Biodiversity Soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring

Douglas W. Yu[1,2,*,**], Yinqiu Ji[1,*], Brent C. Emerson[2,3], Xiaoyang Wang[1], Chengxi Ye[1], Chunyan Yang[1], Zhaoli Ding[4]

[1] Ecology, Conservation, and Environment Center (ECEC), State Key Laboratory of Genetic Resources and Evolution, 32 Jiaochang East Rd., Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223 China
[2] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ UK
[3] Present address: Island Ecology and Evolution Research Group, IPNA-CSIC, C/Astrofísico Francisco Sánchez 3, 38206 La Laguna, Tenerife, Canary Islands, Spain
[4] Kunming Biodiversity Large-Apparatus Regional Center, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
* joint first authors
** corresponding author. Douglas W. Yu, School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ UK, email: dougwyu@gmail.com

Running title: Biodiversity soup

5782 words in Summary to Acknowledgments
1681 words in References
564 words in Table legends
212 words in Figure legends

Summary

1. Traditional biodiversity assessment is costly in time, money, and taxonomic expertise. Moreover, data are frequently collected in ways (e.g. visual bird lists) that are unsuitable for auditing by neutral parties, which is necessary for dispute resolution.

2. We present protocols for the extraction of ecological, taxonomic and phylogenetic information from bulk samples of arthropods. The protocols combine mass trapping of arthropods, mass-PCR amplification of the COI barcode gene, pyrosequencing, and bioinformatic analysis, which together we call ‘metabarcoding.’

3. We construct seven communities of arthropods (mostly insects) and show that it is possible to recover a substantial proportion of the original taxonomic information. We further demonstrate, for the first time, that metabarcoding allows for the precise estimation of pairwise community dissimilarity (beta diversity) and within-community phylogenetic diversity (alpha diversity), despite the inevitable loss of taxonomic information and resolution inherent to metabarcoding.

4. Alpha and beta diversity metrics are the raw materials of ecology and the environmental sciences, facilitating assessment of the state of the environment with a broad and efficient measure of biodiversity.

Keywords: 454 Genome Sequencer FLX System, DNA barcoding, high-throughput sequencing, metagenetics, metagenomics, phylogenetic diversity, OTU picking

Introduction

To manage the forces that affect the levels and distribution of biodiversity, we require the ability to measure biodiversity comprehensively, reliably, repeatedly, and over large scales. Efforts in this direction to date by ecologists and environmental biologists have been impeded by standard survey methodologies that consume large amounts of time, money, and taxonomic expertise, and we are therefore impeded from addressing biodiversity loss as a normal management problem that can be dealt with wherever and whenever it arises. Instead, most biodiversity research remains in the realm of basic science, and even then, scientists typically are forced to rely on proxies (Favreau *et al.* 2006; Lewandowski, Noss & Parsons 2010). One longstanding proxy has been to designate a subset of taxa as indicators, some popular ones being butterflies, dung beetles, birds, and parasitoid wasps (e.g. Gardner *et al.* 2008; Anderson *et al.* 2010).

Proxies might be efficient, but it is a truism in management that we only get what we measure. Schoolteachers evaluated on exam scores have an incentive to 'teach to the test,' and biological proxies are subject to the same narrowing of perspective. As one example, nineteen species of farmland birds have been designated as a biodiversity indicator on UK farmlands (JNCC 2011), the aim being to use birds to indicate overall farmland biodiversity. However, an understandable response has been to 'teach to the test' via supplemental winter feeding of farmland birds (Siriwardena *et al.* 2007; see also Newton 2011). Thus, in addition to proxies, we should tackle the lack of biodiversity information directly.

Here we describe a way to measure arthropod biodiversity rapidly, reliably, cheaply, comprehensively, over large spatial scales, and in ways that can be audited by third-parties, which is a requirement for dispute resolution. The first element is DNA barcoding, in which short gene sequences are used to identify species. The most commonly used barcode for animals is a 658bp section of the mitochondrial Cytochrome c Oxidase Subunit I gene (mtDNA COI) (Hebert *et al.* 2003). Other barcode genes are proposed for plants, protists, and meiofauna (Hollingsworth *et al.* 2009; Creer *et al.* 2010; Medinger *et al.* 2010; Yao *et al.* 2010). Because sequencing is fast and cheap, the barcode approach potentially provides large amounts of species-level inventory data, making it possible to track and measure biodiversity over space and time (e.g. Janzen *et al.* 2005; Waugh 2007; Borisenko *et al.* 2008). However, generating barcodes with Sanger sequencing is inefficient if we want to assign taxonomies to hundreds of thousands of samples, a requirement if we want to measure biodiversity repeatedly and over large spatial scales.

Our protocol therefore includes large-scale trapping for sample acquisition, high-throughput sequencing, and bioinformatic analysis. In short, mass-collected specimens are homogenised ('souped'), and the genomic DNA is extracted, mass-PCR-amplified for the barcode gene of interest, and sequenced on machines that can separate out individual DNA molecules. Bioinformatic tools then process the resulting huge number of sequences down to a dataset of manageable size and high-enough quality that is practical for subsequent analysis.

Altogether, we call this technique *metabarcoding* to distinguish it from the broader term *metagenetics*, which encompasses microbial communities, and from *metagenomics*, which, in addition, refers to the reconstruction of whole genomes. Finally, *environmental barcoding* or *eDNA* is probably best used to refer to the amplification and sequencing of free DNA from soil or water. We note, however, that the terminology is in flux.

Metabarcoding is transforming ecology (Creer 2010; Creer *et al.* 2010), especially of cryptic biodiversity. Recently, Fonseca et al. (2010) compared marine meiofauna (metazoans between 45 and 500 μm long) across beaches in the UK, Porazinska et al. (2010) compared nematode diversity in different rainforest microhabitats in Costa Rica, and Nolte et al. (2010) compared protist diversity across seasons in a lake in Austria. Nolte et al. further showed that for one genus, *Spumella*, species from the clade that is typically found in cold habitats are more abundant in cold months, whereas species belonging to warm-climate clades are more abundant in the summer. In these systems, previous studies had been impeded by the difficulty of measuring very high levels of diversity of very small taxa, and metabarcoding technology has unlocked this diversity, in the same way that microbiome biology has been unlocked by next-generation sequencing (Committee on Metagenomics 2007).

However, precisely because meiofauna and protists were so difficult to study before metabarcoding, independent validation of results has so far been forced to depend on small datasets based on morphospecies (Medinger *et al.* 2010), on laboratory tests (Porazinska *et al.* 2009a; Porazinska *et al.* 2009b), or on BLASTing reads against Genbank (Fonseca *et al.* 2010). These checks have been crucial, but by their nature, they do not fully validate metabarcoding as a method for making general measures of biodiversity.

The field requires further validation because metabarcoding promises important management advantages in addition to increased efficiency. Traditional biodiversity data relies on expertise that is difficult to standardise across multiple individuals, and errors (or even fraud) in direct observational data, such as bird lists, cannot subsequently be corrected or audited. In contrast, metabarcoding requires only that staff be able to carry out protocols using standard collection (e.g. pitfall, malaise, Winkler, and light traps) and laboratory techniques, and the raw sequence data remain available for future analyses. It is also possible to partition aliquots of the original collections, or the extracted DNA, for auditing. Another advantage is that metabarcoding can hitchhike on advances in software and laboratory practises that are being developed for bacterial metagenetics (Kosakovsky Pond *et al.* 2009; Caporaso *et al.* 2010b).

To further the process of turning metabarcoding into a standard management method, we apply the technique to the Arthropoda, especially the Insecta within it, for which it is easier to validate results against independent sampling, as well as other biodiversity proxies, such as vegetation (e.g. Gaspar, Gaston & Borges 2010). Arthropods are also a deserving focal group for direct study, as they form a major component of terrestrial biodiversity, provide important ecosystem services such as pollination, decomposition, and pest control, can themselves be pests and disease vectors, and are potentially indicative of plant diversity, since arthropods are mostly herbivores. Finally, with arthropods, it is easier to use the COI barcode gene, which holds some advantages over 18S and other nuclear rRNA genes (Emerson *et al.* 2011). COI is single copy, present in all taxa of interest, with the exception of a few protozoa, capable of being amplified across a wide range of taxa with a small set of primers (Folmer *et al.* 1994), especially with degenerate primer pairs (Rose, Henikoff & Henikoff 2003; Boyce, Chilana & Rose 2009), and has a faster substitution rate, compared to nuclear rRNA genes, which increases taxonomic resolution. Mitochondrial 12S and 16S genes satisfy these criteria, but COI has additional advantages. There exists a fast-growing taxonomic reference database (www.boldsystems.org, accessed 10 Sep 2011) with over 1.3 million specimen-vouchered records so far (Ratnasingham & Hebert 2007), and finally, the mutational properties of COI offer the opportunity to eliminate most pyrosequencing error (Emerson *et al.* 2011; Ranwez *et al.* 2011), a phenomenon that, if uncorrected, results in overestimates of diversity (Quince *et al.* 2009; Reeder & Knight 2010).

In light of this, a useful step forward was provided by Hajibabaei et al. (2011), who pyrosequenced the mini-barcode gene (the first 130 bp of COI) in test pools of Trichoptera and Ephemeroptera and BLASTed against reference sequences to recover 17 of 23 input species. They also showed that larval collections, which cannot be identified using morphology, could be identified using metabarcoding and that the collections matched known adult species assemblages from the same locations.

Following Fonseca et al.'s (2010) pioneering work with meiofaunal samples, the next step is to go beyond the recovery of species lists and to devise an efficient and adaptable pipeline that can independently turn huge lists of COI sequences into usable and high-quality taxonomic and ecological information. In particular, we wish to show that, even when some taxonomic information is lost, which is currently unavoidable in metabarcoding, it is still possible to recover precise estimates of alpha diversity and beta diversity.

We provide the research community with model laboratory protocols and bioinformatic scripts that can be adapted to incorporate new technologies and software as they arise. We also provide the original sequence data for software developers to use as test datasets *[note to reviewers: to be uploaded to dryad.org, which seems to require source papers to be in press or published]*. Our main contributions are: (1) new degenerate PCR primers to minimise allelic dropout of terrestrial arthropods (mostly but not only insects), (2) validation of several new software packages for denoising, *de novo* OTU picking, and taxonomic assignment (Table 1) within the QIIME pipeline (Caporaso *et al.* 2010b), which has active developer and user communities, (3) detailed scripts, methods, and datasets for users to learn with, (4) experimental demonstration that beta diversity can be recovered, and (5) experimental demonstration that rarefaction of phylogenetic diversity can recover alpha diversity (Nipperess 2011a; Nipperess 2011b).

Methods

Laboratory protocol

Sample collection. - Arthropods, mostly flying insects, and some small annelids were collected with malaise traps from three prefectures in Yunnan province China, Hong He (HONGHE), Xishuangbanna (XSBN) and Kunming (KMG), and preserved in 100% ethanol.

Sanger dataset. - 318, 795 and 316 individuals were hand-picked from HONGHE, XSBN and KMG. Each individual was extracted for genomic DNA using the HotSHOT method (Truett *et al.* 2000), the Qiagen DNEasy Blood and Tissue Kit, or the Bokun Insect DNA Extraction Magnetic Bead Kit (Changchun Bokun Biotech Co., www.bokunbio.com, Changchun, Jilin, accessed 14 Sep 2011) according to manufacturer's instructions. Individuals were then PCR amplified and Sanger sequenced for the 658bp region near the 5' terminus of the COI gene with Folmer's primers LCO1490 and HCO2198 (Folmer *et al.* 1994) (Table 2). PCR was carried out in 30 µl reaction volumes containing 3 µl of 10× buffer, 1.5mM MgCl2, 0.2mM dNTPs, 0.2 µM each primer, 1U Taq DNA polymerase (TaKaRa Biosystems), and approximately 100ng genomic DNA using a thermocycling profile of 95 °C for 2 min, 35 cycles of 95 °C for 15 s; 49 °C for 30 s; 72 °C for 1 min; and finally 72 °C for 7 min. Products were visualized on 2% agarose gels and were bidirectionally sequenced using BigDye version 3.1 on an ABI 3730xl DNA Analyser (Applied Biosystems). We obtained a total of 673 unique Arthropod and Annelid haplotypes (GENBANK accession numbers XXX - XXX). Sequences were truncated to 615 bp from the 5' end.

'454' dataset. - Genomic DNA sampled from individuals corresponding to the 673 unique haplotypes were pooled into seven mixtures, mimicking different ecological communities: HONGHE (n=197 unique haplotypes), XSBN (n=292), KMG (n=184), 2H1K (n=149), 1H1X

(n=198), 2K1X (n=150), 5K1X (n=121) (Figure 1). HONGHE, XSBN, and KMG share no haplotypes, and the latter four are mixtures of the first three, with the numbers indicating (approximate) ratios and letters indicating sources. Thus, 2H1K contains 99 haplotypes from HONGHE and 50 from KMG. For this test, extracting DNA individually and then combining increases our confidence in the composition of the mixtures. The implicit assumption is one that underlies all metagenetic and metagenomic biology: that the efficacy of DNA extraction kits is not affected by the number of species being extracted. We refer to the mixtures as 'MIDs' (Multiplex IDentifiers), following the terminology of the Genome Sequencer FLX System (454 Life Sciences, Roche Applied Science). Throughout, we use '454' to refer to this sequencing technology.

PCR amplification and pyrosequencing. - To maximize amplification of a diverse set of target sequences, we designed the degenerate primers, *Fol-degen-for* and *Fol-degen-rev*, to which we attached the standard A and B Roche adaptors and a MID tag for each community (Table 2). The primers are modifications of Folmer's (1994) primers and were created from an alignment of all 215 complete mtDNA COI gene sequences for Insecta that were present in Genbank (Supplementary Information). Across the 215 sequences, amino acid residues coded for by LCO1490 are conserved, with only a few exceptions involving species with no more than two divergent nucleotide positions. Based on the alignment, we designed *Fol-degen-for* to be fully degenerate to accommodate all possible codon variation for amino acid residues coded for by LCO1490. Amino acid residues coded for by HCO2198 were all conserved across the 215 sequences, so we designed *Fol-degen-rev* to be fully degenerate to accommodate all possible codon variation for amino acid residues coded for by HCO2198.

Each MID was amplified in five independent reactions and pooled. PCRs were performed in 20 μl reaction volumes containing 2 μl of 10× buffer, 1.5mM MgCl2, 0.2mM dNTPs, 0.4 μM each primer, 0.6U Taq DNA polymerase (TaKaRa Biosystems), and approximately 60ng of pooled genomic DNA. We used 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 17 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. We used non-proofreading Taq and fewer, longer cycles to reduce chimera production (following Lenz & Becker 2008). For pyrosequencing, all PCR products of all seven MIDs were gel purified by using a QIAquick PCR purification kit (QIAgen, Hilden, Germany), quantified by using the Quant-iT PicoGreen dsDNA Assay kit (Invitrogen), pooled and A-amplicon-sequenced twice on a 454, using two separate 1/8 regions of a plate.

Bioinformatics protocol: Recovery of input sequences

Sequence files from the two 1/8 plate regions were pooled to maximise coverage. Rather than produce a new analysis pipeline, we augment the QIIME pipeline (Caporaso *et al.* 2010b), which was designed for microbial metagenetics, with a number of new software packages (Table 1). Pyrosequencing data contain PCR chimeras (Lenz & Becker 2008), contaminant sequences, nuclear mitochondrial pseudogenes (Numts), PCR error and sequencing noise. The challenges for processing pyrosequencing data are to 'denoise' the sequences, remove chimeras, contaminants, and Numts, and quantify operational taxonomic units (OTUs). The latter phrase means to cluster the large number of sequences down to, ideally, the same number of unique sequences as there were species in the original samples. Thus, in this field, species are defined operationally as a cluster of similar sequences, and the clustering step is known as 'OTU picking.'

The seven major steps of our pipeline for denoising and quantifying OTUs, plus associated software, are summarised in Figure 2. Example scripts used to transform the output of a given step into the input for the subsequent step are provided as Supplementary Information. Most are from the QIIME pipeline, with some custom scripts that we have written.

Step 1. - Library splitting by MID and quality control. Primer and MID sequences are removed from the raw 454 reads, and the MID information is placed in the header line of each sequence (Table 2). Reads are also passed through a quality control filter that removes sequences with ambiguous nucleotides, with low quality scores (provided by the sequencer), with long repeats that are indicative of 'homopolymer' errors, and/or sequences that are too short or too long. Homopolymer errors occur because the 454 counts nucleotide additions via light bursts, and adding multiple nucleotides at once, e.g. AAA, in theory produces a three-times brighter burst, but in practise, often results in over- or under-estimates.

Step 2. - Initial denoising and de novo chimera removal. We first use PyNAST (Caporaso *et al.* 2010a) to align the post-quality-control sequences against a high-quality, aligned dataset of 17,087 Arthropod sequences (Supplementary Information) at a minimum similarity of 60%. Sequences that fail to align are discarded. The remaining sequences are clustered at 99% similarity with USEARCH (Edgar *et al.* 2011), and a consensus sequence is chosen for each cluster. The clustering step runs very quickly and more than halves the number of sequences (Table 3), speeding up downstream processing. We then apply the *de novo* chimera detection function UCHIME in USEARCH (Edgar *et al.* 2011), which exploits the prediction that small-size clusters are more likely to be chimeras.

Step 3. - Denoising. Sequences are denoised using MACSE (Ranwez *et al.* 2011), which takes advantage of the fact that COI is a coding gene by using the presence of stop codons to infer frameshift mutations caused by homopolymer errors and aligning at the amino-acid level to high-quality reference sequences. MACSE runs at a rate of ~1000 sequences per CPU-hour (on a 2010 iMac) so sequence files should be split into subfiles and run in parallel. An alternative to MACSE is PyroClean (Ramirez-Gonzalez *et al.* in manuscript), which produces similar results (results not shown). We remove sequences < 100bp, the length below which taxonomic information degrades rapidly (Meusnier *et al.* 2008).

Step 4. - OTU picking at 99% similarity. DNACLUST (Ghodsi, Liu & Pop 2011) is used because it ensures that no pairwise sequence comparison within an OTU differs by more than the user-chosen amount. This step reduces the workload for the next step.

Step 5. - OTU picking at 97% similarity. CROP (Hao, Jiang & Chen 2011) is a Bayesian clustering program that finds clusters "based on the natural organization of data without setting a hard cut-off threshold." CROP produces clusters within which ≥95% of sequences are more similar to the centre sequence than the desired cutoff (here, 97%). The bioinformatic challenge is to choose the sequence 'seeds' (cluster 'centres') that minimise cluster number. Note that sequence pairs within a cluster can differ by more than the cutoff. CROP is slow, requiring ~15-30 hours and 12 CPU cores to process a 30,000 sequence dataset, but we have found that CROP produces five to ten times fewer OTUs at the same similarity cutoff than do better-known programs like Cd-hit (Li & Godzik 2006) and UCLUST (Edgar 2010) (results not shown).

Step 6. - Taxonomic assignment of OTUs. The program SAP (Munch *et al.* 2008) assigns taxonomies by MCMC-sampling ten thousand unrooted phylogenetic trees constructed with a query sequence and its GENBANK homologues. The percentage of times that the query sequence is grouped with a given taxonomic level is the posterior probability that the query belongs to that taxonomic level. SAP runs at ~3 sequences/CPU-hour, so we split the OTU file and run in parallel. We then use a perl script (Supplementary Information) to extract the taxonomic information from SAP output and add it to the OTU table. This is the stage where real but contaminant sequences (e.g. *Homo sapiens*) are detected and removed, and we use the taxonomic data to identify the subset of OTUs assigned to the Arthropoda and Annelida (n = 973).

Step 7. - Final clean-up. Finally, we merge the sequence abundance data from the three OTU picking steps (USEARCH, DNACLUST, and CROP) to build an OTU table with sequence abundances (Table 4), and we delete singleton OTUs, reasoning that single reads are likely to be non-informative, since successfully amplified COI templates should be found in multiple copies. We show in results that this step does not affect the recovery of ecological information. We then use the Arthropoda-only OTUs (n = 598) to build a rooted, Tamura-Nei, gamma distance neighbour-joining tree (using an Onychophora sequence as the root). We suggest examining (and possibly deleting) any OTUs that result in (subjectively judged) very long branches, which are probably either local misalignments due to homopolymer errors or remaining chimeras, and to rebuild the tree. In this dataset, we did not observe any such long branches.

At the end of Step 7, we have OTUs assigned to MID and taxonomy (Table 4), which we call the '454-OTU' dataset, and a neighbour-joining tree of the Arthropoda OTU sequences, which is used to estimate phylogenetic diversity and dissimilarities. To test whether the 454-OTU dataset contains reliable information, we cluster the original 673 Sanger haplotypes into 547 'Sanger-OTUs' at 98% similarity and assign to MID and taxonomy (using SAP), and build a rooted NJ-tree.

Recovery of ecological information

Allelic dropout. - We BLASTed each of the 454-OTUs against the Sanger-OTU dataset at a stringency of 1e-10 and 97% minimum similarity to estimate the percentage of input species that did not amplify or survive the above pipeline. We also built a neighbour-joining tree combining the 454- and Sanger-OTUs to look for lone Sanger-OTUs (dropouts) and/or lone 454-OTU clusters (remaining chimeras, contaminants, Numts or noisy sequences).

Abundance versus presence-absence. - Beta diversity can be estimated using traditional dissimilarity indices that require only a Site X Species table (Table 4) or dissimilarity measures that take phylogeny into account (Faith & Baker 2006; Hamady, Lozupone & Knight 2010). Similarly, alpha diversity estimates range from simple counts of species richness to measures that incorporate evenness and/or phylogenetic diversity (PD). In both cases, we must ask whether it is valid to use sequence abundance (number of sequences per OTU) as a proxy of species abundance or biomass. Our opinion is that PCR amplification bias, although to some extent normalised by degenerate primer design, plus reaction stochasticity, corrupts correlations of sequence abundance with sample abundance in highly diverse datasets, which we support with a preliminary experiment in Supplementary Information (see Amend, Seifert & Bruns 2010). We therefore use presence-absence (unweighted) beta and alpha diversity indices (but see Porazinska *et al.* 2009b, who found that 18S rRNA read numbers did correlate with nematode frequencies).

Beta diversity. - We rarefy the 454-OTU table to equalise the number of reads per MID and then estimate pairwise compositional dissimilarities using the 1-Sørensen-Dice similarity index (an option in QIIME) and the unweighted Unifrac index (Hamady, Lozupone & Knight 2010). The latter incorporates phylogenetic distance. To test if the 454-OTU dataset preserves beta diversity information, we use a Mantel test to correlate the 454-dissimilarity matrix against the dissimilarity matrix produced from the Sanger dataset. We also visualise the dissimilarity matrices with a

Principal Coordinates Analysis (PCoA) and use a Procrustes test to test for correlation between the 454 and Sanger ordinations.

Alpha diversity. - Species diversity naturally increases with sample size (numbers of individuals captured), and sample sizes vary across MIDs. Rarefaction must therefore be used to control for the effects of sample size (Gotelli & Colwell 2001). One concern is that read abundance cannot be used to estimate the number of individuals per OTU. Fortunately, Nipperess (2011a; 2011b) has released R functions, *phylocurve.R* and *phylocurve.perm.R*, that can rarefy phylogenetic diversity over multiple numbers of species as a measure of sampling effort. In other words, the total PD of a sample is the sum of the branch lengths of all the OTUs in the sample, and rarefaction subsamples the phylogeny to allow comparisons across MIDs to be made at equal numbers of OTUs. We use *phylocurve.R* to rarefy the PD of each MID in both the 454 and Sanger datasets, and we compare at a common sampling effort of 101 OTUs (which allows all MIDs to be included). For this purpose, we rooted the NJ tree with an Onychophora sequence.

Results

Recovery of input sequences

Pyrosequencing reads were reduced 222-fold from 133,057 sequences to 598 OTUs at Step 7 (Table 3). There were few *de-novo* detected (and removed) chimeras, making up only 707 of 92,864 post-quality-control/PyNAST reads (0.8%). After the entire pipeline, the more-powerful refdb option of UCHIME, which used the input Sanger sequences as references (not possible under normal circumstances), detected only 20 chimeras out of 598 454-OTUs, although this does represent an increase in the ratio of chimera to non-chimera sequences (3.3%). The final numbers of 454-OTUs per MID are significantly predicted by the numbers of Sanger-OTUs per MID (Table 3), which argues that the pipeline has successfully reduced the dataset to mainly represent the input sequences.

Still, there are more 454-OTUs (598) than input OTUs (547), and not all of the former correspond to the latter (see *Allelic dropout* below). Hajibabaei et al. (2011) also report novel OTUs. Because arthropods were collected with malaise traps, some OTUs might represent tissue from species that had been in the same collecting bottles but not Sanger-sequenced and/or from food items and parasites, and it is also possible that some of the extra OTUs are laboratory contaminants, which suggests, unfortunately, that ancient-DNA protocols will be necessary for legally-sensitive work.

Allelic dropout. - Of the 547 Sanger-OTUs, 76% were BLAST-matched by one or more 454-OTUs, with most of the dropout in the Hymenoptera (only 57% matched) (Table 5). Allelic dropout subdivided by MID shows higher dropout percentages, as sequencing coverage per MID is necessarily lower (Supplementary Information). These results are achieved after omitting singleton OTUs (Table 1, Step 7). If singleton OTUs are included, overall dropout is reduced slightly to 19% (with ≥1-read OTUs), but many more 454-OTUs fail to BLAST-match to a Sanger-OTU (without singletons: 153/602=25.4% OTUs failed to match; with: 416/973 = 42.8% failed).

The important question is whether this perceived level of dropout causes loss of ecological information. First, we note that with any hard cutoff, we lose power. Thus, some Sanger-OTUs might indeed be represented by 454-OTUs, but so noisily as to fail to be BLAST-matched. In Supplementary information (FigTree datasets), inspection of the combined tree finds that many of the Sanger sequences that received no 454 BLAST-matches (putative dropouts) nonetheless cluster with one or more 454-OTUs. Nonetheless, there again are clearly more dropouts in the Hymenoptera. Similarly, all but a few of the 454-OTUs cluster close to a Sanger sequence, suggesting that there are few chimeras, Numts, or excessively noisy sequences in our dataset. In short, the 454-OTU dataset contains much the same phylogenetic structure as the input dataset, but with more branching near the tips. Below, we find that the 454-dataset allows us to recover most ecological information.

Recovery of ecological information

Taxonomy. - Despite being on average half the length of the Sanger-OTUs, 969/973 (99.6%) 454-OTUs at Step 6 could be identified to class, and 96% could be identified to order, which are only slightly lower than the success rates of Sanger-OTUs. However, at the family, genus, and species levels, taxonomic assignment of 454-OTUs is less than half that of Sanger-OTUs (Table 6). As barcode databases grow (www.boldsystems.org) and read lengths increase, we expect that assignment success at lower taxonomic levels will increase.

Beta diversity. - Unifrac dissimilarity matrices of the Sanger- and 454-OTUs are highly significantly correlated (Table 7). We can visualise this correlation by using a Procrustes analysis to overlay PCoA ordination diagrams constructed from the dissimilarity matrices, and we find that the first three axes of the two ordinations are also highly significantly correlated (Figure 3). These results appear to be robust, as dissimilarity matrices calculated using the non-phylogenetically-informed 1-Sørensen-Dice index are also highly correlated (Mantel, 9999 permutations, $p < 0.001$). (PCoA ordinations of the Sørensen-Dice dissimilarity matrices are in Supplementary Information). In short, community differences between sampling locations appear to be well preserved in pyrosequencing data.

Alpha diversity. - After rarefaction to control for sampling effort, we find that the phylogenetic diversity of each MID calculated from the 454-dataset highly significantly predicts PD in the corresponding Sanger MID (Figure 4).

Discussion

After mass-PCR amplification and high-throughput sequencing of arthropod DNA, we demonstrate how a denoising and *de novo* OTU-picking pipeline makes it possible to recover taxonomic information from a wider range of taxa than tested in Hajibabaei et al. (2011) and, for the first time, to recover alpha and beta diversity information. This information is the raw material of basic and applied research in ecology and the environmental sciences. As examples, we are now using this protocol for the following applications: (1) projecting the effects of climate change on insect species compositions with light-trap collections along an altitudinal transect (beta diversity); (2) measuring the conservation value of shade-tea versus natural subtropical forests, of once- and twice-logged versus unlogged rainforests, and of Buddhist sacred mountains versus control sites in an alpine habitat (alpha diversity); and (3) determining the management treatments that are most successful in restoring endangered heathland habitat and maintaining insect biodiversity in a temperate forest (alpha and beta diversity). A typical dataset has required one to two months to process, from DNA extraction to bioinformatic analysis, and the bioinformatic analyses for multiple studies can be conducted in parallel. We are also using these studies as ‘field-validation,’ which is to match our estimates of alpha and beta diversity against independent biodiversity measures collected with standard census techniques.

While we have great optimism for the approach we have outlined, we do caution that a number of drawbacks remain. Transport of samples can be legally complicated and expensive, especially if preserved in alcohol and moved across borders. Additionally, the extraction protocol requires sample destruction (although with additional effort, one can use a leg for all but the smallest of samples and retain the rest of the sample as a voucher). Sequencing is costly (although this is balanced against the time and effort of taxonomic experts, and costs should plummet in the next few years). Few OTUs are identified to species (but this will improve as a function of the growth of the BOLD database). Care needs to be taken in the field and in the lab to reduce sample contamination. Abundance data are not available (but subsampling sites, at extra cost, provides an abundance index (Jerde *et al.* 2011)). Finally, the bioinformatics stage is time-consuming, and as yet there is no best-practise pipeline. A consequence is that it remains unclear to what extent metagenetic data will be robust to legal challenges, if used for environmental monitoring and planning.

Another consequence is that there are alternative pipelines, and we have only presented one (see also Fonseca *et al.* 2010). For instance, Hao et al. (2011) report that CROP is robust to non-denoised datasets, which is a time-saving option, as is using a lower similarity threshold such as 95%. Our purpose here is not to define a specific pipeline but to give the community a validated starting point and added confidence that degenerate primers plus already existing hardware and software can recover useful ecological information from bulk samples of arthropods.

Because this is a fast-moving field, we end by listing anticipated and desired future improvements.

1) Better PCR primers required. This is most necessary for Hymenoptera and for soil fauna (e.g. Protura, Collembola, Annelida, Arachnida). One possibility is a divide and conquer approach using different sets of primers for the amplification of major faunal groups, and then combining the result pcr products for sequencing. We note that despite designing the primers (Table 2) with only Insecta sequences, we have been finding that they amplify many OTUs subsequently assigned to non-Insecta arthropods, including Collembola, Protura, and Arachnida (authors' unpublished results). Careful consideration of existing COI sequence data that spans COI priming sites will dictate the best approach for a given set of taxa.

2) Avoid PCR. In bacterial metagenetics, it is feasible to sequence genomic DNA directly and search for barcode sequences bioinformatically (Sharpton *et al.* 2011), which avoids dropout and might even provide reliable abundance information. With animals, the presence of very large nuclear genomes probably needs a protocol to concentrate mitochondria before DNA extraction and sequencing.

3) Targeted species detection. In a remarkable study, Ficetola et al. (2008) used custom PCR primers and Sanger sequencing to detect invasive American bullfrogs in samples of pond water alone. Jerde et al. (2011), Goldberg et al. (2011) and Thomsen et al. (2011) have further validated the use of environmental DNA in water bodies, even in fast-moving streams, for detecting a variety of vertebrate and invertebrate species, including a mammal species. In our own work, we have detected several vertebrate species in our insect malaise trap samples (bats, frogs, birds, and ungulates) that are known to exist in the trapping area (authors' unpublished data). We suspect that we are amplifying blood borne by mosquitoes, and this suggests that terrestrial vertebrate diversity might be measurable with mass mosquito trapping. It will be necessary to validate laboratory and statistical protocols for assigning probabilities of assignment of reads to target templates and to design standard controls.

4) New software packages arise constantly. For instance, a new pipeline, *otupipe.pl*, uses only USEARCH to denoise, remove chimeras, and pick OTUs (drive5.com/otupipe, accessed 10 Sep

2011). The pipeline is very fast (minutes versus days with our bioinformatic protocol) but cannot yet handle multiple MIDs, nor has it been validated.

5) Hardware improves rapidly. Illumina and Ion Torrent sequencers produce orders of magnitude more sequences per run (and/or dollar) but currently are limited to shorter read lengths or accept only short amplicons. However, these sequencers are advancing so quickly that they will probably be competitive with 454 sequencers for many uses in just a few years.

6) Better databases required. Bacterial metagenomics enjoys large, curated databases for taxonomic assignment (DeSantis *et al.* 2006), and while a similar database exists for COI (Ratnasingham & Hebert 2007), it is not yet integrated with taxonomic assignment programs (e.g. SAP Munch *et al.* 2008), nor downloadable to local computers.

7) Coverage estimates required. We do not currently have a good handle on how much sequencing depth (number of reads) is required for sequencing a given number of individuals at given probabilities (a separate problem from PCR bias).

Acknowledgments

For support, we thank Yunnan Province (20080A001), the Chinese Academy of Sciences (0902281081, KSCX2-YW-Z-1027, Y002731079), the National Science Foundation of China (31170498), and the University of East Anglia. We thank Ricardo Ramirez-Gonzalez for early access to PyroClean, Sam Ma for advice, and Kasper Munch, Mohammadreza Ghodsi, and Xiaolin Hao for help with their respective software packages. DY and BE initiated the project; DY and YQJ designed the bioinformatics pipeline and performed data analysis; YQJ designed the laboratory protocol; BE designed the primers; XYW and CXY wrote additional programs; ZLD performed the 454 sequencing; YQJ and CYY conducted the laboratory work; DY wrote the paper, and DY, YQJ, and BE contributed revisions.

References

Amend, A.S., Seifert, K.A. & Bruns, T.D. (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Molecular Ecology, 19, 5555-5565.

Anderson, A., Mccormack, S., Helden, A., Sheridan, H., Kinsella, A. & Purvis, G. (2010) The potential of parasitoid Hymenoptera as bioindicators of arthropod diversity in agricultural grasslands. Journal of Applied Ecology.

Borisenko, A., Lim, B., Ivanova, N., Hanner, R. & Hebert, P. (2008) DNA barcoding in surveys of small mammal communities: a field study in Suriname. Molecular Ecology Resources, 8, 471-479.

Boyce, R., Chilana, P. & Rose, T.P. (2009) iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucleic Acids Research, 37, w222-w228.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L. & Knight, R. (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics, 26, 266-267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010b) QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7, 335-336.

Committee on Metagenomics (2007) The New Science of Metagenomics: Revealing the secrets of our microbial planet. pp. 170. National Research Council of the National Academies, Washington, D.C.

Creer, S. (2010) Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. Molecular Ecology, 19, 2829-2831.

Creer, S., Fonseca, V.G., Porazinska, D.L., Giblin-Davis, R.M., Sung, W., Power, D.M., Packer, M., Carvalho, G.R., Blaxter, M.L., Lambshead, P.J.D. & Thomas, W.K. (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Molecular Ecology, 19, 4-20.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied Environmental Microbiology, 72, 5069-5072.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics.

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.

Emerson, B.C., Cicconardi, F., Fanciulli, P.P. & Shaw, P.J.A. (2011) Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. Philosophical Transactions of the Royal Society B, in press.

Faith, D.P. & Baker, A.M. (2006) Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evolutionary Bioinformatics Online, 2, 121-128.

Favreau, J.M., Drew, C.A., Hess, G.R., Rubino, M.J., Koch, F.H. & Eschelbach, K.A. (2006) Recommendations for Assessing the Effectiveness of Surrogate Species Approaches. Biodiversity and Conservation, 15, 3949-3969.

Ficetola, G.F., Miaud, C., Pompanon, F. & Taberlet, P. (2008) Species detection using environmental DNA from water samples. Biology Letters, 4, 423-425.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3, 294-299.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P., Packer, M., Blaxter, M.L., Lambshead, P.J.D., Thomas, W.K. & Creer, S. (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nature Communications, 1, 1-8.

Gardner, T.A., Barlow, J., Araujo, I.S., Avila-Pires, T.C., Bonaldo, A.B., Costa, J.E., Esposito, M.C., Ferreira, L.V., Hawes, J., Hernandez, M.I.M., Hoogmoed, M.S., Leite, R.N., Lo-Man-Hung, N.F., Malcolm, J.R., Martins, M.B., Mestre, L.A.M., Miranda-Santos, R., Overal, W.L., Parry, L., Peters, S.L., Ribeiro-Junior, M.A., da Silva, M.N.F., Motta, C.d.S. & Peres, C.A. (2008) The cost-effectiveness of biodiversity surveys in tropical forests. Ecol Lett, 11, 139-150.

Gaspar, C., Gaston, K.J. & Borges, P.A.V. (2010) Arthropods as surrogates of diversity at different spatial scales. Biological Conservation, 143, 1287-1294.

Ghodsi, M., Liu, B. & Pop, M. (2011) DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics, 12, 271.

Goldberg, C.S., Pilliod, D.S., Arkle, R.S. & Waits, L.P. (2011) Molecular detection of vertebrates in stream water: a demonstration using rocky mountain tailed frogs and IDAHO giant salamanders. PLoS ONE, 6, e22746.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters, 4, 379-391.

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.

Hamady, M., Lozupone, C. & Knight, R. (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J, 4, 17-27.

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

Hebert, P., Cywinska, A., Ball, S. & Dewaard, J. (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society London B, 270, 313-321.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.-J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., Chacón, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanakurti, P.R., Kim, J.S., Kim, Y.-D., Lahaye, R., Lee, H.-L., Long, D.G., Madriñán, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C.-W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D.-K. & Little, D.P. (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences, 106, 12794-12797.

Janzen, D.H., Hajibabaei, M., Burns, J.M., Hallwachs, W., Remigio, E. & Hebert, P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philosophical Transactions of the Royal Society B-Biological Sciences, 360, 1835-1845.

Jerde, C.L., Mahon, A.R., Chadderton, W.L. & Lodge, D.M. (2011) “Sight-unseen” detection of rare aquatic species using environmental DNA. Conservation Letters, 4, 150-157.

JNCC (2011) Joint Nature Conservation Committee: UK Biodiversity Indicators. Joint Nature Conservation Committee.

Kosakovsky Pond, S., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W.-Y., Taylor, J. & Nekrutenko, A. (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline. Genome Research, 1-11.

Lenz, T. & Becker, S. (2008) Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis. Gene, 427, 117-123.

Lewandowski, A.S., Noss, R.F. & Parsons, D.R. (2010) The effectiveness of surrogate taxa for the representation of biodiversity. Conservation Biology, 24, 1367-1377.

Li, W. & Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22, 1658-1659.

Medinger, R., Nolte, V., Pandey, R.V., Jost, S., Ottenwälder, B., Schlötterer, C. & Boenigk, J. (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Molecular Ecology, 19, 32-40.

Meusnier, S., Singer, G.A.C., Landry, J.-F., Hickey, D.A., Hebert, P.D.N. & MHajibabaei, M. (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics, 9, 214.

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

Newton, A.C. (2011) Implications of Goodhart's Law for monitoring global biodiversity loss. Conservation Letters, 4, 264-268.

Nipperess, D. (2011a) phylocurve: an R function for generating a rarefaction curve of Phylogenetic Diversity. pp. webpage for downloading phylocurve.R.

Nipperess, D. (2011b) phylocurve.perm: an R function for generating a rarefaction curve of Phylogenetic Diversity by randomisation.

Nolte, V., Pandey, R.V., Jost, S., Medinger, R., Ottenwälder, B., Boenigk, J. & Schlötterer, C. (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Molecular Ecology, 19, 2908-2915.

Porazinska, D.L., Giblin-Davis, R.M., Esquivel, A., Powers, T.O., Sung, W. & Thomas, W.K. (2010) Ecometagenetics confirms high tropical rainforest nematode diversity. Molecular Ecology, 19, 5521-5530.

Porazinska, D.L., Giblin-Davis, R.M., Faller, L., Farmerie, W., Kanzaki, N., Morris, K., Powers, T.O., Tucker, A.E., Sung, W. & Thomas, W.K. (2009a) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Molecular Ecology Resources, 9, 1439-1450.

Porazinska, D.L., Sung, W., Giblin-Davis, R.M. & Thomas, W.K. (2009b) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. Molecular Ecology Resources, 10, 666-676.

Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F. & Sloan, W.T. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods, 6, 639-641.

Ramirez-Gonzalez, R., Heavens, D., Caccamo, M. & Emerson, B.C. (in manuscript) Community barcoding for rapid, species-level biodiversity assessment.

Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E.J.P. (2011) MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS ONE, 6, e22594.

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes, 7, 355-364.

Reeder, J. & Knight, R. (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods, 7, 668-669.

Rose, T.M., Henikoff, J.G. & Henikoff, S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Research, 31, 3763-3766.

Sharpton, T.J., Riesenfeld, S.J., Kembel, S.W., Ladau, J., O'Dwyer, J.P., Green, J.L., Eisen, J.A. & Pollard, K.S. (2011) PhylOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Computational Biology, 7, e1001061.

Siriwardena, G.M., Stevens, D.K., Anderson, G.Q.A., Vickery, J.A., Calbrade, N.A. & Dodd, S. (2007) The effect of supplementary winter seed food on breeding populations of farmland birds: evidence from two large-scale experiments. Journal of Applied Ecology, 44, 920-932.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Wiuf, C., Rasmussen, M., Gilbert, M.T.P., Orlando, L. & Willerslev, E. (2011) Monitoring endangered freshwater biodiversity using environmental DNA. Molecular Ecology.

Truett, G.E., Heeger, P., Mynatt, R.L., Truett, A.A. & Walker, J.A. (2000) Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). BioTechniques, 29, 52-54.

Waugh, J. (2007) DNA barcoding in animal species: progress, potential and pitfalls. BioEssays, 29, 188–197.

Yao, H., Song, J., Liu, C., Luo, K., Han, J. & Hansson, B. (2010) Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals. PLoS ONE, 5, 370-375.

Supplementary Information

1. Table S1. Allelic dropout by MID and taxon.
2. Figure S1. Principal coordinates ordination of the 454 OTUs.
3. Figure S2. Principal coordinates ordination of the Sanger OTUs.

Other Supplementary information.

4. Read vs sample abundances analysis (XSBN_abundance_analysis-20111228.xlsx, see script commands in “Example script commands.txt,” subheading: Read abundances vs sample abundances, using XSBN)
5. figtree datasets (CombinedTreeWithColour_mutual_BLAST.figtree and CombinedTreeWithColour_Sanger.figtree, figtree_file_descriptions.txt)
6. Example script commands.txt

The following perl scripts and datasets are used in the script commands and are provided as inputs and examples.

7. perl script: OTU_filter_trans.pl
8. perl script: Otu_table_withtax_from_sap_modified2.pl
9. perl script: Pick_cluster_num_V1.pl
10. perl script: Replace_Seq_name_V2.pl
11. perl script: Sequence_filter_V2.pl
12. perl script: Split_seq.pl
13. dataset: 454_Map.txt
14. dataset: Reference sequences: Arthropoda_ref.fasta
15. dataset: Reference sequences: Priming site only alignment.txt
16. dataset: Reference sequences: Sanger_orig.fasta
17. dataset: 454 OTU table: 454_CROP97_withtax.txt
18. dataset: Sanger OTU table: otu_table_withtax.txt
19. dataset: 454-OTUs: 97Arthropoda_Annelida_seqs_80_noSingletons.fas
20. dataset: Sanger-OTUs: Sanger_reprset.fas

Raw sequence data to be archived at http://datadryad.org, not included in submission

1. Sequence_files/original_454_files/split_library_output_1/seqs.fna
2. Sequence_files/original_454_files/split_library_output_2/seqs.fna

Table 1. Software packages and pipelines used. Software dependencies not listed. (OTU = operational taxonomic unit).

Package name	Reference	Used here for	URLs
QIIME v. 1.3.0	Caporaso et al. (2010b)	Main pipeline: library splitting and quality control, PyNAST sequence alignment, OTU picking, OTU table creation and rarefaction, beta diversity analyses (Procrustes, Mantel), network visualisation	qiime.sourceforge.net, Mac OSX implementation available at www.wernerlab.org/software/macqiime
FastUnifrac v. 1.5.1	Hamady et al. (2010)	Phylogenetic beta diversity estimation	included in QIIME
PyNAST v. 1.1	Caporaso et al. (2010a)	Fast sequence alignment to template alignment	included in QIIME
USEARCH v. 4.2.66	Edgar (2010); Edgar et al. (2011)	Denoising, chimera removal, sequence sorting and initial OTU picking	drive5.com/usearch/usearch4.0.html
MACSE 0.8b	Ranwez et al. (2011)	Denoising	mbb.univ-montp2.fr/macse
PyroClean v. 0.1	Ramirez-Gonzalez et al., in manuscript	Denoising	not yet available
DNACLUST v. 1	Ghodsi et al. (2011)	OTU picking	dnaclust.sourceforge.net
CROP v. 1.31	Hao et al. (2011)	OTU picking	code.google.com/p/crop-tingchenlab
SAP v. 1.0.12.	Munch et al. (2008)	Taxonomic assignment	www.daimi.au.dk/~kmt/StatisticalAssignmentPackage.html
phylocurve.R & phylocurve.perm.R	D. Nipperess, unpublished	Rarefaction of phylogenetic diversity	homepage.mac.com/davidnipperess/page2/page2.html
Geneious v. 5.4.6	Drummond et al. (2011)	Neighbour-joining tree construction, sequence file and tree visualisation	www.geneious.com

Table 2. Primers and MIDs used in this study. LCO1490 and HCO2198 are the Sanger primers. *Fol-degen-for/rev* are the degenerate primers for mass amplification. Adaptors A and B are used by the '454' sequencer to attach individual DNA molecules to microscopic beads, for subsequent sequencing. MIDs (Multiplex Identifiers) are 10 bp sequences that allow different samples to be sequenced together on a single '454' plate and then separated bioinformatically for downstream analysis. There is no need to add MIDs to *Fol-degen-rev*, because we only pyrosequenced from the forward direction. The last two rows contain an example of a 454 read, with the MID underlined and the forward primer dotted underlined. Both are removed in the split_libraries.py step, and the corresponding MID info is added to the header line (i.e. ACGCTCGACA = HONGHE, see the 454_Map.txt file in Supplementary Information), plus a unique number for that read, together making the sequence_ID. The reverse primer was never in the read because it was too short; 454 reads rarely exceed 600 bp.

Primer	Sequence (5' → 3')
LCO1490	GGTCAACAAATCATAAAGATATTGG
HCO2198	TAAACTTCAGGGTGACCAAAAAATCA
Fol-degen-for	Adaptor A + MID + TCNACNAAYCAYAARRAYATYGG
Fol-degen-rev	Adaptor B + TANACYTCNGGRTGNCCRAARAAYCA
Raw 454 sequence (adaptor sequence omitted)	>GWL4WKW01A3F7T rank=0000061 x=332.5 y=839.0 length=101 ACGCTCGACATCAACCAACCATAAGGATATTGGTTGTGGTAATACATCAAGGGGTCACACATTTAGTGATTTT TGGACACCCGGAAGTATACTGAGCGGCT
Post split_libraries.py sequence	>HONGHE_1 GWL4WKW01A3F7T orig_bc=ACGCTCGACA new_bc=ACGCTCGACA bc_diffs=0 TTGTGGTAATACATCAAGGGGTCACACATTTAGTGATTTTTGGACACCCGGAAGTATACTGAGCGGCT

Table 3. Basic statistics and OTU picking progression. **A.** Numbers of raw pyrosequencing reads before and after quality control (Step 1, Fig. BIOINFORMATICS), subdivided by MID (Multiplex Identifier) and sequencing region. Each region is 1/8 of a 454 plate. The name of each mixture MID (e.g. 2H1K) indicates the ratio of haplotypes from the source MIDs (HONGHE, KMG, XSBN, Figure 1). **B.** Numbers of OTUs (operational taxonomic units) after each major bioinformatic step (Figure 2). Starting with Step 4, OTU-picking is performed on the combined dataset, so the sum of OTUs per MID is greater than the number of unique OTUs, indicating that OTUs are shared amongst MIDs, as designed. **C.** Numbers of unique haplotypes and 98% similarity OTUs, subdivided by MID, in the Sanger (input) dataset. Across the 7 MIDs, the number of Sanger-OTUs predicts the final number of 454-OTUs (linear regression, $F_{1,5}$=6.6, p=0.039, R^2=0.61). Software package details in Table 1.

A. 454 raw data				Post-quality-control sequences (number of reads)							
454 plate	Raw reads	Length$_{Max}$	Length$_{Avg}$	All MIDs	HONGHE	KMG	XSBN	1H1X	2H1K	2K1X	5K1X
Region 1	65,554	617 bp	350.3 bp	48,531	5,239	8,898	5,464	5,147	9,560	5,987	8,236
Region 2	67,503	604 bp	333.8 bp	48,128	5,126	8,724	5,444	5,155	9,471	5,950	8,258
Total	133,057	⟶		96,659	10,365	17,622	10,908	10,302	19,031	11,937	16,494

Step 1 (convert raw to post-quality-control sequences, QIIME)

B. 454 dataset	454-OTUs (operational taxonomic units, total and per MID)							
	Unique OTUs	HONGHE	KMG	XSBN	1H1X	2H1K	2K1X	5K1X
Step 2 (PyNAST, 60%)	92,864	9,906	16,872	10,686	9,798	18,189	11,547	15,866
Step 2 (USEARCH) [1]	41,087 [707]	4,930 [33]	7,568 [23]	4,903 [105]	4,369 [181]	8,321 [29]	4,949 [125]	6,047 [211]
Step 3 (MACSE)	41,057	4,925	7,561	4,900	4,366	8,318	4,945	6,042
Step 4 (DNACLUST, 99%)	34,905	4,540	6,735	4,348	3,934	7,261	4,313	5,140
Step 5 (CROP, 97%)	1,047	278	277	231	236	240	171	144
Step 6 (SAP)	973	258	254	223	224	218	157	133
Step 7 (Final clean-up) [2]	598 [20]	192 [2]	174 [3]	183 [2]	162 [3]	153 [1]	128 [5]	98 [4]

C. Sanger dataset	Sanger-OTUs (total and per MID)							
Total haplotypes	673	197	184	292	198	149	150	121
OTUs (UCLUST, 98%)	547	167	153	230	159	140	134	106

[1] Numbers in brackets indicate deleted *de novo* chimeras (detected using the USEARCH --uchime *de novo* function].

[2] Numbers in brackets indicate chimeras remaining (detected using the USEARCH --uchime --refdb function with the Sanger dataset).

Table 4. Example rows from an OTU table, with assigned taxonomy, edited for clarity. Full tables are in Supplementary Information.

#OTU ID	1H1X	2H1K	2K1X	5K1X	HONGHE	KMG	XSBN	Taxonomy assigned at probability ≥ 80% by SAP
34278	0	1	0	0	1	0	0	Eukaryota;Metazoa;Arthropoda;Hexapoda;Insecta;Neoptera;Hymenoptera
29894	1	0	0	0	0	0	5	Eukaryota;Metazoa;Arthropoda;Hexapoda;Insecta;Neoptera;Hemiptera

Table 5. Allelic dropout. Figures indicate the number of Sanger-OTUs that were successfully BLAST-matched by at least one of the 598 454-OTUs (at 1e-10 and ≥97% similarity, Step 7, Table 3). ≥2-read and ≥5-read OTUs indicate the number of OTUs with that minimum cluster size. Note that singleton-OTUs (1-read OTUs) were removed in Step 7. Across all taxa, 76% of Sanger-OTUs were matched by ≥1 454-OTU (i.e. 24% dropout), with the greatest dropout in the Hymenoptera. Tables of allelic dropout subdivided by taxon and MID are provided in supplementary information.

		Number of 454-OTUs (after Step 7) successfully BLAST-matched to Sanger-OTUs at 1e-10, 97% similarity			
Taxa	Sanger-OTUs	≥ 2-read OTUs		≥ 5-read OTUs	
Lepidoptera	172	127	74%	112	65%
Diptera	169	144	85%	126	75%
Hymenoptera	108	58	54%	39	36%
Coleoptera	39	31	79%	28	72%
Hemiptera	38	33	87%	29	76%
Psocoptera	7	7	100%	7	100%
Arachnida	5	4	80%	2	40%
Blattaria	2	2	100%	2	100%
Plecoptera	2	2	100%	2	100%
Trichoptera	2	2	100%	2	100%
Ephemeroptera	1	1	100%	1	100%
Odonata	1	1	100%	1	100%
Annelida	1	1	100%	1	100%
All Taxa	547	413	76%	352	64%

Table 6. Taxonomic assignment of Arthropoda and Annelida OTUs to four lower taxonomic levels. Class- and ordinal-level assignment success is similar between the two datasets. Somewhat more than twice as many Sanger-OTUs as 454-OTUs are assigned with posterior probability ≥ 80% to family, genus, and species. Two of the 973 454-OTUs at Step 6 (Table 3) were assigned only to the level of Arthropoda, one was assigned to Collembola, and one was assigned to Amphipoda.

SAP 80% posterior probability

			% Identified to			
Class		OTU count	Order	Family	Genus	Species
Insecta						
	Sanger	541	99%	37%	36%	35%
	454	951	96%	17%	16%	16%
Arachnida						
	Sanger	5	100%	80%	80%	80%
	454	12	92%	33%	33%	33%
Clitellata (Annelida)						
	Sanger	1	100%	0%	0%	0%
	454	6	100%	0%	0%	0%
Total						
	Sanger	547	98%	36%	35%	35%
	454	969	96%	17%	16%	16%

Table 7. Estimation of beta diversity. Unweighted, Unifrac dissimilarity matrices from the Sanger (lower) and 454 (upper) datasets. A pair of corresponding cells is indicated by the two boxes. The two matrices are highly significantly correlated (Mantel, 9999 permutations, p < 0.001).

		Unweighted, Unifrac distance matrix from 454						
		1H1X	2H1K	2K1X	5K1X	HONGHE	KMG	XSBN
Unweighted, Unifrac distance matrix from groundtruth	1H1X		0.939	0.860	0.862	0.707	0.928	0.746
	2H1K	0.959		0.896	0.898	0.603	0.743	0.941
	2K1X	0.895	0.906		0.395	0.939	0.626	0.812
	5K1X	0.896	0.899	0.181		0.935	0.615	0.887
	HONGHE	0.690	0.505	0.944	0.941		0.920	0.947
	KMG	0.952	0.756	0.500	0.431	0.937		0.946
	XSBN	0.690	0.958	0.798	0.893	0.961	0.946	

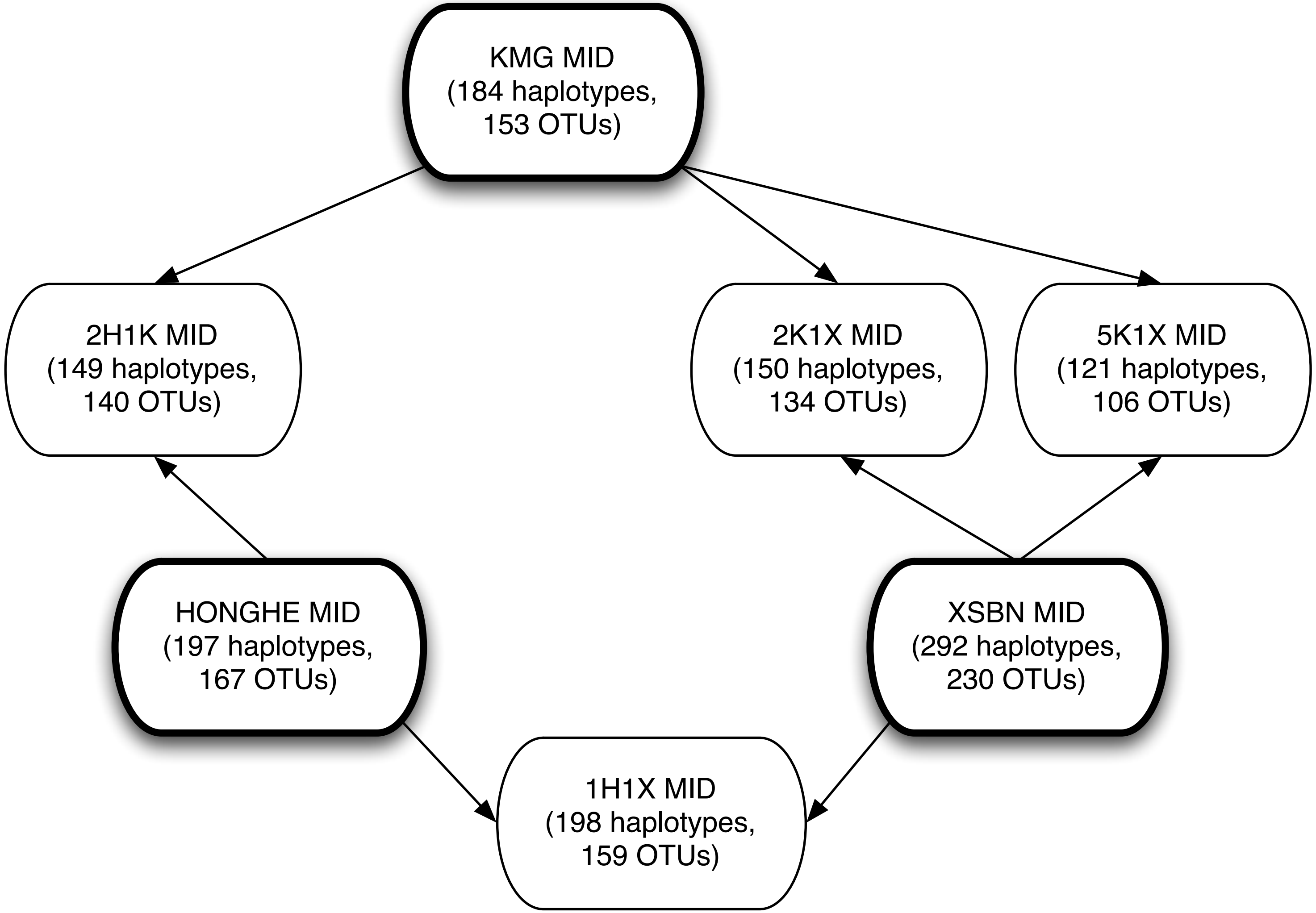

Figure 1. Schematic relationship of the four mixture communities and the three source communities (bold outline). For brevity, the communities are referred to as MIDs in the text (Multiplex IDentifiers). OTUs represent 98%-similarity clusters of haplotypes.

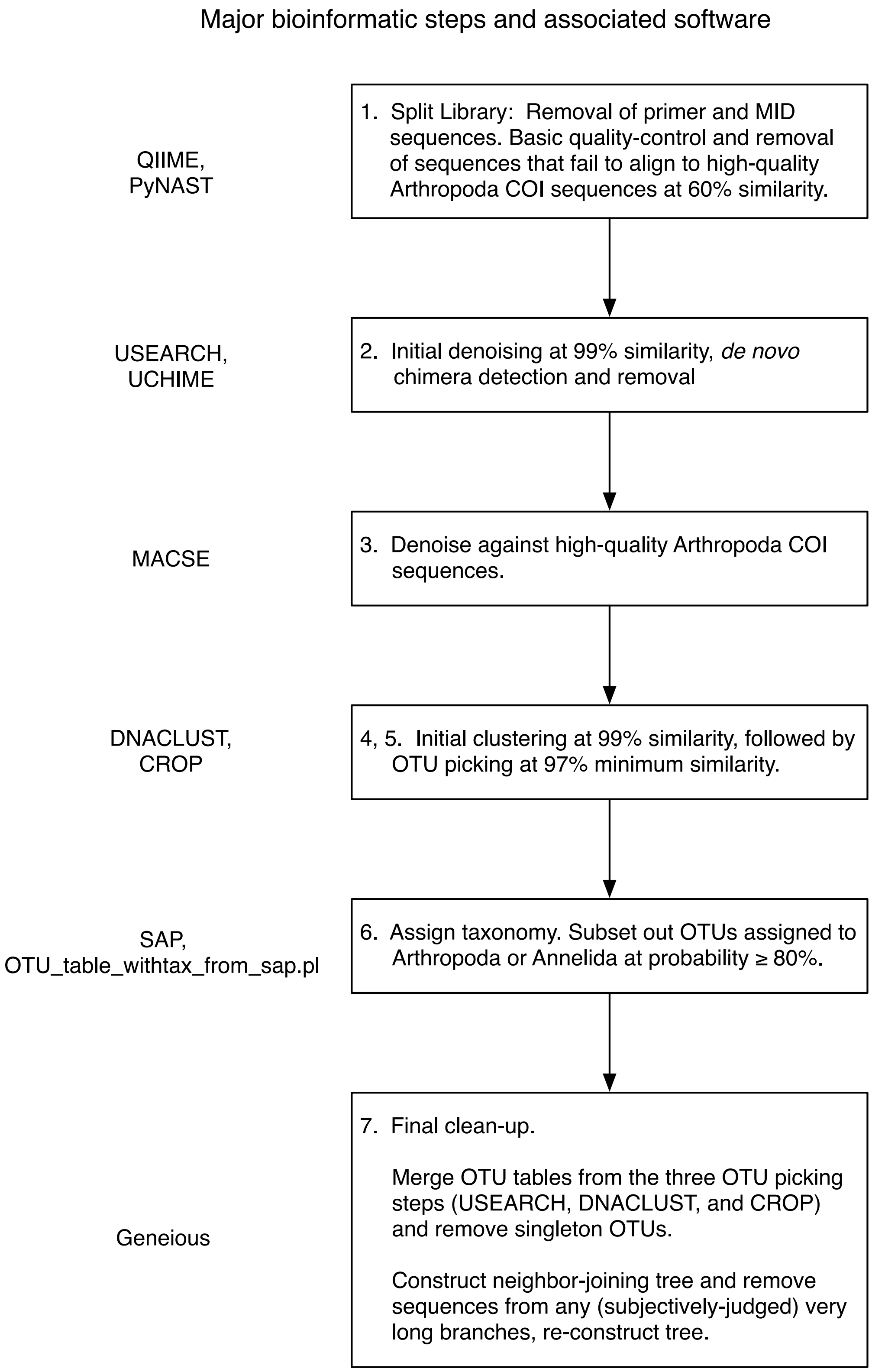
Major bioinformatic steps and associated software
QIIME, PyNAST
1. Split Library: Removal of primer and MID sequences. Basic quality-control and removal of sequences that fail to align to high-quality Arthropoda COI sequences at 60% similarity.
USEARCH, UCHIME
2. Initial denoising at 99% similarity, *de novo* chimera detection and removal
MACSE
3. Denoise against high-quality Arthropoda COI sequences.
DNACLUST, CROP
4, 5. Initial clustering at 99% similarity, followed by OTU picking at 97% minimum similarity.
SAP, OTU_table_withtax_from_sap.pl
6. Assign taxonomy. Subset out OTUs assigned to Arthropoda or Annelida at probability ≥ 80%.
Geneious
7. Final clean-up.
Merge OTU tables from the three OTU picking steps (USEARCH, DNACLUST, and CROP) and remove singleton OTUs.
Construct neighbor-joining tree and remove sequences from any (subjectively-judged) very long branches, re-construct tree.

Figure 2. Schematic of the major bioinformatic steps (numbering corresponds to Table 3), with associated software packages and pipelines.

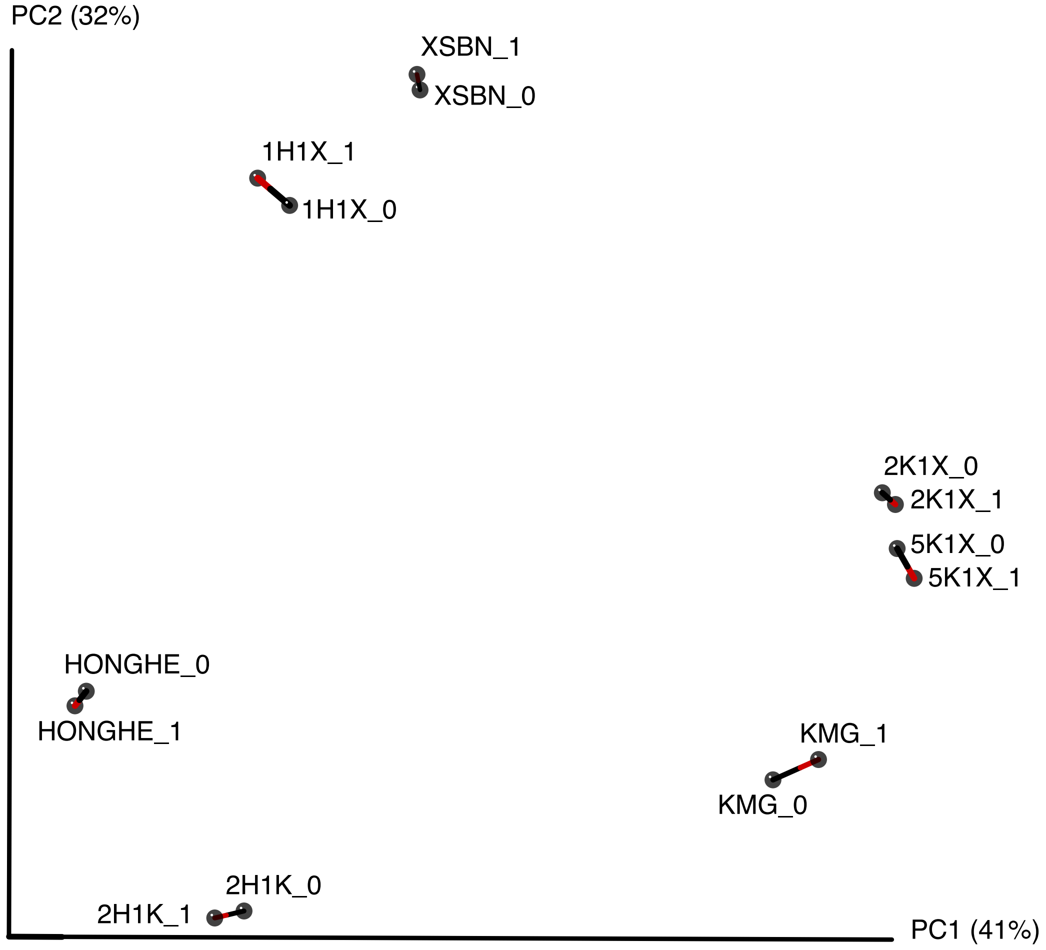

Figure 3. Estimation of beta diversity. Highly significant correspondence between Principal Coordinates ordinations of the two unweighted Unifrac dissimilarity matrices in Table 7 (Procrustes, 9999 permutations, p < 0.001). "0" indicates the 454 dataset, and "1" indicates the Sanger dataset. The Procrustes analysis was run on the first three PCs, but we show the first two PCs for clarity. Note that the mixture MIDs (e.g. 1H1X) lie between the corresponding source MIDs (HONGHE and XSBN), as expected (Fig. 1).

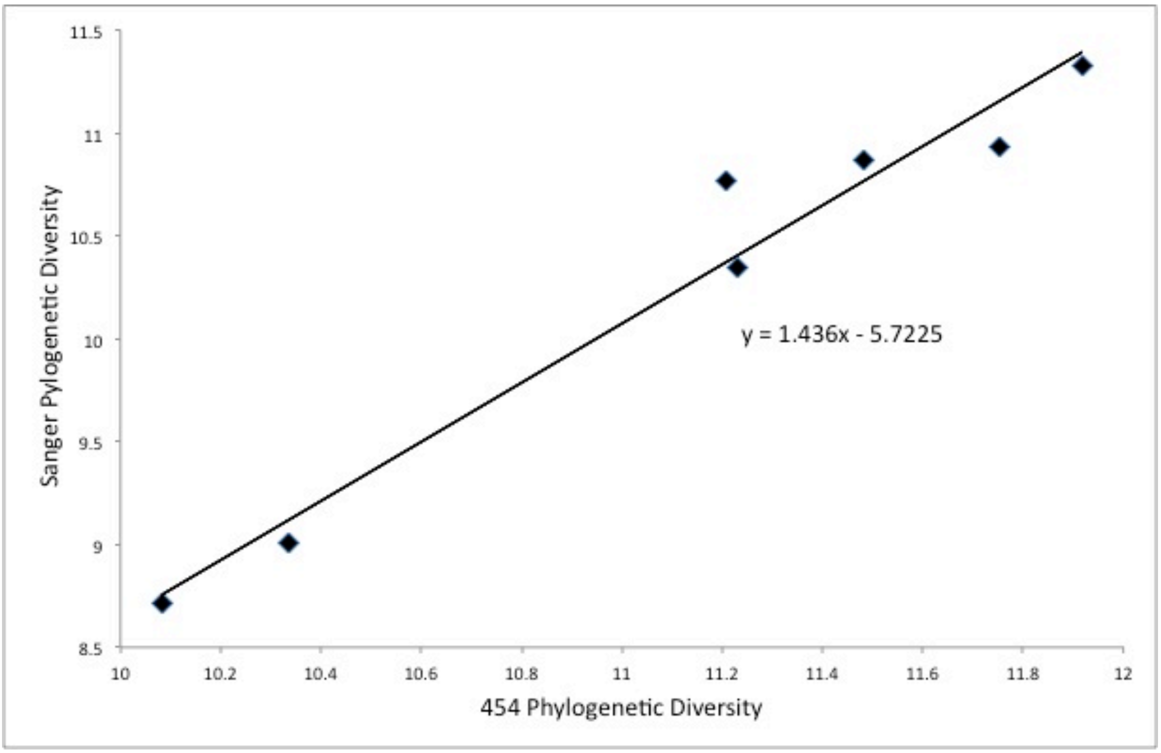

Figure 4. Estimation of alpha diversity. Local phylogenetic diversity (PD) is estimated using Phylocurve.R rarefaction (Nipperess 2011a) for each of the seven MIDs in both the 454 and Sanger datasets. Sanger PD is significantly predicted by 454 PD (linear regression, $F_{1,4}$=76.2, p<0.001, R^2=0.95). PD is estimated at a sample size of 101 OTUs because the 5K1X MID has only 106 Sanger-OTUs (Figure 1); the relationship holds for higher numbers of OTUs if 5K1X is omitted (data not shown).

Table S1. Allelic dropout, subdivided by MID and taxon.

	HongHe											
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Odonata	Plecoptera	Psocoptera	Trichoptera	Arachnida	Annelida	Total
Total input haplotypes	15	86	9	28	19	1	2	3	2	1	1	167
≥1-read OTUs	13	76	7	13	14	1	2	3	2	1	1	133
≥2-read OTUs	12	74	6	7	14	1	2	3	1	1	1	122
≥5-read OTUs	9	61	5	7	9	1	2	3	1	1	1	100
% Total input haplotypes												
≥1-read OTUs	87%	88%	78%	46%	74%	100%	100%	100%	100%	100%	100%	80%
≥2-read OTUs	80%	86%	67%	25%	74%	100%	100%	100%	50%	100%	100%	73%
≥5-read OTUs	60%	71%	56%	25%	47%	100%	100%	100%	50%	100%	100%	60%

	KMG								
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Psocoptera	Arachnida	Ephemeroptera	Total
Total input haplotypes	12	57	24	26	25	4	4	1	152
≥1-read OTUs	11	50	23	14	22	4	3	1	127
≥2-read OTUs	9	46	23	10	20	4	2	1	114
≥5-read OTUs	9	40	21	5	17	4	1	1	97
% Total input haplotypes									
≥1-read OTUs	92%	88%	96%	54%	88%	100%	75%	100%	84%
≥2-read OTUs	75%	81%	96%	38%	80%	100%	50%	100%	75%
≥5-read OTUs	75%	70%	88%	19%	68%	100%	25%	100%	64%

	XSBN						
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Blattaria	Total
Total input haplotypes	12	28	5	54	129	2	230
≥1-read OTUs	10	22	2	31	89	2	156
≥2-read OTUs	8	20	2	26	85	2	143
≥5-read OTUs	4	13	1	11	65	2	96
% Total input haplotypes							
≥1-read OTUs	83%	79%	40%	57%	69%	100%	68%
≥2-read OTUs	67%	71%	40%	48%	66%	100%	62%
≥5-read OTUs	33%	46%	20%	20%	50%	100%	42%

	1H1X									
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Blattaria	Odonata	Plecoptera	Trichoptera	Total
Total input haplotypes	3	67	3	22	59	1	1	2	1	159
≥1-read OTUs	3	49	3	13	34	1	1	2	1	107
≥2-read OTUs	3	43	3	13	31	1	1	2	1	98
≥5-read OTUs	3	34	3	11	29	1	1	1	1	84
% Total input haplotypes										
≥1-read OTUs	100%	73%	100%	59%	58%	100%	100%	100%	100%	67%
≥2-read OTUs	100%	64%	100%	59%	53%	100%	100%	100%	100%	62%
≥5-read OTUs	100%	51%	100%	50%	49%	100%	100%	50%	100%	53%

	2H1K									
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Psocoptera	Trichoptera	Arachnida	Annelida	Total
Total input haplotypes	18	43	16	35	21	4	1	1	1	140
≥1-read OTUs	16	41	14	20	19	4	1	1	1	117
≥2-read OTUs	15	36	12	16	18	4	1	1	1	104
≥5-read OTUs	14	31	11	11	17	4	1	1	1	91
% Total input haplotypes										
≥1-read OTUs	89%	95%	88%	57%	90%	100%	100%	100%	100%	84%
≥2-read OTUs	83%	84%	75%	46%	86%	100%	100%	100%	100%	74%
≥5-read OTUs	78%	72%	69%	31%	81%	100%	100%	100%	100%	65%

	2K1X								
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Blattaria	Ephemeroptera	Arachnida	Total
Total input haplotypes	11	24	16	29	48	1	1	4	134
≥1-read OTUs	8	16	15	11	37	1	1	1	90
≥2-read OTUs	8	15	14	9	35	1	1	1	84
≥5-read OTUs	8	11	7	8	30	1	1	0	66
% Total input haplotypes									
≥1-read OTUs	73%	67%	94%	38%	77%	100%	100%	25%	67%
≥2-read OTUs	73%	63%	88%	31%	73%	100%	100%	25%	63%
≥5-read OTUs	73%	46%	44%	28%	63%	100%	100%	0%	49%

	5K1X							
Taxa	Coleoptera	Diptera	Hemiptera	Hymenoptera	Lepidoptera	Ephemeroptera	Arachnida	Total
Total input haplotypes	8	22	15	21	35	1	4	106
≥1-read OTUs	5	16	13	6	25	1	1	67
≥2-read OTUs	5	12	12	5	22	1	0	57
≥5-read OTUs	3	10	7	2	21	1	0	44
% Total input haplotypes								
≥1-read OTUs	63%	73%	87%	29%	71%	100%	25%	63%
≥2-read OTUs	63%	55%	80%	24%	63%	100%	0%	54%
≥5-read OTUs	38%	45%	47%	10%	60%	100%	0%	42%

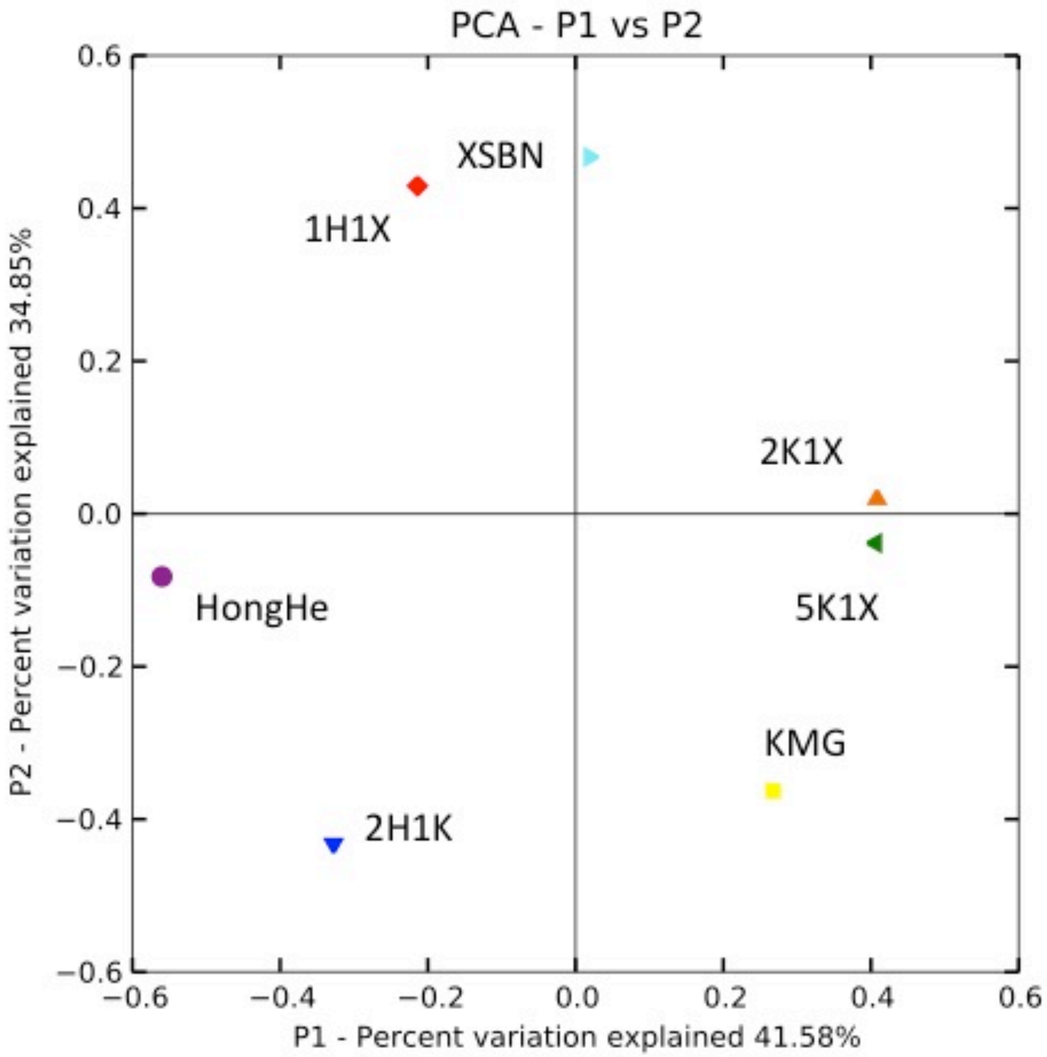

Figure S1. Principal coordinates ordination of the 454 dataset, using the 1-Sørensen-DICE (presence-absence) measure of compositional dissimilarity. MID names indicate source communities (HONGHE, KMG, XSBN) and mixture communities (1H1X, 2H1K, 2K1X, 5K1X).

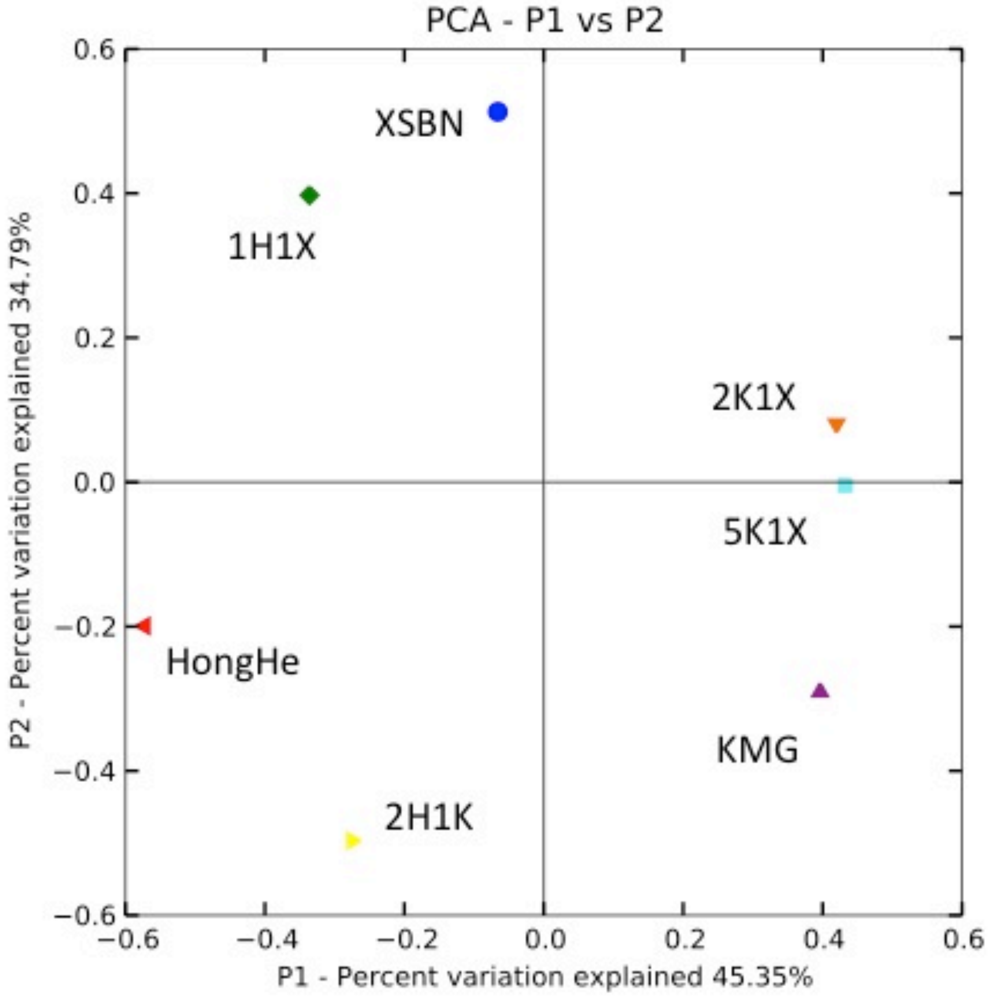

Figure S2. Principal coordinates ordination of the Sanger dataset, using the Sørensen-DICE (presence-absence) measure of compositional dissimilarity. MID names indicate source communities (HONGHE, KMG, XSBN) and mixture communities (1H1X, 2H1K, 2K1X, 5K1X). Note that the mixture communities lie between the respective source communities and that Figures S2 and S3 are very similar, after rotation (Use the labels, not the symbols to match MIDs).