit assignments. More of the mini-barcodes were assigned to non-Asian species (109/193), which are presumably false positives, than were full-barcode sequences (39/112). This difference is statistically significant ($\chi^2 = 13{\cdot}31$, d.f. = 1, $p < 0{\cdot}0003$). In short, this analysis suggests full barcodes are more likely to return reliable taxonomic assignments, at least on BOLD.

COMMUNITY DISSIMILARITIES

We used Unifrac to estimate pairwise community dissimilarities among the three samples (HONGHE, XSBN and KMG)

analysed by three sequencing technologies (Sanger, 454 and Illumina), for a total of (3 + 3 + 2=) 8 communities (see Table S5, Supporting Information). Principal coordinates analysis clusters the two Illumina data sets near their corresponding Sanger and 454 communities, despite the presence of novel OTUs, and does not place the Illumina data sets near the HONGHE communities, despite cross-contamination with some HONGHE sequences (Fig. 4). Thus, despite the presence of novel and cross-contaminant sequences, Illumina data sets contain sufficient information such that Unifrac analysis is able to extract the correct ecological relationships among communities.

Discussion

ILLUMINA PIPELINE ENABLES BETTER SPECIES RECOVERY AT LOWER COST

The PCR-based Illumina pipeline using SOAPBarcode can recover whole barcode sequences from amplicons of bulk arthropod samples. Just *c.* 1·1% of the full sequencing capacity of a single Illumina HiSeq 2000 run (600 Gbp per run) allows recovery of 84·9% of the input OTUs in two bulk samples (Fig. 3), which is higher than the species recovery rate of using a previously published Roche 454 pipeline (Yu *et al.* 2012). This higher recovery rate of the SOAPBarcode pipeline, especially for Hymenoptera, combined with full-barcode sequences, allows for accurate recovery of accurate community composition information (Fig. 4) and fewer false-positive taxonomic assignments (Tables S3 and S4, Supporting Information).

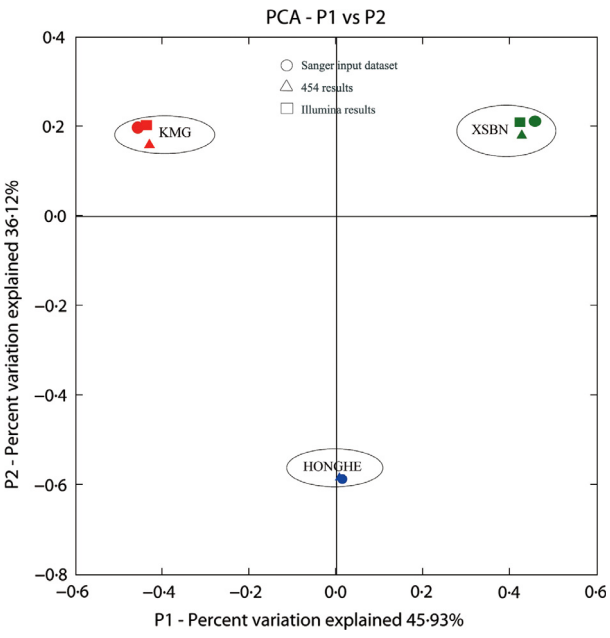

Fig. 4. Principal coordinates analysis (PCoA) of beta diversities evaluated by the dissimilarities between samples using Unifrac. The KMG and XSBN communities analysed by the three methods cluster together, and the Illumina communities are located slightly closer to the Sanger reference communities, reflecting the closer taxonomic and sequence match of the Illumina and Sanger data sets (Fig. 3).

The improved recovery rate for true positives can be attributed to Illumina's high-throughput sequencing capacity. In other work, we have demonstrated that the deep sequencing capacity of the Illumina HiSeq 2000 allows the detection of insect specimens that are represented by only minute quantities of mitochondrial sequences (Zhou *et al.* 2013). This increased sensitivity suggests that Illumina sequencing can partially alleviate taxonomic biases in metabarcoding caused by either low individual abundance or inefficient PCR primer performance. To show this, we divide true positive OTUs that were detected by the Illumina pipeline into two categories: those also detected by 454 (Yu *et al.* 2012) and those only detected by Illumina. Not surprisingly, the read abundances of the Illumina-*only* OTUs are significantly lower than those of the OTUs also detected by the 454 (Wilcoxon rank test, $p < 0{\cdot}01$; Table S6, Supporting Information). In fact, the higher data volume per taxon unit in the KMG sample, compared to XSBN, might explain the difference in OTU recovery rates (97·4% and 75·2%, respectively); the KMG sample produced 3·71 Gbp raw data (including both insert-size libraries) for 153 species compared to 3·28 Gbp raw data for 230 species in XSBN. The raw data were also aligned to the reference Sanger sequences to trace the 'culprit OTUs' causing dropout. Only one reference sequence could not be matched by any raw data, which means that all the reference templates generated at least some PCR product and raw data. However, the dropout OTUs generated too little read data to achieve adequate coverage for assembly. Low read numbers can be due to taxonomic bias in the primer, which is suggested by the lower recovery rate for Hymenoptera.

The Illumina OTU data sets also result in ever-so-slightly more accurate Unifrac beta diversity estimates, since the two Illumina communities clustered more closely to the Sanger reference data sets (Fig. 4). On the other hand, the downside of this higher sensitivity remains the detection of cross-sample contaminants that originate from sample collection, transportation or laboratory processing.

A precise cost comparison between the HiSeq 2000 and 454 protocols is not possible because laboratory work and informatics analysis were carried out at separate institutes and because it is difficult to parse out one-time costs of protocol development. However, a rough estimate based on read volume and average cost per read in both platforms suggests that the overall sequencing cost in the Illumina protocol is *c.* 1/2 that of the 454 (this study and Yu *et al.* 2012). In fact, a preliminary test has found that just 200 Mbp FLS and 200 Mbp SLS can recover 128/153 (83·7%) of the OTUs in the KMG data set, which promises even lower sequencing costs after further optimization of the pipeline (S. Liu, unpublished data). Although the current SOAPBarcode pipeline is already affordable, we continue to optimize protocols to further reduce the overall cost and increase efficiency. For instance, we are investigating the possibility to incorporate the barcoding multiplex identifier (MID) tag into the current pipeline. As MID tags will be added to each of the primer sets for various bulk samples,

which are subsequently pooled and sequenced in the same Illumina lane, multiple bulk samples can be analysed simultaneously and then be sorted out respectively.

ILLUMINA METABARCODING OF STANDARD FULL-LENGTH BARCODES

Given the high data output of Illumina sequencing, an increasing amount of work has been carried out on this platform. To date, most NGS biodiversity analyses involving Illumina sequencing and PCR amplifications have been focused on degraded DNA using short sequence markers, for example fragments of the mitochondrial 12S and 16S rRNA genes (Shehzad *et al*. 2012a) and the P6 loop of the plastid DNA *trnL* intron (Baamrane *et al*. 2012), all of which are typically <100 bp. The preference for a short, non-standard DNA marker lies in its improved primer universality (Coissac, Riaz & Puillandre 2012) and the fact that these markers can be easily sequenced using current Illumina technology. For instance, there have been diet analysis of herbivores (Baamrane *et al*. 2012) and carnivores (Shehzad *et al*. 2012a,b) via faecal samples, and similar strategies have been applied on soil samples (Bienert *et al*. 2012; Yoccoz *et al*. 2012). A drawback of these short markers, however, is that they do not take advantage of the large, high-quality and growing data base of standard DNA barcode markers (ibol.org; www.barcodeoflife.org). With the rapid developments in NGS technologies and methodological innovations in constructing reference data bases for mitochondrial genomes (Timmermans *et al*. 2010; Taberlet *et al*. 2012a; Zhou *et al*. 2013), it is possible now to marry the classic DNA barcoding concept with genomics methods. However, as shown in Zhou *et al*. (2013), the application of genomics methods for biodiversity assessment still poses logistical obstacles, such as (i) DNA quality criteria are difficult to meet for typical field samples, and (ii) sequencing data requirements are ten times greater than that in PCR amplicon studies, not to mention the massive computational resources required in genomics-related analyses. An Illumina approach focusing on standard barcode markers will not only take advantage of existing barcode references but also bridge the gap between current and future barcoding methods.

Our study is the first to test the utility of the Illumina HiSeq 2000 platform for analysing biodiversity via the standard animal barcode gene amplicons (658 bp). Our results showed that the HiSeq 2000 coupled with the SOAPBarcode pipeline can recover more accurate taxonomic and ecological information by allowing the construction of scaffolds of full-length COI barcodes. It is also worth noting that this new pipeline will likely be helpful for Illumina sequencing of other standard barcode markers, such as those for plants and fungi.

The Illumina MiSeq platform, which generates less data at higher per-base-pair costs (compared to HiSeq 2000), but has the advantages of smaller run sizes and (increasingly) longer read lengths (currently 2 × 250 bp), could also be used to generate longer FLS reads, where gaps to be filled with SLS can be much shortened, making it easier to be assembled by SOAPBarcode. It is also possible that MiSeq will eventually be improved to the point that it can generate such a long read lengths that full-length COI amplicons can be sequenced directly. When this desired transition happens, it depends on the extent to which sequence quality is maintained at those longer read lengths.

MODIFICATIONS IN SAMPLE COLLECTION, PREPARATION AND PROCESSING ARE NEEDED TO AVOID CONTAMINATION

Although the new Illumina protocols produce improved results, several common issues remain in existing NGS approaches (reviewed in Yu *et al*. 2012), for example PCR bias and false-positive OTUs. Additionally, cross-sample contamination seems to be a bigger issue compared to the 454, probably due to the increased sensitivity of the HiSeq 2000. Multiple causes could have contributed to the introduction of novel OTUs, for example, sharing of collecting equipment between sampling, ambient DNA captured in chemical preservatives (Shokralla *et al*. 2012), failure of detection and separation of small specimens attached to those used in reference construction via Sanger sequencing, detection of food items or parasitoids, escape of arthropods from the collecting trap while leaving residual tissues or eggs. The improved sensitivity in NGS technologies has enabled detection of ambient DNA that has been ignored by conventional molecular methods. However, in order to achieve a consistent NGS methodology that is comparable to classic morphology-based biodiversity assessment approach, modification on current protocols in sample collection, preservation, transportation and laboratory procedures seems necessary. For instance, collecting jars used in Malaise traps should not be shared in between sampling sites; dip nets used in benthic sampling should be decontaminated after each use. Similarly, new laboratory protocols, such as mixing tissue in sealed chamber and decontaminating laboratory ware between samples, will reduce chances for cross-sample contamination.

In summary, the Illumina HiSeq 2000 coupled with SOAPBarcode pipeline can achieve a high recovery rate and assemble full COI barcodes and, consequently, deliver reliable and taxonomically informative metabarcoding outcomes for environmental bulk samples. The high-throughput sequencing capacity of the Illumina platform enables the detection of NGS reads for nearly all reference taxa from the mixed arthropod samples. This offers a significant potential to alleviate taxon dropout for metabarcoding analysis. Of course, further endeavour (laboratory and bioinformatics) will need to balance the desire to reveal all real taxa and to avoid false positives.

Acknowledgements

This study is supported by the National High-tech Research and Development Project (863) of China (2012AA021601) and by BGI. YQJ and DWY were supported by Yunnan Province (20080A001), the Chinese Academy of Sciences (0902281081, KSCX2-YW-Z-1027), the National Natural Science Foundation of China (31170498), the Ministry of Science and Technology of China (2012FY110800), the University of East Anglia, and the State Key Laboratory of Genetic Resources and Evolution at the Kunming Institute of Zoology.

Authors' contribution

XZ, SL, and DWY designed the experiments and wrote the manuscript; JL and SL developed SOAPBarcode; SL, YL, JL and MT analysed the data; QY, XS, CZ, and LZ performed experiments; RZ edited the manuscript; YJ and DWY collected samples and provided data.

Conflict of interest

The authors declare that they have no competing interests.

Reference

Baamrane, M.A.A., Shehzad, W., Ouhammou, A., Abbad, A., Naimi, M., Coissac, E., Taberlet, P. & Znari, M. (2012) Assessment of the food habits of the Moroccan Dorcas Gazelle in M'Sabih Talaa, West Central Morocco, using the trnL approach. *PLoS ONE*, **7**, e35643.

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

Bienert, F., Dedanieli, S., Miquel, C., Coissac, E., Poillot, C., Brun, J.J. & Taberlet, P. (2012) Tracking earthworm communities from soil DNA. *Molecular Ecology*, **21**, 2017–2030.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R. & Thomas, W.K. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Coissac, E., Riaz, T. & Puillandre, N. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Creer, S., Fonseca, V.G., Porazinska, D.L., Giblin-Davis, R.M., Sung, W., Power, D.M. *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Felsenstein, J. (2002) {*PHYLIP*}*(Phylogeny Inference Package) Version 3.6 a3*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Hao, X., Jiang, R. & Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, **27**, 611–618.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, R., Li, Y., Kristiansen, K. & Wang, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Liu, B., Yuan, J., Yiu, S.M., Li, Z., Xie, Y., Chen, Y. *et al.* (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, **28**, 2870–2874.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J. & Knight, R. (2010) UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal*, **5**, 169–172.

Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E. & Nielsen, R. (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, **57**, 750–757.

Pevzner, P.A., Tang, H. & Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the USA*, **98**, 9748–9753.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. & Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: the Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Shehzad, W., McCarthy, T.M., Pompanon, F., Purevjav, L., Coissac, E., Riaz, T. & Taberlet, P. (2012a) Prey preference of snow leopard (*Panthera uncia*) in South Gobi, Mongolia. *PLoS ONE*, **7**, e32104.

Shehzad, W., Riaz, T., Nawaz, M.A., Miquel, C., Poillot, C., Shan, S.A., Pompanon, F., Coissac, E. & Taberlet, P. (2012b) Carnivore diet analysis based on next generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Molecular Ecology*, **21**, 1951–1965.

Shokralla, S., Spall, J.L., Gibson, J.F. & Hajibabaei, M. (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, **21**, 1794–1805.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012a) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Taberlet, P., Prud'homme, S.M., Campione, E., Roy, J., Miquel, C., Shehzad, W. *et al.* (2012b) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, **21**, 1816–1820.

Timmermans, M.J., Dodsworth, S., Culverwell, C., Bocak, L., Ahrens, D., Littlewood, D.T., Pons, J. & Vogler, A.P. (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, **38**, e197–e197.

Yoccoz, N.G., Bråthen, K.A., Gielly, L., Haile, J., Edwards, M.E., Goslar, T. *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y. & Ding, Z.L. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhou, X., Li, Y., Liu, S., Su, X., Yang, Q., Zhou, L. *et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

Received 9 July 2013; accepted 19 September 2013
Handling Editor: David Orme

Supporting Information

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

Data S1. User's manual of SOAPBarcode pipeline.

Table S1. Taxonomic information of all COI sequences used in Pro_C.

Table S2. Taxonomic information of different OTUs.

Table S3. Taxonomic assignment of arthropod OTUs to four lower taxonomic levels using SAP.

Table S4. The taxonomic assignment and the geographic distribution of the full-length COI barcodes and the 5′ mini-barcode region.

Table S5. Dissimilarity matrix from different sequencing platforms for all sample sites.

Table S6. Wilcoxon rank test of the abundance of two OTU categories.

Fig. S1. The hydrophily interval of all 302 476 COI genes.

Data S2. The Newick format Neighbour-Joining tree built from Sanger, 454 and Illumina data.

Data S3. Terminology.