SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

January 2013 Vol.56 No.1: 73–81
doi: 10.1007/s11427-012-4423-7

Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity

YANG ChenXue[1], JI YingQiu[1], WANG XiaoYang[1], YANG ChunYang[1] & YU Douglas W.[1,2*]

[1]*Ecology, Conservation, and Environment Center, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;*
[2]*School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ, UK*

Received October 23, 2012; accepted November 29, 2012; published online December 20, 2012

A number of basic and applied questions in ecology and environmental management require the characterization of soil and leaf litter faunal diversity. Recent advances in high-throughput sequencing of barcode-gene amplicons ('metabarcoding') have made it possible to survey biodiversity in a robust and efficient way. However, one obstacle to the widespread adoption of this technique is the need to choose amongst many candidates for bioinformatic processing of the raw sequencing data. We compare three candidate pipelines for the processing of 18S small subunit rDNA metabarcode data from solid substrates: (i) USEARCH/CROP, (ii) Denoiser/UCLUST, and (iii) OCTUPUS. The three pipelines produced reassuringly similar and highly correlated assessments of community composition that are dominated by taxa known to characterize the sampled environments. However, OCTUPUS appears to inflate phylogenetic diversity, because of higher sequence noise. We therefore recommend either the USEARCH/CROP or Denoiser/UCLUST pipelines, both of which can be run within the QIIME (Quantitative Insights Into Microbial Ecology) environment.

454 Genome Sequencer FLX System, DNA barcoding, high-throughput sequencing, soil fauna, metabarcoding, 18S rDNA

Citation: Yang C X, Ji Y Q, Wang X Y, et al. Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. Sci China Life Sci, 2013, 56: 73–81, doi: 10.1007/s11427-012-4423-7

Our goal in this study was to selectively amplify, sequence, and assign taxonomies to metazoan barcode genes in order to characterize faunal communities in soil and leaf litter habitats. In the process, we compare two DNA extraction methods (bead beating and liquid nitrogen), two barcoding markers (18S rDNA and CO1 mtDNA), three bioinformatic pipelines (USEARCH [1]+CROP [2], Denoiser [3]+ UCLUST [1], and OCTUPUS [4]), and two taxonomic assignment methods (BLAST searches against the small subunit (SSU) rDNA seed of the SILVA database release 108 [5], and the program SAP [6]). The first two pipelines can be run (mostly) within the QIIME environment (Quantitative Insights Into Microbial Ecology; qiime.org, accessed 12 Nov. 2012) [7]. Our primary focus is on comparison of the bioinformatic pipelines, as other researchers [4,8,9] have amply demonstrated the utility of 18S rDNA amplification from solid substrates for assessing metazoan diversity.

We chose these three particular pipelines because they are the ones most widely used in recent metabarcoding research, and they are all actively maintained [4,10,17–19,20]. Ideally, we would also have assessed the OBITOOLS pipeline [21], which is the most widely used pipeline for 18S amplicons [22,23]. However, while the software is readily available (www.grenoble.prabi.fr/trac/OBITools/wiki, accessed 12 May 2012), there is no instruction manual and the system is thus not straightforward to use.

*Corresponding author (email: dougwyu@gmail.com)

Our major findings were that (i) bead beating is superior to liquid nitrogen as a DNA extraction method for soil samples, but cannot be used for leaf litter; (ii) the 18S primers selectively amplify a wide range of metazoans, whilst CO1 primers mainly amplify bacteria; and most importantly, (iii) the three pipelines and two taxonomic assignment methods produce similar results as measured by taxonomic profiles and community dissimilarities. However, the OCTUPUS pipeline appears to overestimate phylogenetic diversity (PD) because it lacks a 'denoising' step. These results corroborate a series of recent papers [4,10–16] which collectively demonstrate that metabarcoding can be used to survey soil biodiversity quickly, comprehensively, and robustly.

1 Materials and methods

1.1 Sample collection and preparation

We collected topsoil from the lawn of the Kunming Institute of Zoology, Kunming, Yunnan, China, and leaf litter from a tropical forest reserve in Meng Lun, Xishuangbanna, Yunnan. The samples were immersed in a 10× volume of 100% ethanol and stored at 4°C until ready for processing. No attempt was made to characterize the biodiversity of these samples using traditional methods, as we are interested only in the differences among the protocols that we test in this study.

1.1.1 Soil fauna filtering

Following Creer et al. [24], 400 g dry soil was washed with distilled water through a 1 mm cylindrical sieve (to remove large particles) nested above a 63 μm sieve. Inspection by stereoscope showed that the retained fraction (280 g wet weight) contained soil particles. We washed this retained fraction with 75 mL distilled water into a graduated cylinder, added water to 500 mL, inverted 4–7 times, allowed sand and clay particles to settle for 20 s, then decanted the supernatant through a 63-μm sieve. This was repeated 4–5 times, which removed most but not all soil particles, as revealed by stereoscopic inspection. We further cleaned the sample in a 3:2 (v/v) ratio of water:Ludox solution (Sigma-Aldrich, St. Louis, MO, USA; specific gravity=1.13–1.18) [25], decanting twice, each time vigorously mixing the solution in a beaker, pouring it into a graduated cylinder, allowing it to settle for 40 min, then pouring the supernatant through the 63-μm sieve. The final retained fraction, looking like wet, black cotton, was divided into roughly equal halves of approximately 15 g. One half was subjected to tissue homogenization via bead beating for 20 s in a Qiagen FastPrep®-24 tissue homogenizer (Qiagen GmbH, Hilden, Germany), using steel beads at 45 m s^{-1}. The other half was hand-homogenized with liquid nitrogen using a mortar and pestle.

1.1.2 Leaf litter fauna filtering

Because leaf litter is such a low-density substrate, mass decantation cannot be used to separate leaf fragments from fauna, nor was it feasible to use bead-beating to homogenize the samples. Thus, 150 g leaf litter was sifted through a 1 mm sieve, and the resulting fine mixture was hand-homogenized with liquid nitrogen using a mortar and pestle.

1.2 DNA extraction, PCR, and pyrosequencing

We used the PowerSoil® DNA isolation kit (MO BIO Laboratories, Inc., Carlsbad, CA, USA) to extract DNA from 0.25 g of soil and leaf litter samples. We created five samples: two replicates of leaf litter homogenized with liquid nitrogen, two replicates of soil homogenized with bead beating, and one replicate of soil homogenized with liquid nitrogen (one sample failed during DNA extraction).

Following Hamilton et al. [8] and Wu et al. [9], we amplified an approximately 830-bp segment of the small subunit (SSU) 18S rDNA gene, with the metazoan-specific forward primer 18S11b (5′-GTCAGAGGTTCGAAGGCG-3′) and the universal eukaryotic 18S2a reverse primer (5′-GATCCTTCCGCAGGTTCACC-3′). To allow multiplexing during pyrosequencing, the forward primer was prefixed by a unique 10 bp sequence for each sample, known as a multiplex identifier (MID), and the standard A-adaptor for pyrosequencing. For each sample, separate amplifications were carried out for the CO1 gene using Folmer et al.'s [26] CO1 primers LCO1490 and HCO219.

PCRs were carried out in 10 μL reaction volumes containing 0.8 μL dNTP mixture (1.25 mmol L^{-1} each base), 6.05 μL distilled water, 0.05 μL Taq DNA polymerase (Takara Biosystems, Dalian, China), 1.0 μL 10× PCR buffer (100 mmol L^{-1} Tris-HCl (pH 8.3), 500 mmol L^{-1} KCl, 15 mmol L^{-1} $MgCl_2$), 0.2 μL each primer (20 μmol L^{-1}), 0.5 μL DMSO, 0.2 μL BSA and 1.0 μL DNA template. We used non-proofreading Taq and fewer, longer cycles to reduce chimera production [20]. Thermal cycling conditions for amplification of 18S included 2 min at 95°C, followed by 30 cycles of 15 s at 95°C, 30 s at 57°C and 3 min at 72°C, and a final elongation stage of 10 min at 72°C. For amplification of the CO1 gene, we used 3 min at 95°C, followed by 30 cycles of 30 s at 94°C, 30 s at 49°C and 1 min at 72°C, and a final elongation stage of 5 min at 72°C. Each sample was amplified three times independently, and the products were pooled for sequencing (a total of 30 μL per sample).

For pyrosequencing, all PCR products were gel-purified using a QIAquick PCR Purification Kit (Qiagen), quantified by using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA, USA), pooled and A-amplicon-sequenced on a Roche GS FLX '454' System (454 Life Sciences, Branford, CT, USA). The two pooled amplicons (18S and CO1) were sequenced on separate 1/8 regions of a plate.

1.3 Recovery of operational taxonomic units (OTUs) and taxonomies

We tested three processing pipelines for sequence quality control, denoising, chimera removal, and sequence clustering into OTUs. Example command scripts are provided as Supporting Information. For the first two pipelines, we used QIIME scripts to perform initial quality control and to remove primer and MID sequences from the 454 reads. Reads were discarded if they were shorter than 100 bp, longer than 700 bp, or had more than two primer errors and/or a homopolymer run of more than six nucleotides. Thus, the 97839 original reads were cut to 63185.

1.3.1 Pipeline 1: USEARCH/CROP

The USEARCH program [1] provides a very fast pipeline for sequence denoising and chimera detection (note that we use the multiplex identifier (MID) to distinguish the five separate samples). Sequences were denoised in this system by clustering reads at 99% similarity and constructing a consensus sequence for each cluster by a majority-nucleotide-or-gap rule. Subsequently, *de novo* and reference-based chimera detection and deletion were conducted with the UCHIME function in USEARCH [27]. The QIIME pipeline [7] recently incorporated the USEARCH functions, using QIIME scripts to keep track of which sequences belong to which MID.

We then used a two-step protocol to select OTUs (i.e., to reduce the dataset to one representative sequence per OTU). USEARCH already clusters at 99% similarity, reducing the workload for the CROP program [2] which was used to cluster sequences at 96% similarity following Fonseca et al. [4]. CROP applies Bayesian clustering methods to find clusters "based on the natural organization of data without setting a hard cut-off threshold" [2]. CROP can be slow, but we have found that it produces five to ten times fewer OTUs than do programs like Cd-hit [28] and UCLUST [1,29]. CROP outputs a representative sequence for each OTU, of which we kept all sequences ⩾100 bp.

1.3.2 Pipeline 2: Denoiser/UCLUST

The QIIME pipeline [7] includes a standard system for processing 18S amplicons that uses the denoiser algorithm [3] to remove noise from sequences. Operating at a rate of ~1500 sequences per hour, denoiser is slower than USEARCH but is one of the most widely used denoising programs for 454 reads. The UCLUST algorithm [1] was then used at 96% similarity to select OTUs. The longest sequence within each OTU cluster was chosen as the representative sequence. Finally, the blast_fragments option [30] in QIIME was used to detect and delete chimeric OTUs, after aligning the OTUs against the SSU rDNA seed of the SILVA reference database release 108 [5].

1.3.3 Taxonomic assignment for USEARCH/CROP and Denoiser/UCLUST pipelines

We tested two contrasting methods to assign taxonomies to OTUs. The first involves the program SAP [6], which uses Markov chain Monte Carlo (MCMC) to sample 10000 unrooted phylogenetic trees constructed with the query OTU and its GenBank homologues. SAP is slow (~3 sequences/CPU-hour) but provides posterior assignment probabilities to different taxonomic levels (e.g., a sequence might be assigned with 97% probability to the spider family Araneae, but at <90% probability to any particular spider family). The second method is to match OTUs against a curated reference dataset. Here, we BLAST at a stringency of 1×10^{-3} against a 97%-similarity UCLUST-clustered version of the SILVA SSU rDNA database release 108 [5], available at www.arb-silva.de/download/archive/qiime/ (accessed 11 March 2012). Our question is whether the second, much faster method can assign taxonomies similar to those produced by SAP, which we consider to be more reliable.

1.3.4 Pipeline 3: OCTUPUS

The third pipeline (octupus.sourceforge.net, accessed 11 March 2012) was developed to process a metagenetic dataset of marine benthic meiofauna [4]. In short, the program LUCY [31] was used to conduct initial quality filtering, and MEGABLAST (www.ncbi.nlm.nih.gov/staff/tao/URLAPI/megablast.html, accessed 11 March 2012) and MUSCLE [32] were used to select OTUs at 96% similarity. MEGABLAST against a local copy of the GenBank nucleotide database was used to detect chimeras under the assumption that only the 5′ or 3′ end of a chimeric sequence will match any given reference sequence, and to assign taxonomies according to the OCTUPUS pipeline.

1.4 Pipeline comparisons

1.4.1 Alpha diversity

The number of sequence reads per OTU is reported to correlate with true OTU abundance and biomass in samples of nematodes [14], but is probably correlated only weakly (if at all) with taxonomically diverse samples [29,33]. Thus, abundance-coverage estimators of alpha diversity [34] among MIDs are probably unreliable. We therefore used the R function phylocurve.R [35,36] to rarefy phylogenetic diversity (total branch length) over the number of OTUs. For each MID, we constructed a rooted maximum likelihood tree (based on the Kimura 2-parameter genetic distance model) of sequences representing the various OTUs. The trees were rooted with an *Onychophora* sequence from the SILVA database. Highly variable positions (entropy >75%) were masked, and the tree was built using the PhyML plugin v. 2.0.12 [37] in Geneious v. 5.6.4 [38].

1.4.2 Beta diversity

For each of the three pipelines, we used QIIME scripts to

generate a dissimilarity matrix using the 1-Sørensen-Dice index on a table of OTUs×MID. We visualized the dissimilarity matrix with a Principal Coordinates Analysis (PCoA) and used a Procrustes analysis to test for pairwise community correlations among the three pipelines.

2 Results

2.1 Sequencing output, denoising, and chimera detection

2.1.1 18S

Pyrosequencing produced a total of 97839 reads. Of these, 63185 reads with an average length of 270 bp were retained after initial quality control, before processing by the USEARCH/CROP and Denoiser/UCLUST pipelines. After denoising and chimera detection, the USEARCH/CROP pipeline produced 9995 reads (317 chimeric reads detected), whilst the Denoiser/UCLUST pipeline produced 49494 reads (14332 chimeric reads detected). The former pipeline denoises by clustering and generating consensus sequences, so this pipeline outputs fewer reads and fewer chimeras than the latter, before the OTU picking stage. On a 2010 iMac (processor: 2.66 GHz Intel Core i5; memory: 8 GB 1067 MHz DDR3; software: Mac OS X Lion 10.7.5 (11G63)), both pipelines required 1–2 h of processing to reach this stage.

The OCTUPUS pipeline required approximately 20 min of processing on the same iMac system. This method retained 75195 sequences at an average read length of 277 bp after initial quality control, and retained 445 sequences after OTU picking and chimera detection. Note that in OCTUPUS it is not straightforward to determine the time spent at each stage of the pipeline.

2.1.2 CO1

Pyrosequencing produced a total of 74884 reads. Of these, 52808 reads with an average length of 285 bp were retained after initial quality control in the Denoiser/UCLUST pipeline.

2.2 OTU picking

Using different algorithms or different similarity thresholds can result in quite different estimates of OTU richness. For each pipeline, we trialed three thresholds on the 18S dataset (over each of the five MIDs): 96%, 97%, and 99%. For these three respective values, the USEARCH/CROP pipeline generated 419, 528, and 1177 OTUs, the Denoiser/UCLUST pipeline generated 671, 822, and 1242 OTUs, and OCTUPUS generated 445, 1000, and 14000 OTUs (Table 1). Fonseca et al. [4] recommended a 96% similarity cut-off for their benthic meiofauna dataset, and seeing that this threshold did not greatly reduce OTU richness relative to the 97% cut-off in our dataset, we also used a value of 96% for downstream analyses. Running on the 2010 iMac system

Table 1 Taxonomic assignment of OTUs by each pipeline, using the SILVA SSU rDNA database[a)]

	Taxa	Total OTUs	Eukaryota	Metazoa	Arthropoda				Nematoda	Annelida
18S	USEARCH/CROP	419	224 (53.5% of total)	222 (53.0% of total)	159 (71.6% of Metazoa)				42 (18.9% of Metazoa)	7 (3.2% of Metazoa)
					Arachnida 30.6%	Hexapoda 27.5%		Myriapoda 7.2%		
						Insecta 12.6%	Collembola 13.1%			
	Denoiser/UCLUST	671	458 (68.3% of total)	455 (67.8% of total)	359 (78.9% of Metazoa)				65 (14.3% of Metazoa)	14 (3.1% of Metazoa)
					Arachnida 35.6%	Hexapoda 30.1%		Myriapoda 8.6%		
						Insecta 10.8%	Collembola 16.9%			
	OCTUPUS	445	141 (31.7% of total)	141 (31.7% of total)	121 (63.8% of Metazoa)				9 (6.4% of Metazoa)	20 (14.2% of Metazoa)
					Arachnida 43.3%	Hexapoda 29.8%		Myriapoda 5.0%		
						Insecta 14.2%	Collembola 12.8%			
	CO1	10908	105 (0.96% of total)							

a) Less than 1% of CO1 OTUs were assigned to the Metazoa, compared to over 30% of 18S OTUs. Within the Arthropoda, the three pipelines generally assigned similar proportions of OTUs to Arachnida, Insecta, Collembola, and Myriapoda, though OCTUPUS produced a higher frequency of Arachnid OTUs. Percentages do not add to 100 because we omit some small taxonomic groups.

described above, CROP required around 10 min to process 10000 18S sequences.

For the CO1 datasets, we also used a 96% threshold and generated 10908 OTUs over the five MIDs using the USEARCH/CROP method. This is the pipeline that produced the lowest number of 18S OTUs (Table 1) and is essentially the same as the pipeline used successfully for CO1 amplicons from malaise trap samples [29].

2.3 Taxonomic assignment

For the CO1 OTUs, SAP assigned only 105 of 10908 OTUs (0.96%) to Eukaryota (Table 1). Clearly, generic CO1 primers are not suitable for material that has a high soil or leaf litter component, and we therefore do not consider this dataset further.

In contrast, across the three pipelines, 31.7%–67.8% of the 18S OTUs were assigned to Metazoa, and 63.8%–78.9% of the metazoan OTUs were assigned to Arthropoda (Table 1). Importantly, only a small proportion of these arthropod OTUs was assigned to Insecta (Table 1), with most of the rest allocated to Arachnida, Collembola, and Myriapoda, which dominate soil and leaf litter fauna [39].

2.3.1 Taxonomic reliability of SILVA assignments

All sequences in the SILVA 108 database are assigned a species identity, meaning that all matching OTUs are also automatically assigned a species-level taxonomic identity. How reliable are assignments at lower taxonomic levels? For the 455 metazoan OTU sequences from the Denoiser/UCLUST pipeline (Table 1), we compared the SILVA taxonomies with those assigned by SAP [6], which can assign posterior probabilities at each taxonomic level (here, we allow only assignments ⩾80% probability). Both methods assigned almost all sequences to order, but the SAP assignment success declined to between 7% and 50% at lower taxonomic levels (Table 2), comparable with levels seen for CO1 OTUs [29]. This result suggests that SILVA assignments below the ordinal level should be treated with low confidence (i.e., assigned ⩽80% posterior probability). However, it is encouraging to note that SILVA taxonomic assignments (which are much faster than SAP assignments) appear to be reliable at ordinal and higher levels, since we find that at these levels SAP and SILVA assignments almost always agree (Table 3).

2.4 Alpha diversity

2.4.1 Homogenization method

The soil samples homogenized using the liquid nitrogen method produced only 91–168 OTUs per MID within each of the three pipelines, whereas bead-beating produced 101–312 OTUs per MID. The bead-beating approach therefore seems to release more soil faunal DNA.

2.4.2 Phylogenetic diversity and pipeline

Despite producing the fewest Arthropoda OTUs (Table 1), the OCTUPUS pipeline returned by far the highest phylogenetic diversity (PD) for any given number of OTUs. The

Table 2 Comparison of SAP and BLAST-to-SILVA (release 108) methods for taxonomic assignment of OTUs from the Denoiser/UCLUST pipeline[a)]

Taxonomy probability			% Identified			
Group		OTU count	Order	Family	Genus	Species
Nematoda	SAP	55	92%	38%	32%	26%
	Silva	66	100%	→	→	→
Collembola	SAP	60	100%	46%	24%	11%
	Silva	77	100%	→	→	→
Insecta	SAP	43	100%	50%	35%	33%
	Silva	49	100%	→	→	→
Arachnida/Acari	SAP	138	100%	20%	12%	7%
	Silva	158	100%	→	→	→
Arachnida/Araneae	SAP	5	100%	20%	20%	20%
	Silva	4	100%	→	→	→
Annelida	SAP	13	100%	48%	10%	10%
	Silva	14	100%	→	→	→
Total Metazoa	SAP	403	85%	28%	16%	13%
	Silva	455	100%	→	→	→
No taxonomy	SAP	219				
	Silva	214				

a) Ordinal-level assignment success is similar between the two methods. The SAP method allows an estimate of the posterior probability of assignment at each taxonomic level. Every sequence in the SILVA reference database is identified to the level of species, so all assignments of OTUs using SILVA are automatically at the species level. However, even at a low posterior probability threshold (⩾80%), the more conservative SAP method assigns only 7%–50% of OTUs to lower taxonomic levels (family to species). Thus, SILVA taxonomic assignments should be treated as high-confidence only down to the ordinal level.

Table 3 Taxonomic concordance between SILVA and SAP assignment methods, using OTUs from the Denoiser/UCLUST pipeline[a)]

Taxonomy level	Main groups	SAP	Silva	Agreement percentage
Domain	Eukaryota	408	403	98.8%
Kingdom	Metazoa	403	387	98.7%
	Viridiplantae	0	0	100%
Phylum	Annelida	13	12	92.3%
	Arthropoda	296	291	98.3%
	Nematoda	55	53	96.4%
Class	Insecta	43	41	95.3%
	Arachnida	143	138	96.5%
Order	Acari	121	118	97.9%
	Araneae	5	4	80%
	Collembola	61	60	98.4%
No taxonomy hit		219	214	97.7%

a) At ordinal and higher taxonomic levels, SAP and SILVA assignments are almost entirely in agreement, with the greatest disagreement in the Order Araneae.

other two pipelines returned PD estimates that are similar to each other, though their Arthropoda OTU richness differs by ~2× (Figure 2). Very high PD values from the OCTUPUS pipeline are caused by long terminal branches resulting from unavoidably poor alignments, in turn due to the absence of a denoising step in the OCTUPUS approach. To illustrate this, we used PyNAST [40] to align OTU sequences from the three pipelines against the SILVA 108 reference database. Mean pairwise similarity among OCTUPUS OTUs after PyNAST alignment was only 41.4%, substantially lower than pairwise similarities between USEARCH/CROP OTUs (66.3%) and Denoiser/UCLUST OTUs (77.7%).

2.5 Beta diversity

Despite the inevitable loss of taxonomic information due to sequencing noise and bioinformatic processing, all three pipelines produced compositionally distinct liquid nitrogen-extracted soil, bead-beaten soil, and leaf litter communities, as expected (Figure 1). Moreover, and reassuringly, the two DNA extraction replicates for the bead-beaten soil and for the liquid nitrogen-extracted leaf litter samples cluster together (Figure 1), and the community dissimilarity matrices produced by the three pipelines are significantly correlated with each other, as visualized by a Procrustes test of the PCoA ordinations (Figure 1). In short, despite the differences in the numbers of OTUs generated (Table 1), the three pipelines separate the samples in similar ways.

3 Discussion

The particular challenge of metabarcoding soil and leaf litter fauna is to separate the taxa of interest (metazoans) from an abundance of other biological diversity. Physical separation is a necessary first step [24] but is insufficient. In addition, CO1 primers are likely to be of limited utility as they appear to primarily amplify bacteria. However, as expected, a metazoan-specific 18S SSU rDNA primer is able to target animals in soil and leaf litter samples, and the taxa amplified in this study are characteristic of these habitats (higher frequencies of Arachnida, Collembola, and Myriapoda and a low frequency of Insecta; Table 1). Other candidate 18S primers that could be investigated in future have been described by Fonseca et al. [4]: SSU_FO4 (5′-GCTTGTCTCAAAGATTAAGCC-3′) and SSU_R22 (5′-GCCTGCTGCCTTCCTTGGA-3′).

We found that bead-beaten samples produced more OTUs than samples prepared using liquid nitrogen, and the former did not appear to have high rates of chimera formation [24]. Chimeras were also rare in a previous experiment using CO1 amplicons from bead-beaten arthropod samples [29].

The major conclusions can be drawn from our results, as follows: (i) the three bioinformatic pipelines assessed in this study produced very similar community compositions, as shown by (a) similar numbers of OTUs (Table 1), (b) similar taxonomic assignments for these OTUs (Table 1) and (c) highly correlated Principal Coordinates Analysis ordinations (Figure 1). (ii) BLAST searches against the SILVA 18S database assigns taxonomies to the ordinal and higher levels as reliably as does the more conservative but slower SAP program. Below the ordinal level, assignment confidence drops considerably. (iii) The OCTUPUS pipeline results in artifactually high estimates of phylogenetic diversity (Figure 2), which appears to be caused by greater sequence noise due to the absence of a denoising step in this method. Thus, we can recommend either of the other two pipelines for processing of 18S metabarcode data.

In general, it is a greater challenge to estimate alpha diversity from metabarcode datasets than to estimate beta diversity. The number of OTUs is highly sensitive to the choice of similarity threshold value, which is to some extent an arbitrary decision. In addition, our results indicate that if we increase the threshold to 99%, not only does the number

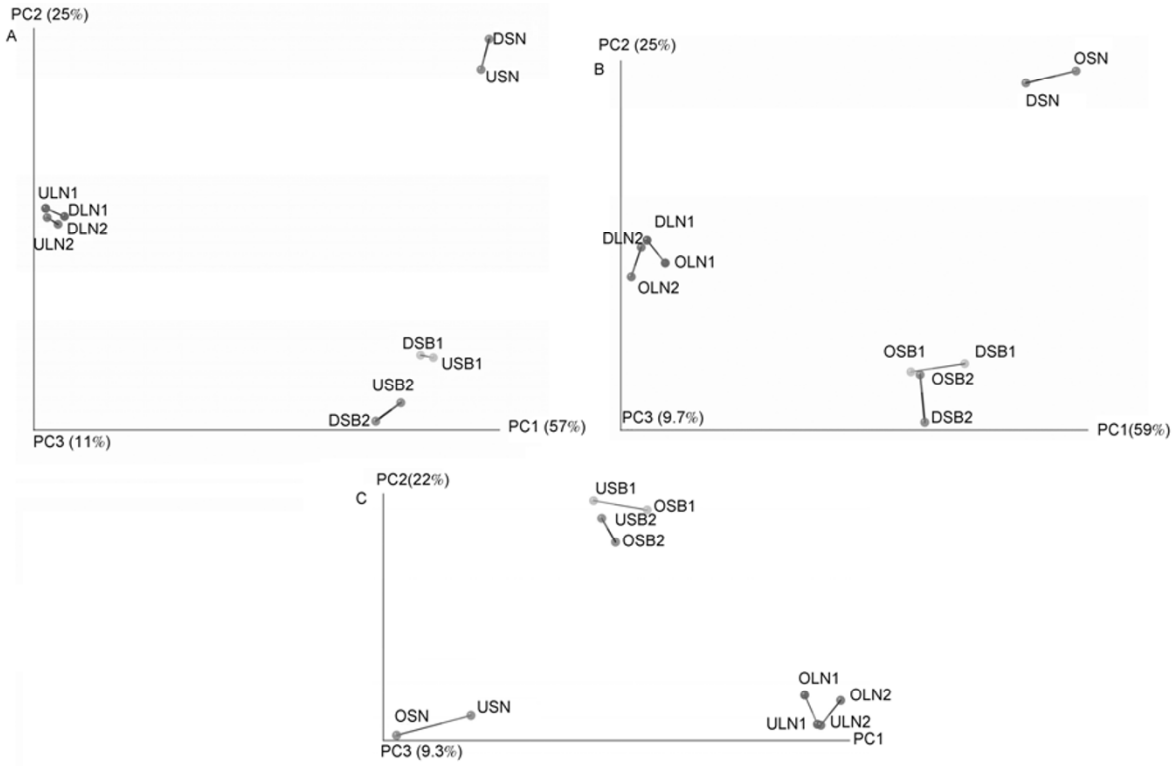

Figure 1 Comparisons of community compositions from the three processing pipelines, pairwise compared with Procrustes tests. Shown here are Principal Coordinates Analyses (PCoAs) of Sørensen-Dice dissimilarities (which are based on presence/absence only). All pairwise comparisons are highly significantly correlated. A, Denoiser/UCLUST versus USEARCH/CROP (Procrustes P=0.01). B, Denoiser/UCLUST versus OCTUPUS (P=0.03). C, OCTUPUS versus USEARCH/CROP (P=0.03). D, U and O on the diagrams denote, respectively, the Denoiser/UCLUST, USEARCH/CROP, and OCTUPUS pipelines. S, soil samples; L, leaf litter samples; B, bead-beating extraction; N, liquid-nitrogen extraction; 1 and 2, sample replicates. Soil and leaf litter samples are separated, and within the soil samples, bead-beaten and liquid nitrogen extracted samples are separated. Replicate extracts (1 vs. 2) cluster closely together.

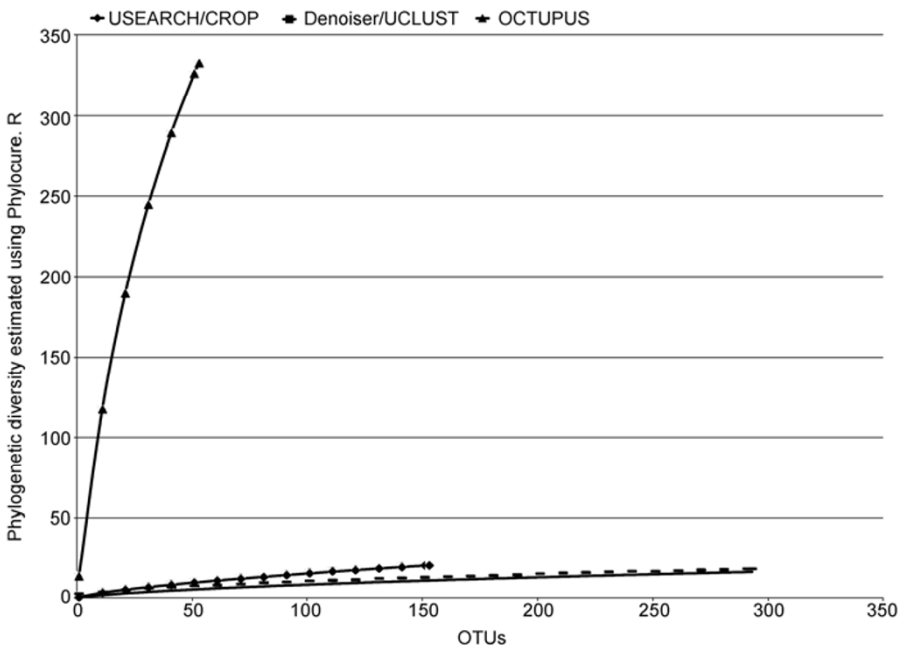

Figure 2 Arthropoda-only phylogenetic diversity (PD) rarefaction curves for the three pipelines, generated using Phylocurve.R and rarefied over the number of OTU. The OCTUPUS pipeline estimated the highest PD, despite extracting the fewest OTUs. In the Supporting Information, PD curves for each pipeline are broken down by sample within the pipeline (Figures S1 and S2).

of OTUs increase but the proportion assigned to Eukaryota also increases, changing community composition. At a 99% threshold the USEARCH/CROP pipeline assigned 82.4% of 1,177 OTUs to Eukaryota (vs. 53.5% at 96%) and the Denoiser/UCLUST pipeline assigned 72.9% of 1242 OTUs (vs. 68.3% at 96%). The choice of an optimal threshold for soil faunal datasets requires considerable further investigation.

Moreover, the relationship between true species abundance or biomass and the number of reads per OTU is, at best, complex [14,29,33], especially since 18S is a multiple-copy gene. However, metabarcoding makes it easy to take multiple samples and thus to determine the incidence (presence/absence) of an OTU across samples. Accordingly, we suggest that 'incidence-coverage estimators' (e.g., ICE and Chao2 in Gotelli & Colwell [34]) could be more robust measures of alpha diversity than are 'abundance-coverage estimators' (ACE and Chao1 in Gotelli & Colwell [34]), which are based on the read numbers per OTU. Unfortunately, we do not have sufficient replicate samples in our dataset to test this idea. Another estimator of alpha diversity is phylogenetic alpha diversity, PD [41]. With Phylocurve.R, we rarefied phylogenetic diversity (based on a common tree of the OTUs) over the total number of OTUs, and found that PD estimates were similar between the USEARCH/CROP and Denoiser/UCLUST pipelines, despite a 2× difference in raw OTU number. Thus, PD appears to be robust to artifactual differences in clustering techniques used to estimate OTUs, and the choice of threshold values. However, it remains a challenge to build an acceptable tree from out of the OTUs, especially with 18S sequences of differing lengths and high frequencies of indels. One promising possibility is to build a high-quality 18S tree from the SILVA database and then to 'place' OTU sequences on that tree, using, for instance, the software package pplacer [42], which is now available in the QIIME environment.

This work was supported by Yunnan Province (20080A001), Chinese Academy of Sciences (0902281081, KSCX2-YW-Z-1027), the National Natural Science Foundation of China (31170498), Ministry of Science and Technology of China (2012FY110800), Kunming Institute of Zoology, and the University of East Anglia.

1 Edgar R C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 2010, 26: 2460–2461
2 Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics, 2011, 27: 611–618
3 Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods, 2010, 7: 668–669
4 Fonseca V G, Carvalho G R, Sung W, et al. Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nat Commun, 2010, 1: 98
5 Pruesse E, Quast C, Knittel K, et al. 2007 SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res, 2007, 35: 7188–7196
6 Munch K, Boomsma W, Huelsenbeck J, et al. Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst Biol, 2008, 57: 750–757
7 Caporaso J G, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 2010b, 7: 335–336
8 Hamilton H C, Strickland M S, Wickings K, et al. Surveying soil faunal communities using a direct molecular approach. Soil Biol Biochem, 2009, 41: 1311–1314
9 Wu T, Ayres E, Bardgett R D, et al. Molecular study of worldwide distribution and diversity of soil animals. Proc Natl Acad Sci USA, 2011, 108: 17720–17725
10 Bienert F, De Danieli S, Miquel C, et al. Tracking earthworm communities from soil DNA. Mol Ecol, 2012, 21: 2017–2030
11 Epp L S, Boessenkool S, Bellemain E P, et al. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol Ecol, 2012, 21: 1821–1833
12 Porazinska D L, Giblin-Davis R M, Esquivel A. Ecometagenetics confirms high tropical rainforest nematode diversity. Mol Ecol, 2010a, 19: 5521–5530
13 Porazinska D L, Giblin-Davis R M, Faller L, et al. Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Mol Ecol Resources, 2009, 9: 1439–1450
14 Porazinska D L, Sung W, Giblin-Davis R M, et al. Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. Mol Ecol Resources, 2010b, 10: 666–676
15 Taberlet P, Coissac E, Hajibabaei M, et al. Environmental DNA. Mol Ecol, 2012, 21: 1789–1793
16 Yoccoz N G, Bråthen K A, Gielly L, et al. DNA from soil mirrors plant taxonomic and growth form diversity. Mol Ecol, 2012, 21: 3647–3655
17 Koskinen J P, Holm L. SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. Bioinformatics, 2012, 28: 438–443
18 Bik H M, Porazinska D L, Creer S, et al. Sequencing our way towardsunderstanding global eukaryoticbiodiversity. Cell, 2012, 27: 4
19 Smith B C, McAndrew T, Chen Z, et al. The cervical microbiome over 7 years and a comparison of methodologies for its characterization. PloS one, 2012, 7: 7
20 Lenz T, Becker S. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci—implications for evolutionary analysis. Gene, 2008, 427: 117–123
21 Coissac E, Riaz T, Puillandre N. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol, 2012, 21: 1834–1847
22 Taberlet P, Prud'Homme S M, Campione E, et al. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol Ecol, 2012, 21: 1816–1820
23 Yoccoz N G, Bråthen K A, Gielly L, et al. DNA from soil mirrors plant taxonomic and growth form diversity. Mol Ecol, 2012, 21: 3647–3655
24 Creer S, Fonseca V G, Porazinska D L, et al. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Mol Ecol, 2010, 19: 4–20
25 Somerfield P J, Warwick R M, Moens T. Meiofauna techniques. In: Methods for the Study of Marine Benthos. Oxford: Blackwell Science Ltd., 2005. 229–272
26 Folmer O, Black M, Hoeh W, et al. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Marine Biol Biotechnol, 1994, 3: 294–299
27 Edgar R C, Haas B J, Clemente J C, et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 2011, 27: 2194–2200
28 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 2006, 22: 1658–1659
29 Yu D W, Ji Y Q, Emerson B C, et al. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and

biomonitoring. Methods Ecol Evol, 2012, 3: 613–623
30 Haas B J, Gevers D, Earl A M, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res, 2011, 21: 494–504
31 Chou H H, Holmes M H. DNA sequence quality trimming and vector removal. Bioinformatics, 2001, 17: 1093–1104
32 Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 2004, 32: 1792–1797
33 Amend A S, Seifert K A, Bruns T D. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Mol Ecol, 2010, 19: 5555–5565
34 Gotelli N J, Colwell R K. Estimating species richness. In: Meagurran A E, McGill B J, eds. Biological Diversity: Frontiers in Measurement and Assessment. Oxford: Oxford University Press, 2011. 39–54
35 Nipperess D. Phylocurve: an R function for generating a rarefaction curve of phylogenetic diversity. http://davidnipperess.blogspot.com/2012/07/phylocurve-r-function-for-generating.html, 2011
36 R Development Core Team. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria, 2012
37 Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, 2003, 52: 696–704
38 Kearse M, Moir R, Wioson A, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinf Appl Note, 2012, 28: 1647–1649
39 Yang X D, Sha L Q. Species composition and diversity of soil mesofauna in the 'Holy Hills' fragmentary tropical rain forest of Xishuangbanna, China. Chin J Appl Ecol, 2010, 12: 261–265
40 Caporaso J G, Bittinger K, Bushman F D, et al. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics, 2010a, 26: 266–267
41 Faith D P, Baker A M. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evol Bioinf Online, 2006, 2: 121–128
42 Matsen F A, Kodner R B, Armbrust E V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics, 2010, 11: 538

Supporting Information

1 Example script commands for each pipeline
   a. USEARCH_CROP_commands.txt
   b. Denoiser_UCLUST_commands.txt
   c. OCTUPUS_commands.txt
2 18S OTU trees for estimating phylogenetic diversity
   a. USEARCH_CROP.nex
   b. Denoiser_UCLUST.nex
   c. OCTUPUS.nex
3 perl scripts used in the pipelines
   a. split_seq.pl
   b. otu_filter_trans.pl
   c. otu_table_withtax_from_sap_modified2.pl
   d. Seq_extract.pl
   e. octu_table_tax_V1.pl
4 Figures S1 and S2

Figure S1 compares alpha PD from the USEACRCH/CROP and Denoiser/UCLUST pipelines. Red dotted lines, leaf litter using liquid nitrogen; green lines, soil, using bead-beating; brown line, soil, using liquid nitrogen; blue line, all MIDs together.

Figure S2 plots by the OCTUPUS MIDs. Note that the MID-specific lines differ in total OTU numbers by more than a factor of 3, but PD is more similar amongst the MIDs, suggesting that total OTU number is perhaps a reliable index of alpha diversity.

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.