ReceivedDate:13-May-2014

RevisedDate:26-Jan-2015

AcceptedDate:23-Feb-2015

ArticleType : Research Article

Editor : Douglas Yu

Title: A DNA Barcoding system integrating multi-gene sequence data

Douglas Chesters[1], Wei-Min Zheng[2], Chao-Dong Zhu[1]

[1] Key Laboratory of Zoological Systematics and Evolution (CAS), Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China.

[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China.

Corresponding author:

Professor Chao-Dong Zhu

Key Laboratory of Zoological Systematics and Evolution (CAS),

Institute of Zoology,

Chinese Academy of Sciences,

1 Beichen West Road, Chaoyang District,

Beijing 100101,

P. R. China.

+86-10-64807085

zhucd@ioz.ac.cn

Running Title: Multi-Gene DNA Barcoding

SUMMARY

1. A number of systems have been developed for taxonomic identification of DNA sequence data. However, in eukaryotes these systems are largely based on single predefined genes, and thus are vulnerable to biases from limited character sampling, and are not able to identify most sequences of genomic origin.

2. We here demonstrate an implementation for multi-gene DNA barcoding. First a reference framework is built of frequently sequenced loci. Query sequence data are then organized by excising sequences homologous to references and assigning species names where the level of sequence similarity between query and reference falls within the (gene-appropriate) level of intra-specific variation usually observed. The approach is compared to some existing methods including 'bagpipe_phylo', a re-implementation for taxonomic assignment on phylogenies.

3. 78% of the species and 94% of the genera known to be present in arthropod test queries were correctly inferred by the proposed multi-gene system. Most critically, the rate of species identification was increased over using a COI-only approach. 24% of species in the queries were found only in non-COI genes, with no clear reduction in the accuracy of species assignment at many of these other loci. Similarly, additional species assignments were made for a pooled metagenomic dataset by using non-COI columns. On a smaller query dataset of 273 bee sequences, the accuracy of species assignment using modified calculation of

distances was indistinguishable from phylogeny-based taxonomic identification.

4. Standardized single fragment DNA barcoding remains an invaluable tool in species identification for PCR-generated sequence data. The approach developed here supplements the established species dense DNA barcode backbone with other genomic data, reducing error via integration of independent genetic loci, and permitting additional identifications for non-barcode fragments. The latter will be particularly relevant in monitoring of community genomics using next generation sequencing platforms.

KEYWORDS: metagenomics, biodiversity monitoring, species clustering

INTRODUCTION

Taxonomic organization of DNA sequences is an essential component in much of ecology and evolutionary research, including in the description of biodiversity. A DNA-based system for taxonomic profiling of communities emerged firstly during study of microbial diversity (Stackebrandt and Goebel 1994), and has since become standard, with wide availability of comprehensive reference libraries for the universal 16S-rRNA gene in particular (Maidak et al. 2001; Pruesse et al. 2007). DNA-based characterization of eukaryote communities is now also well established, with single gene systems in fungi (O'Brien et al. 2005) and animals (Hebert et al. 2003) and a two-locus barcoding system adopted in plants (CBOL Plant Working Group 2009).

The bioinformatics tools for analysis of community DNA data are diverse, variously performing tasks such as sequence processing (often platform specific), OTU clustering,

taxonomic assignment and ecological comparisons (for a review focused on eukaryote community tools, see Bik et al. 2012). Computational description of diversity in a sample of query DNA data can take the form of OTU clusters (e.g. 'Esprit', Sun et al. 2009), assignment of taxonomies via comparison to members of a reference library ('Megan', Huson et al. 2007), or a combination of these (the 'jMOTU'/' Taxonerator' companion software, Jones et al. 2011). Further, these may be carried out within a phenetic ('Marta', Horton et al. 2010) or phylogenetic framework ('pplacer', Matsen et al. 2010).

The tradition of community profiling with PCR-based sequencing is reflected in the predominance of methods and tools designed around the single-gene. However, the sequencing of genomic DNA directly obtained from communities in the environment (Venter et al. 2004), coupled with increased use of high-throughput technology has necessitated multi-gene taxonomic assignment (Segata et al. 2012; Mende et al. 2013). Again these techniques have been adapted from microbes to community monitoring of eukaryotes (Hajibabaei et al. 2011; Yu et al. 2012; Ji et al. 2013), which includes non-targeted sequencing of genomic data (Timmermans et al. 2010; Nock et al. 2011; Taberlet et al. 2012; Zhou et al. 2013). However, the systems for taxonomic organization of *genomic* DNA from eukaryote communities have been slow to appear.

Here we implement a DNA barcoding system based around the widely used Blast (Camacho et al. 2009) and Usearch (Edgar 2010) tools, but integrating multi-gene arthropod sequence data. The protocol is implemented as a Linux pipeline in which queries can be automatically assigned with little user input. The pipeline consists of a set of scripts which carry out specific steps, free to be used as designed or adapted for other applications.

METHODS

Overview

Figure 1 outlines the approach for multi-gene DNA barcoding. Query sequences (Fig. 1a) are present of variously overlapping species and genes. A small number of sequences represent each gene in the references (Fig. 1b, with each of four genes depicted by differently colored blocks). Using this set of gene representatives, homologous sequences in the queries can be efficiently identified, excised and organized to columns (Fig. 1c, colored segments denote regions in the queries which are homologous to references). A set of references with all available species and genes (Fig. 1d) permits species level assignments in queries (Fig. 1e, with hits to two rows/species in the example).

Building a Reference Framework from Mined Data (Fig. 1b, d)

A framework was built representing both species and gene diversity in publically available Arthropod data, broadly using approaches outlined earlier (Peters et al. 2011; Papadopoulou et al. 2014; Chesters and Zhu 2014). The NCBI invertebrate release (gbinv*.seq.gz) was downloaded from ftp.ncbi.nih.gov/genbank/ and the taxonomy database (taxdump.tar.gz) from ftp.ncbi.nlm.nih.gov/pub/taxonomy/ (the December 2013 release, downloaded on 12th January 2014). All arthropod sequences were extracted from the files, then species-level labels were parsed from the taxonomy database and assigned to sequence entries via the NCBI taxon identifiers (script parse_ncbi_tax_database.pl). The file of arthropod sequences was processed to remove redundant sequences (those both identical and labeled with the

same species name) using Usearch v4.2.66 (Edgar 2010).

Sequences were grouped into genes using an approach modified from Peters et al. (2011) (script multiple_sequence_splitter.pl). A hash of gene synonyms was developed for standardization of gene names (supplementary file SF1), then sequence entries were split and grouped into files according to gene label. The set of reference genes produced by this were processed in several ways. First, genes sparsely sampled at the species level were omitted (discarding genes containing species numbering <1% of all species found in the database). Next, sequences for each gene were oriented and filtered by first finding a most representative sequence (MRS, the entry with the greatest aligned length against all other tested entries), then using this as a seed for orientation via Blast alignment (scripts most_representative_seq.pl, orient_sequences.pl). This step highlighted problematic sequence sets (globally dissimilar), which were omitted from further analysis. Finally, mislabeled and chimerical sequences were identified also using the MRS. To achieve this, for each gene file we used as Blast queries the MRS of the other loci, where hits >150 bases were presumed mislabeled and removed. In addition to organization of the references, parameters appropriate for species level clustering of each gene (the linkage clustering threshold that results in sequence clusters most closely resembling species) were inferred as described in Chesters and Zhu (2014). The optimal parameter was selected as that with the highest taxonomic congruence, to be used for later query assignments.

Taxonomic Assignments for Query Sequence Data (Fig. 1)

User supplied query sequences were assigned gene and taxonomy via the references

(pseudocode for this proposed multi-gene barcoding approach is given in supplementary Fig. 1 and a text file containing all commands can be found in supplementary materials). In order to avoid unnecessary computations when processing multi-gene data, queries were placed in two steps. As assignment of queries to gene does not require exhaustive comparison to all references, an initial Blast search is performed (on all queries) against a small number of sampled reference sequences for each gene ('gene representatives'; Fig. 1b; using the script sample_db.pl). The Blast settings at this step were liberal, as to include sequences homologous although distantly related; word size 12, percent identity 25, e-value 1e-6. The Blast output for each gene was parsed, with hits invoking Blastdbcmd (Camacho et al. 2009) to excise the homologous subsequence in the queries and place them into gene-specific files (script parse_hits.pl, Fig. 1c).

In the second step, queries were compared to homologous sequences for all available species on a gene by gene basis (Fig. 1d, e), with attempts to assign taxonomies. For each gene we performed a species-level search using the Usearch software. Under default settings the search is terminated after only a small number of successful hits (via command line setting '-maxaccepts 20'), since any one hit is in principle a successful species assignment, while we ensured the whole database would be searched where hits were sparse (setting '-maxrejects' to outnumber all references). The gene-specific results were parsed for hits within species-level thresholds (as determined in the previous section, although generic thresholds can be used where these are not available, and liberal settings where broad-level taxonomies are acceptable). Where there was more than a single hit within the threshold, either the single best hit was selected (that with the highest percent identity), or a consensus

taxon was inferred. In addition to using Uclust calculation of similarities between references and queries, for smaller datasets we tested the similarity score described in Papadopoulou et al. (2014); pairwise global Needleman-Wunsch (NW) alignment using Emboss v6.3.1 (Rice et al. 2000), then percent identities calculated scoring indels as single events and ignoring terminal gaps and positions with ambiguous characters (script nw_align_blast_hits.pl). As with Uclust, hits outside the clustering thresholds were discarded. Species clustering between queries and references was repeated for all genes using the similarity threshold appropriate for each, then results from each gene integrated (Fig. 1e; script integrate_taxonomic_assignments.pl) . This script primarily extracts all species inferred for queries, and lists these in an output file along with the genes found for each. An additional parameter for this stage is the length of alignment below which a hit is deemed spurious. To account for difference in length of genes, we calculated the average length of each gene from the reference data, then set a lower limit on the length of alignment permitted below which the hit was dismissed. Different permitted lengths were tested.

Formation of New MOTU for Unassigned Sequences

While less amenable to the production of a simple-to-use barcoding tool (and thus not included in main pipeline), we examined means by which queries that could not be assigned at the species level might be organized into new MOTU. For each gene, query sequences unassigned to reference species were extracted and all-against-all Blast carried out at 95% identity. Hits were then aligned by NW, percent identities calculated, then hierarchical clustering performed in Esprit (Sun et al. 2009) under gene specific thresholds, each as

described in previous sections. Congruence between MOTU produced at different genes was examined, although strategies for formation of integrated MOTU from these have been discussed elsewhere (Chesters and Vogler 2013; Sunagawa et al. 2013)

Testing

Three datasets were used as test queries. First we selected species annotated sequences typical of Genbank DNA uploads ('QD1'; query dataset 1), from an arbitrary number (all from the month of February 2014) of the NCBI daily releases. For the second test dataset ('QD2') we obtained a metagenomic dataset of a pooled arthropod sample from a subtropical region in China (Zhou et al. 2013). We used the assembled formal sample (published under the file name k61dR2_k45.scafseq) which are consensus sequences of short reads, assembled and clustered where >=98% in similarity. Finally we used a set of newly generated sequences from a survey of bee diversity in tropical Xishuangbanna, south China (Xiuwei Liu et al., manuscript in preparation). This dataset ('QD3') was composed of 273 sequences in total (148 28S-rRNA, 104 COI and 20 CytB) from 157 specimens, with morphological identification to the level of genus or species. The new bee sequences were combined with the corresponding (28S-rRNA, COI, CytB) reference data (bee sequences taken from the insect data described in *Building a Reference Framework from Mined Data*), aligned with muscle (Edgar 2004), and a phylogeny inferred using RAxML v7.2.8 (Stamatakis 2006) under the gene partitioned GTRCAT model.

The multi-gene barcoding method described herein was compared to Claident (Tanabe and Toju 2013), RAxML EPA (Berger et al. 2011), and a new implementation of a key

component of the BAGpipe system (Papadopoulou et al. 2014). We rewrote the phylogeny-based taxonomic assignment routine of the latter as a standalone program (bagpipe_phylo.pl, made freely available, see supplement). Originally separate components of a lengthy pipeline, the code for reading the taxonomic hierarchy from NCBI and the code for taxonomic assignment on a phylogeny (specifically, the scripts parse_ncbi_tax_database.pl, see Fig. 1a of Papadopoulou et al. 2014, and parse_clades.pl, Fig. 2f, similarly) were developed into a single tool. The input requirements of the new program are a single rooted phylogeny and the NCBI taxonomic database as downloaded from NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/). The phylogeny required is one containing both references and queries, with references indicated by standard binomial labels. The software is run as described originally, with taxon names assigned to each query according to the shared taxonomy of the reference clade to which it is placed in the tree. Although we newly implement a feature for indicating reliability of assignments according to distance of query to nearest reference leaf.

RESULTS

Building a Reference Framework from Mined Arthropod Data

After initial processing and removal of sequences without complete species labels, the database contained 771,332 entries. The database was split by gene name. After assessment of the primary ~20 name based partitions, one was omitted due to difficulties in alignment. The name assignment for this gene was Large Subunit Ribosomal RNA, an ambiguous label encompassing 5S, 5.8S and 28S (while not containing these three labels in the annotation of

these entries). The percentage of apparently mislabeled/chimerical sequences identified and removed was very low (0.11%), although in the context of multi-gene barcoding the stringent partitioning of homologous sequences is critical. This is in order to prevent duplicating sequence diversity and overestimating the number of species.

Each of the 19 genes was clustered at the level of species. Suitable clustering parameters were inferred on reference data for later application in species assignment of query data. Statistics in relation to the reference data are given in supplementary table 1.

Multi-Gene Taxonomic Assignment in the Test Datasets

Multi-gene barcoding was tested on two mined datasets, the results of which are summarized in table 1. The first (QD1) was composed of sequences with full taxonomic annotation obtained from mining the NCBI daily release files (selected from the month of February 2014) downloaded from ftp://ftp.ncbi.nih.gov/genbank/daily-nc/. After removal of identical sequences and sequences with accessions also contained in the reference framework (recently updated entries), 9523 sequences remained (from 1848 named species and 282 different arthropod families). Accession information for QD1 is given in supplementary file 2. Queries were assigned to genes by comparison to the gene representatives. Since the approach used here for mining genes aims to return those most widely used (Chesters and Zhu 2014), it is expected that assignment of queries to genes would be near-complete (in contrast to assignment to species, which would be sensitive to taxonomic completeness of references). The number of sequences required to represent any particular gene in order to capture homologous queries was modest, with little benefit of using many. For example of the 9523

QD1 sequences, 78.5% (91.9% of species contained in queries) could be assigned to columns using 60 reference sequences per gene, and 79.5% (92.6% of species) using 600 per reference gene. Further, a feature on this implementation is automatic splitting of a single query sequence where it contains sub-sequences homologous to different reference genes (for instance a query sequence spanning much of the mitochondrial genome would be split to the widely used fragments of COI, 16S-rRNA, CytB etc.). This applied to 3.2% of sequences of the 9523 query set used here.

With QD1 queries assigned to genes, we performed a search of query subsequences against the full complement of references, on a gene by gene basis under the species appropriate thresholds. Blast was tested for this purpose (reducing the command line parameter '–max_target_seqs' to 20) but was prohibitively slow for COI, in which ~4000 sequences were queried against an Arthropod database of ~220000. In the Usearch analysis, 27.9% (2657) of the 9523 query sequences (and 44.4% of the 1848 query species) could be assigned at the species level. The rate of 'species description' for queries was highest but not limited to COI, with 75.8% being hits to the COI column, 32.5% to non-COI columns (some of which also hit the COI column), and 24.2% that were hits only to a non-COI column.

More pertinent than the absolute number of queries assigned a species (which is determined by completeness of reference data in addition to search settings) is the accuracy. In QD1 this can be inferred based on the match of species names originally given to queries and to those assigned in the current analysis (via sequence similarity). Of the 2657 queries assigned a species name, 77.6 +/-1.6 % had matching ('correct') species assignment and 94.0 +/-0.9% had matching genus assignment. The rate of accuracy in species assignment differed

substantially across loci and taxonomic group (Fig. 2). Notably, the rate of accurate species identification was only 12.5 +/-6.4% for 18S, and 56.5 +/-8.1% in the Orthoptera. Many of the errors in species assignment were caused by spurious short alignments between query and references. Therefore, simply applying a lower bound for the aligned sequence length between the query and the candidate reference (actually, the proportion of the average length of the two sequences) increases the accuracy of species assignment. For example, an aligned proportion of 12% or 60% gives a rate of accurate species assignment of 77.0 +/-1.6, 80.6 +/-1.7%, respectively. Accuracy could be further increased to 85.1 +/-1.6% by making only unambiguous species assignments, i.e. those in which a query was associated only with a single reference species within the given threshold. The current implementation assigns higher-level taxonomies in such cases. Naturally, increasing stringency in these ways reduced the number of queries that could be assigned.

Formation of MOTU for queries unassigned at the species level

Queries that could not be assigned at the species level were clustered into MOTU. For simplicity we focused on the predominant gene (COI in the current case), for which 965 MOTU were inferred. This was a ~10percent overestimation of the actual number of species in the file; 869. Clustering was also performed for other genes in order to test for congruence in MOTU in the overlapping members. The number of MOTU inferred in the COI queries was the same as found in the remaining genes, although in the current dataset this was based on a limited number of overlapping members (75); mostly derived from whole mitochondrial sequences.

The approach was also tested on a second mined dataset (QD2), a processed file of 561120 sequences from 73 pooled arthropod individuals. While reported to contain 37 MOTU based on morphotypes of input specimens and analysis of fragments spanning COI, extra were suspected, perhaps derived from gut content, residual tissue or mixed DNA (Zhou et al. 2013). In the analysis here, 12 sequences could be assigned directly to the references (table 1 column QD2, and table 2), representing 8 species units. The column assignments numbered 5 (COI), 2 (28S), 1 (16S), 2 (18S), 1 (CytB). Sequences assigned only to column were further grouped into MOTU for the COI gene. Ignoring a single known contaminant sequence (C2058678), the COI sequences comprised 35 MOTU, giving 43 species units in total, with 5 of the 8 assignments to genes other than COI (table 2 columns 3, 4, 6).

Comparison to other methods

For QD1 we compared the multi-gene barcoding to Claident (Tanabe and Toju 2013), being representative of eukaryote taxonomic software and not optimized for multi-gene data. The arthropod references were combined to a single file and input into Claident, then queried with the QD1 set using both the QC and NNC algorithms. The exhaustive Blast of queries to all references gave a running time orders of magnitude greater than the approach proposed herein (~48hr compared to 1.5hr for the latter). Further, no species assignments could be made (likely since this software favors more complete reference data), although 2476 of the queries could be assigned genus names at a greater accuracy (96.6%) to that herein (94.0%).

Finally, a 'mid-sized' aligned dataset (QD3; three genes, 1145 mined reference species and 273 new query sequences, subject to independent morphological identification by

taxonomic experts) permitted comparison of the phenetic method proposed herein, to phylogeny-based identification. For the latter we used EPA (Berger et al. 2011) and the new bagpipe_phylo implementation. Using our multi-gene barcoding, species assignments were conservative with the default Uclust scores of sequence similarity, with only half of the specimens (26 of the 46 which had a species name also contained in the references) given species labels. However making species matches instead using NW alignments and more appropriate gap treatments (as described in Papadopoulou et al. 2014) led to 44 correct species assignments out of 46, a rate identical to the phylogenetic methods EPA and bagpipe_phylo.

DISCUSSION

Herein we describe a method for species level organization of multi-gene data. This utilizes a reference framework consisting simply of a set of genes, each containing sequence data of known taxonomic origins. Approaches for building frameworks from mined sequence data have been described previously (e.g. McMahon and Sanderson 2006; Peters et al. 2011), although Chesters and Zhu (2014) attempt to maximize the return of species dense homologs in particular, and in the process infer parameters broadly appropriate for species organization. We here demonstrate the set of steps necessary for utilizing this species/gene framework for the purpose of multi-gene DNA barcoding. The approach is efficient and scalable, with a number points described at which the process can be tuned depending on context and requirements.

With the framework built, query sequences homologous to any of the reference genes can be assigned and grouped according to the level of sequence variation appropriate for the evolutionary characteristics of that locus. Species clustering and assignment to references is commonly performed for single locus data in eukaryotes. For example the jMOTU software computes distances between query pairs based on global NW alignment, which are used for clustering input members into MOTU according to a user defined cutoff, with MOTU then assigned taxonomic annotation where possible by a Blast search of reference data (Jones et al. 2011). Extending this procedure in order to include the genomic dimension (in other words adding additional genes to the usual DNA barcode marker) requires determining appropriate species clustering parameters for additional columns, but also how units might be integrated between columns. The former is achieved in an established way (Göker et al. 2009; Sauer and Hausdorf 2012; Mende et al. 2013); that is, for each gene select the parameters or method which produces the most reasonable species units. Here species clustering is performed on the percentage of identical sites between pairs of homologous sequences. While not sophisticated, the use of percent identity for the current purpose does not clearly underperform comparative to scores which are more so (Srivathsan and Meier 2012).

A two step implementation for species assignment (assign to gene representatives, then assign to species using the complete reference matrix) reduces computations substantially, both due to reducing the number of alignments made, and via allowing precise similarity cutoffs to be used when aligning each homolog. The placement of queries to a set of references in which multiple genes have been integrated (Fig. 1e) gives a species-level organization of queries in which genes are (therefore) also integrated. However, integrating

different genes can be more problematic for new MOTU (those built from queries that are not assigned to reference species). We describe simply forming MOTU of the predominant gene (usually COI in the animals) for queries not assigned species. In cases where multi-gene MOTU are required, the availability of sequence labels permits integration of MOTU from different genes (Chesters and Vogler 2013). In the absence of labels other features can assist in integrating loci. Where much of the data is of genomic origin, a contig that spans several genes can be used as a reference point to match individuals (and therefore MOTU) from different genes. Further, where using metagenomic samples, read abundances of a given MOTU are expected to covary across samples, and thus MOTU covarying between genes can be united as likely from the same genome (Sunagawa et al. 2013).

The utility of a multi-gene barcoding system is dependent on the structures both of the currently available reference data and the queries used. Barcode markers dominate sequence databases, although despite many favorable characteristics, few are truly universal (Kondo et al. 2009; Sun et al. 2012). For example 28S-rRNA is often favored by Hymenoptera workers, this gene having a higher rate of PCR success (Andersen and Mills 2012) and lower incidence of sequencing problems compared to COI (Li et al. 2010). While the lower substitution rate means less species level information content, the advantage of sequencing success has resulted in a great deal of 28S data for potential use in species identification. As the multi-gene reference framework demonstrated herein permits a consistent and objective species assignment criteria for any gene present and implicit means of integration, these allow seamless use of the standard barcode in addition to 28S (for example) and many other informative loci.

While the benefits of the continued focus on a standardized single-gene system for species identification are clear, we demonstrate that development of a multi-gene framework around the DNA barcode backbone is easily achieved and can increase the rate of species identification often at an indistinguishable rate of error. The potential exists for species dense genomic arthropod references in the near future, with 5000 insect genomes (i5k Consortium 2013), 1000 transcriptomes (1KITE; www.1kite.org) and 10000 mitochondrial genomes (X. Zhou pers. communication) undergoing sequencing. The reference framework built from these data will greatly improve our ability to organize genomic fragments from pooled samples, in turn enabling the next generation of metagenomic community studies (e.g. Davies et al. 2012).

ACKNOWLEDGMENTS

The authors would like to thank Xin Zhou, Shanlin Liu and Yiyuan Li, who were helpful in providing additional details of their analysis and published data (Zhou et al. 2013), and useful comments on the first draft of this manuscript. We would also like to thank Douglas Yu for valuable suggestions on phylogeny-based taxonomic assignment and editing later versions of this manuscript, and Xiuwei Liu, Qingyan Dai and Zeqing Niu for providing access to their preliminary dataset of tropical Chinese bees.

DATA ACCESSIBILITY

-Supplementary table 1 (Statistics for each gene during formation of the reference matrix); uploaded as online supporting information.

-Supplementary figure 1 (pseudocode); uploaded as online supporting information.

-Supplementary files 1 (gene synonyms), 2 (accessions for QD1 sequences), 3 (Linux pipeline); uploaded as online supporting information.

-Source code; the scalable phenetic system is made freely available under the GNU general public license at http://sourceforge.net/projects/multilocusbarcoding, and the redeveloped bagpipe_phylo at http://sourceforge.net/projects/bagpipe/files/.

- Bee DNA sequence data; Genbank accession numbers KP258999-KP259269.

FUNDING

This work was supported by the National Science Foundation of China (Grants No. 31172048, J1210002); the Key Laboratory of Zoological Systematics and Evolution, CAS (No. Y229YX5105); the Program of Ministry of Science and Technology of the People's Republic of China (2012FY111100); and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KSXC2-EW-B-02) to CDZ.

REFERENCES

-Andersen, J.C. & Mills, N.J. (2012) DNA extraction from museum specimens of parasitic Hymenoptera. *PloS One*, **7**, e45549.

-Berger, S.A., Krompass, D. & Stamatakis, A. (2011) Performance, accuracy, and Web server

for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, **60**, 291-302.

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

-Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics, **10**, 1-421.

-CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**, 12794-12797.

-Chesters, D. & Vogler A.P. (2013). Resolving Ambiguity of Species Limits and Concatenation in Multi-locus Sequence Data for the Construction of Phylogenetic Supermatrices. *Systematic Biology,* **62**, 456-466.

-Chesters, D. & Zhu, C.D. (2014) A Protocol for Species Delineation of Public DNA Databases, Applied to the Insecta. *Systematic Biology*, **63**, 712-725..

-Davies, N., Meyer, C., Gilbert, J.A., Amaral-Zettler, L., Deck, J., Bicak, M., Rocca-Serra, P., Assunta-Sansone, S., Willis, K. & Field, D. (2012) A call for an international network of genomic observatories (GOs). *GigaScience*, **1**, 5.

-Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* **32**, 1792-1797.

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

-Göker, M., García-Blázquez, G., Voglmayr, H., Tellería, M.T. & Martín, M.P. (2009)

Molecular taxonomy of phytopathogenic fungi: a case study in Peronospora. *PLoS One*, **4**, e6319.

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

-Hebert, P.D.N., Ratnasingham, S. & DeWaard, J.R. (2003) Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, **270**, S596–S599.

-Horton, M., Bodenhausen, N. & Bergelson, J. (2010) MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences. *Bioinformatics*, **26**, 568-9.

-Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377-86.

-i5K Consortium (2013) The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *Journal of Heredity*, **104**, 595-600.

-Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245-1257.

-Jones, M.O., Ghoorah, A. & Blaxter, M.L. (2011) jMOTU and Taxonerator: Turning DNA Barcode Sequences into Annotated Operational Taxonomic Units. *PLoS One*, **6**, e19259.

-Kondo, T., Gullan, P.J. & Williams, D.J. (2009) Coccidology. The study of scale insects

(Hemiptera: Sternorrhyncha: Coccoidea). *Revista Corpoica–Ciencia y Tecnología Agropecuaria*, **9**, 55-61.

-Li, Y., Zhou, X., Feng, G., Hu, H., Niu, L., Hebert, P.D.N. & Huang, D. (2010) COI and ITS2 sequences delimit species, reveal cryptic taxa and host specificity of fig-associated Sycophila (Hymenoptera, Eurytomidae). Molecular Ecology Resources, **10**, 31-40.

-Maidak, B.L., Cole, J.R., Lilburn, T.G., Parker Jr., C.T., Saxman, P.R., Farris, R.J., Garrity, G.M., Olsen, G.J., Schmidt, T.M. & Tiedje, J.M. (2001) The RDP-II (ribosomal database project). *Nucleic Acids Research*, **29**, 173-174.

-Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, **11**, 538.

- McMahon, M.M., Sanderson, M.J. (2006) Phylogenetic Supermatrix Analysis of GenBank Sequences from 2228 Papilionoid Legumes. *Systematic Biology*, **55**, 818-836.

-Mende, D.R., Sunagawa, S., Zeller, G. & Bork, P. (2013) Accurate and universal delineation of prokaryotic species. *Nature Methods*, **10**, 881-884.

-Nock, C.J., Waters, D.L., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M. & Henry, R.J. (2011) Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal*, **9**, 328-333.

-O'Brien, H.E., Parrent, J.L., Jackson, J.A., Moncalvo, J.-M. & Vilgalys, R. (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology*, **71**, 5544–5550.

-Papadopoulou, A., Chesters, D., Coronado, I., De la Cadena, G., Cardoso, A., Reyes, J.C.,

Maes, J.M., Rueda, R.M. & Gómez-Zurita, J. (2014) Automated DNA-based plant identification for large-scale biodiversity assessment. *Molecular Ecology Resources,* **15**, 136-152.

-Peters, R.S., Meyer, B., Krogmann, L., Borner, J., Meusemann, K., Schutte, K., Niehuis, O. & Misof, B. (2011) The taming of an impossible child - a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biology*, **9**, 55.

-Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. & Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, **35**, 7188-7196.

-Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, **16**, 276-277.

-Sauer, J. & Hausdorf, B. (2012) A comparison of DNA-based methods for delimiting species in a Cretan land snail radiation reveals shortcomings of exclusively molecular taxonomy. *Cladistics*, **28**, 300-316.

-Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**, 811-814.

-Stackebrandt, E. & Goebel, B.M. (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, **44**, 846-849.

-Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics,* **22**, 2688-90.

-Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M.L., McKendree, W. & Farmerie, W. (2009) ESPRIT: Estimating Species Richness Using Large Collections of 16S rRNA Shotgun Sequences. *Nucleic Acids Research*, **37**, e76-e76.

-Sun, Y., Kupriyanova, E.K. & Qiu, J.W. (2012) COI barcoding of Hydroides: a road from impossible to difficult. *Invertebrate Systematics*, **26**, 539-547.

-Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R., Coelho, L.P., Arumugam, M., Tap, J., Nielsen, H.B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., de Vos, W.M., Wang, J., Li, J., Doré, J., Ehrlich, S.D., Stamatakis, A. & Bork, P. (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, **10**, 1196-1199.

-Srivathsan, A. & Meier, R. (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics*, **28**, 190-94.

-Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045-2050.

-Tanabe, A. S. & Toju, H. (2013) Two new computational methods for universal DNA barcoding: A benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. *PLoS ONE*, **8**, e76910.

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

-Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D.,

Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. & Smith, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.

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

- Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. & Huang, Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 1-4.

Table 1: Summary results for the Arthropod query datasets. 'Number of species inferred via hits to references' gives the number of named species assigned by the current method to the queries. Breakdown of these species counts are given for the predominant gene ('of which, via COI gene') and the remainder, where the latter includes cases where a given species is assigned to both a COI and non-COI query ('of which, via genes other than COI'), or where the species is assigned only to a non-COI sequence ('of which, only via genes other than COI'). 'Inference accuracy for species ; genus' is calculated on these subset of queries to which species were assigned. 'MOTU inferred for unassigned hits' is not part of the core protocol, although shown to give a proxy for number of species in sequences that are assigned to gene but not assigned a species name.

Mined query dataset		QD1; NCBI species labeled uploads	QD2; Arthropod metagenomic scaffolds
Number of query sequences		9523	561120
Presumed number of species in queries		1848	37
Number of species inferred via hits to references		820	8
	of which, via COI gene	75.8%	63%
	of which, via genes other than COI	32.5%	63%
	of which, only via genes other than COI	24.2%	25%
Inference accuracy for species ; genus		77.6 +/-1.6 ; 94.0 +/-0.9	NA
MOTU inferred for unassigned hits (COI-based)		965	35

Table 2: Pooled arthropod metagenomic sequences (QD2) assigned at the species level. First column (Query IDs) gives details of the 12 assigned query sequences, with sequence IDs (as used in the data published by Zhou et al. 2013) followed by the gene to which they were assigned. Remaining columns are the relevant reference genes, containing the accession and species information for the available sequence data. Entry 'N' indicates no available sequence data for the given gene/species.

Query IDs	Reference data to which queries are assigned				
	CO1	28S	16S	18S	CytB
scaffold6419 (COI)	Catoblemma semialba JN401205	N	N	N	N
scaffold1324 (COI)	Cerynea trogobasis KF395041	N	N	N	N
C1753954 (16S) C2051434 (18S)	Chironomus tepperi AF192211	Chironomus tepperi KC177658	Chironomus tepperi KC177440	Chironomus tepperi KC177280	Chironomus tepperi KC750582
C2054508 (COI)	Clanis surigaoensis JN677834	N	N	N	N
C2004271 (COI) C2067734 (CytB)	Culex sitiens DQ317598	N	N	N	Culex sitiens FJ025883
C2109518 (18S)	N	Nothodelphax gillettei DQ532594	N	Nothodelphax gillettei DQ532514	N
scaffold3608 (COI) C2036659 C2093220 (28S)	Sogatella furcifera AB572348	Sogatella furcifera HM017358	Sogatella furcifera JX556734	Sogatella furcifera JF773150	Sogatella furcifera JX556861
scaffold7745 (28S)	N	Tagosodes wallacei HM017380	N	Tagosodes wallacei HM017272	N

Figure 1. Species level organization of query genomic fragments, toy example. a) Query metagenome data. b) Reference data (subset). Differently colored blocks correspond to four different genes. This reference subset are used in c) for assignment of queries to gene. Regions of homology are identified (using Blast) for organization of queries to genes. d) All available reference data, used in e) for species level organization of each gene. In the example, the four sequences of the blue gene are clustered to a single species, the sequence of the green gene shows insufficient similarity to reference species so is omitted, and the two sequences of the yellow gene are inferred of two different species. In total two species are identified over all genes.

Figure 2. Breakdown of species assignments. Counts are from assignments to fully named species only. Bar height indicates sum species assignments. Lower section of each bar indicates number of 'accurate' species assignments. This is estimated according to the match between the names originally assigned to sequence submissions, and names assigned via the DNA based approach herein. Loci with > 30 observations are shown. A gives breakdown by locus, and B gives breakdown by taxonomy (rank of 'order' within the arthropods).

Supplementary Figure 1. Pseudocode for the core steps performed in species organization of query data, with commands, input and output files indicated. For each gene three commands are run, these first find homologous sequences and then perform DNA barcode-like species assignments. Blastn finds regions of broad similarity between queries and reference genes, then the parse_hits.pl Perl script uses this information for sequence extraction. Trim option 1 means hits are trimmed to the start and end positions of the longest hit (other options are implemented). Next Usearch aligns queries to references where within species level similarity. An additional Perl script is used to integrate species assignments from the different genes. Abbreviations; q = search queries, db = search database.