Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 961–970 doi: 10.1111/2041-210X.12104

Heuristic optimization for global species clustering of DNA sequence data from multiple loci

Douglas Chesters[1,2], Fang Yu[1], Huan-Xi Cao[1], Qing-Yan Dai[1], Qing-Tao Wu[1], Weifeng Shi[3], Weimin Zheng[4] and Chao-Dong Zhu[1,5]*

[1]*Key Laboratory of Zoological Systematics and Evolution (CAS), Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101, China;* [2]*Guangzhou Institute of Advanced Technology, Chinese Academy of Sciences, Guangzhou 511458, China;* [3]*School of Basic Medical Sciences, Taishan Medical College, Taian, Shandong, 271000, China;* [4]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China; and* [5]*University of Chinese Academy of Sciences, Beijing 100149, China*

Summary

1. Hierarchical clustering of molecular data is commonly used for estimation of species diversity in all forms of life. Parameters appropriate for species-level clustering are usually derived from reference data and applied for the delineation of sequences of unknown species membership, although it is not clear how this should be carried out in a multilocus scenario.

2. We introduce a novel means of concurrent clustering parameter optimization and delineation for multilocus data. A simulated annealing heuristic search is performed, whereby clustering thresholds are independently varied for each locus, but optimized according to the recovery of expected taxonomic species globally over loci. For each iteration of the search, one or more loci are randomly selected and a different threshold is separately proposed to cluster each, then the loci are linked to form global species units. Where the set of thresholds group the reference (species labelled) data with high taxonomic congruence, they are adopted for clustering of the subject (nonlabelled) sequences into global molecular operational taxonomic units (global MOTU). Four mined test data sets composed of both reference and subject sequences are combined with a newly sequenced three gene Apoidea data set, and subject to the proposed method.

3. Even optimizing four loci and thousands of sequences, the approach rapidly convergences on a set of parameters with maximal optimality score, although the method masks a high degree of incongruence, and does not always converge on a single set of thresholds. For example, of the 476 Apoidea sequences, 70 global MOTU were inferred over the heuristic search, although the number of single gene MOTU were much lower for the 28S RNA locus, and a range of equally optimal clustering thresholds were observed for the CytB gene.

4. We demonstrate the approach as a scalable species delineation solution for heterogeneous data sets composed of incompletely and inconsistently labelled data from public DNA data bases, for newly sequenced multilocus data, or both. The delineation over a heuristic search of clustering parameters facilitates the estimation of species diversity in multilocus data, giving species estimates that take into account uncertainty regarding choice of clustering thresholds.

Key-words: Apoidea, heuristic clustering, incongruence, MOTU, multilocus clustering, simulated annealing, species delineation

Introduction

The delineation of species diversity has benefitted from the development of many methods that utilize molecular sequence variation (Ferguson 2002; Blaxter, Elsworth & Daub 2004; Schloss & Handelsman 2005; Pons *et al.* 2006; Knowles & Carstens 2007; Powell *et al.* 2011). In the phenetic approach to sequence-based species clustering, sequences are grouped where they exhibit a relative or absolute level of similarity which is deemed high, for example, grouping sequence pairs as a single putative species if they have <2 base differences (Blaxter *et al.* 2005), <0·5% divergence (Floyd *et al.* 2002) or <10x the divergence usually observed in conspecifics (Hebert *et al.* 2004). While there is some opposition to phenetic species clustering due to theoretical aspects and features of biology (Ferguson 2002; Meier *et al.* 2006; Vogler & Monaghan 2006), it remains popular. This is because implementation is somewhat straightforward, requiring no multiple sequence alignment, sequence evolutionary model, prior population assignments, nor phylogenetic interference, and so tends to be used particularly where the data set is large or heterogeneous (e.g. Schloss & Handelsman 2005; Hibbett *et al.* 2011).

*Correspondence author. E-mail: zhucd@ioz.ac.cn

Species clustering has historically been performed primarily on single-locus data sets (Stackebrandt & Goebel 1994; Hugenholtz, Goebel & Pace 1998; Hebert, Ratnasingham & DeWaard 2003; O'Brien *et al*. 2005; Nilsson *et al*. 2009). However, the increase in both species and genomic coverage from the accumulation of public sequence data means that multilocus data will be more frequently available. Progress has been made on sequence-based species delineation of multilocus data in a model-based phylogenetic framework (Liu *et al*. 2008; Heled & Drummond 2010; O'Meara 2010; Yang & Rannala 2010), although as such methods are computationally intensive and often have strict sampling assumptions, they are less applicable both for large data sets, or data sets in which the sampling is incomplete. The latter is a natural feature of data bases, which are a complex assemblage of species and specimens, labelled in a number of different conventions, and represented by gene sequences which overlap to various degrees.

Any method for species-level clustering of multilocus data sets has to address the effect of genomic variation in substitution rate, as this results in differences across loci of the parameters appropriate for species clustering (Cognato 2006). The standard method to infer the parameters is clustering of data in which species identities are already known (reference data). For validation of species units produced under a given clustering parameter, a measure of the similarity of molecular species clusters to morpho-species groups is required, of which the Rand Index and derived measures have been used previously (Göker *et al*. 2009; Sauer & Hausdorf 2012). The optimal values are those in which the molecular clusters of the reference data set most resemble the taxonomic species (Göker *et al*. 2009; Hibbett *et al*. 2011), and these can then be used to cluster sequences in which species diversity is unknown.

This procedure can be performed for each locus in a multigene data set. However, two obstacles are encountered, the formation of global species units from single-locus units, and determination of the most appropriate *set* of thresholds, which are not necessarily equivalent to those separately derived for loci. Firstly, a complex relationship between species clusters at various genes can exist, where incongruence exists in the form of (i) a given taxonomic species being split into more than one molecular operational taxonomic units (MOTU), (ii) a given MOTU containing more than a single taxonomic species or species ID, (iii) different MOTU being formed at different genes as a result of inconsistent sampling, stochastic effects or differing signal. The difficulty in across-locus combination of incongruent MOTU can be overcome using graph-matching algorithms (Chesters & Vogler 2013), which permits multilocus clustering for the purpose of global threshold assessment and formation of MOTU. Secondly, when assessed for global MOTU, the optimal clustering parameters are not expected to be equivalent to unmatched species units. This is illustrated by the toy example in Fig. 1(e,f), with calculated Rand index. In this example, where MOTU and species from different genes are unlinked, the Rand Index favours stringent clustering (Fig. 1a), whereas the Rand index from linked MOTU favours permissive clustering (Fig. 1d). Finally, as using multiple loci requires the assessment of many combinations of parameters, the computational complexity of inferring optimal values is compounded. For example, where 40 different thresholds (96–100% with step size of 0·1, as used herein) are used on each of three loci, the comparison of all would require assessment of 40^3 different parameter combinations. This parameter space increases exponentially with the number of loci, necessitating the use of heuristic solutions for obtaining a reasonable set of global thresholds.

Herein, we extend earlier work, primarily to gain thresholds more appropriate for global MOTU, while reducing the reliance of diversity estimates on specific thresholds. A heuristic search is implemented (Fig. 2) to obtain reasonable global-clustering parameter estimates for a set of individuals (Fig. 2A). Starting from random values, thresholds are proposed (Fig. 2B) for which the resulting species-level clusters (Fig. 2C) are matched between loci (Fig. 2D) and assessed according to congruence of the reference partition with the species labels (Fig. 2E). We use the simulated annealing search (Osman & Kelly 1996), in which solutions of reduced optimality (lower taxonomic congruence) are initially accepted. This provides the potential to escape local optima by traversing regions of lower optimality and increasing the likelihood of obtaining better global solutions. Clustering of unidentified data is performed simultaneously to that of the species labelled data during the heuristic search, giving a distribution of delineations over a number of different thresholds. The search is applied both to mined unidentified data and newly generated data set of Chinese Apoidea, demonstrating a multilocus species-clustering method with low computational demand, and few assumptions on sampling completeness.

Materials and methods

MINED DATA SETS

Insect sequences were obtained for a number of loci densely sampled at the species level. Two mitochondrial (COI and CytB) and two nuclear (EF1a and 28S rRNA) loci were selected, all with broad spectrum primers available for the insects, thus being favoured for species studies in this group. The invertebrate release (as of July 2012) was downloaded from GenBank (http://ftp.ncbi.nih.gov/genbank/), insect sequences were isolated and used to form a local data base. The data base was searched using the following queries: **EF1a**, *Bombus hypocrita* (JF751028), *Bactrocera dorsalis* (GU339154), *Dendroctonus ponderosae* (BT126614), *Hypermnestra helios* (DQ351106), *Cryptocercus punctulatus* (JQ686946), *Agmina diriwi* (GU966919); **COI barcode fragment**, *Cycnia tenera* (AF549611), *Judolia montivagans* (AY165712), *Oecophylla smaragdina* (AY165697), *Empis sp*. (AY165709), *Gryllus pennsylvanicus* (AY165657), *Isogenoides frontalis* (AY165725), *Parcoblatta pensylvanica* (AY165718), *Ameletus andersoni* (AY165698); **CytB,** *Apis mellifera* (NC_001566), *Daphnia pulex* (NC_000844), *Tribolium castaneum* (NC_003081), *Bombyx mori* (AF149768); **28S**, *D. melanogaster* (M21017), *B. mori* (AY038991), *T. castaneum* (HM156703), *D. pulex* (FJ177015), *A. mellifera* (AY703551). Insect sequences homologous to these queries were identified using UCLUST v4.2.66 (Edgar 2010), with an e-value cut-off of 1e-6. Hit subsequences were then extracted using

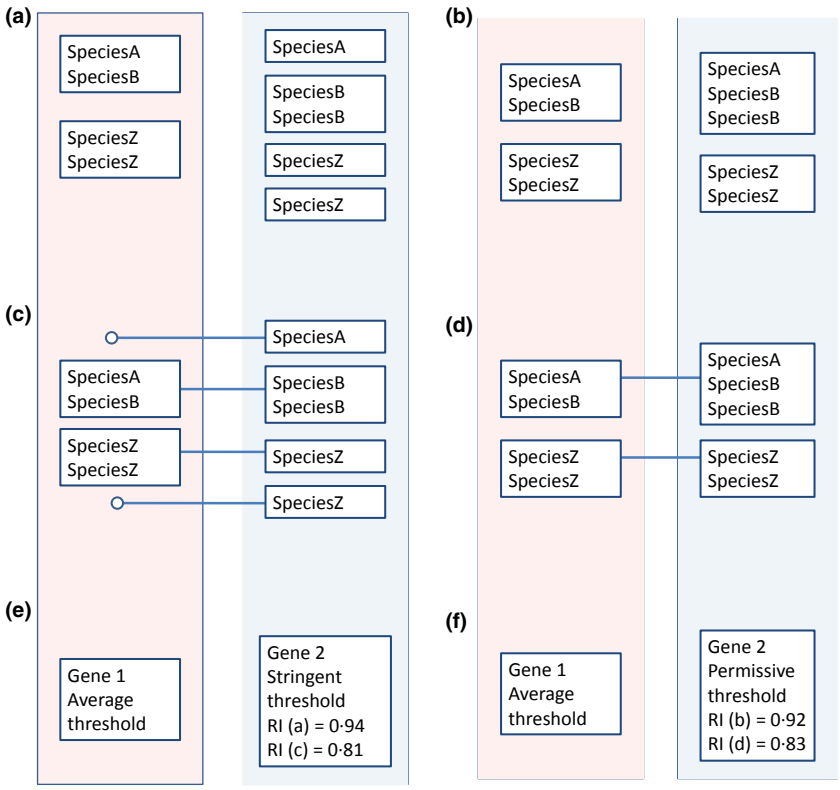

Fig. 1. Contrasting different ways of calculating the Rand Index, for comparing the similarity in MOTU groups and taxonomic groups. These two types of groupings can be treated independently between loci (a, b) or after locus matching (c, d). Species clusters are formed using two loci in all cases, where Gene 1 is given in a red shaded box, and gene 2 in the blue shaded box. Linkage clustering is performed using constant thresholds for gene 1, whereas clustering parameters are varied for gene 2, with stringent (left blue) and permissive (right blue) thresholds used, resulting in either four clusters (left blue) or two clusters (right blue). (e, f) Give calculation of the Rand index. Adjusted versions of the Rand Index (e.g. HA Rand Index) give an equivalent pattern.

BLASTDBCMD v.2.2.25 (Camacho *et al.* 2009) and a custom Perl wrapper script. Finally, we identified four broad scale taxonomic groups to be used as test data sets. Groups were selected using an automated procedure to maximize (i) common species between, (ii) unidentified sequences within and (iii) intraspecific data within, various combinations of the genes listed above.

SAMPLE COLLECTION, DNA EXTRACTION AND SEQUENCING

In addition to the mined data, we generated a multilocus Apoidea data set from an ongoing diversity monitoring study of Orchard insects. Insects were collected by malaise trapping in Beijing, China (40°06·239N, 115°54·773E), between April 2011 and September 2011. Approximately 250 Apoidea specimens were subject to DNA sequencing for the COI, CytB and 28S genes. Genomic DNA was extracted using the QIAGEN DNeasy tissue extraction kit (QIAGEN, Valencia, CA, USA). The COI and CytB genes were amplified via PCR using LA Taq (TAKARA) and 28S using MightyAmp (TAKARA). The primer pair LCO1490 (5′-GGTCA ACAAA TCATA AAGAT ATTGG-3′) and HCO2198 (5′-TAAAC TTCAG GGTGA CCAAA AAATCA-3′) (Folmer *et al.* 1994) were used to amplify COI. The 28S gene utilized the primer pairs D2-3549F (5′-AGTCG TGTTG CTTGA TAGTG CAG-3′) and D2-4068R (5′-TTGGT CCGTG TTTCA AGACG GG-3′) (Campbell, Steffen-Campbell & Werren 1993) or D2-3566F (5′-TGCAG CTCTA AGTTG GTGGT-3′) (Gillespie *et al.* 2005) and D2-4057R (5′-TCAAG ACGGG TCCTG AAAGT-3′) (Heraty *et al.* 2004), and CytB utilized the primer pair CytB F (5′-CGWTT AATTC ATATA AATGG-3′) and CytB R (5′-TATCA TTCWG GTTTA ATATG-3′) (Koulianos 1999). All amplification reactions were performed in a total volume of 50 μL, in which COI and CytB reactions included 5 μL 10× LA buffer, 5 μL MgCl2 (2·5 mM), 5 μL dNTP (2·5 mM), 1 μL each primer (10 mM), 0·5 μL LA Taq polymerase (5 U μL^{-1}), 2–4 μL template DNA and distilled water up to 50 μL. The 28S reaction included 25 μL MightyAmp Buffer version 2, 1 μL MightyAmp DNA Polymerase (1·25 U μL^{-1}), 1 μl of each primer (10 mM), 2–4 μL template DNA and distilled water to 50 μL. The PCR conditions were as following: 94°C for 2 min, 35 cycles of 94°C for 30 s, 48–50°C for 50 s, 72°C for 1 min and a final extension at 72°C for 10 min for the COI and CytB reactions; 98°C for 2 min, 35 cycles of 98°C for 10 s; 58°C for 15 s; 68°C for 1 min; and a final extension at 68°C for 5 min for the 28S reaction. Sequencing was performed with an ABI3130 sequencer. Sequences are made publically available (see Data Accessibility section).

FORMING GLOBAL MOTU FROM MULTIPLE CLUSTERED LOCI

The molecular data were clustered into species-level units according to sequence similarity. Pairwise percent identities were obtained for sequences within each locus by all-against-all alignment using UCLUST (Edgar 2010), and then used for single-linkage clustering under a range

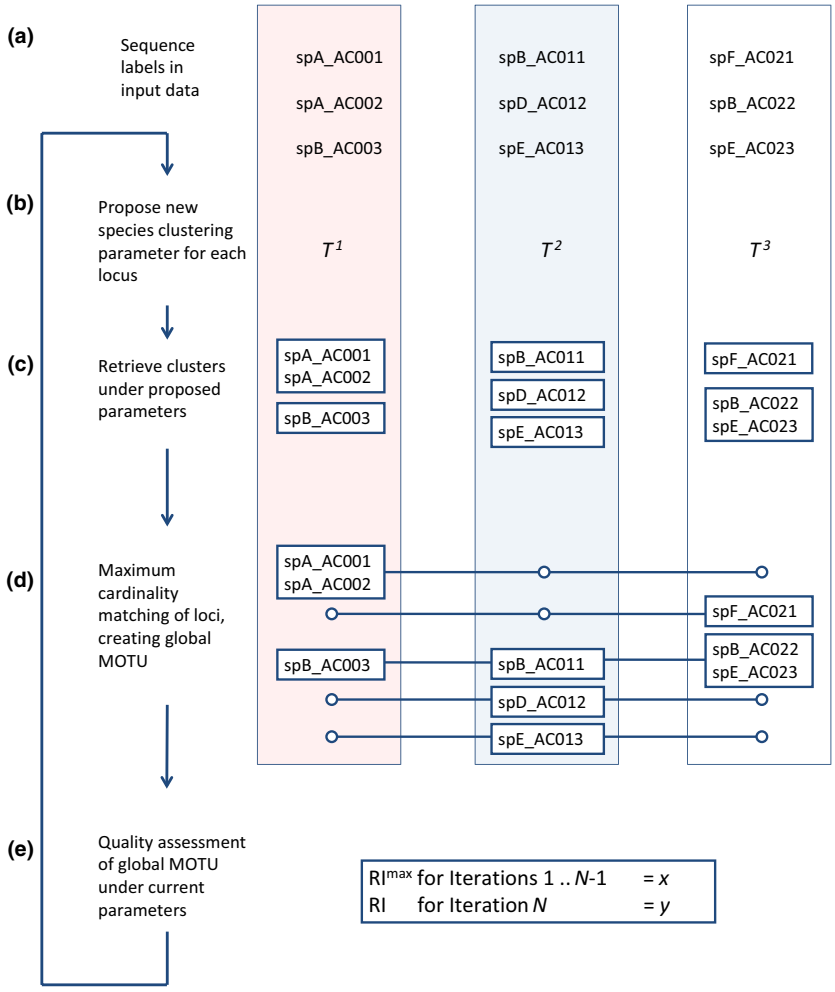

Fig. 2. Flow chart of the heuristic search for three loci (red, blue and white boxes). (A) The parameter optimization is carried out on species labelled sequences. Toy sequences here are illustrated with a species label (spA, spB, spD-spF) followed by accession number, with some species having more than a single sequence present. In a single iteration N of the search (B–E), (B) a set of clustering parameters (T^j = linkage threshold for locus j) is proposed for each of the loci, (C) the set of species clusters under these parameters is retrieved, (D) species units are matched between loci, forming global MOTU, (E) The set of global MOTU is assessed for taxonomic congruence according to the HA Rand Index, and the proposed clustering parameters adopted if the congruence is improved, adopted if congruence is worsened at a probability determined by the annealing temperature, and rejected otherwise. The optimization is run on the named species, although clustering of the unidentified sequences is performed concurrently, under the parameters adopted through the search.

of thresholds. Linkage clustering was performed using a Perl script optimized for BLAST or UCLUST all-against-all outputs (script is made available, see Data Accessibility section).

Where multiple gene fragments are independently clustered into species units, the formation of global MOTU requires the combination of species clusters derived from the different genes. The presence of incongruent clusters and different representative sampling for the genes prevents straightforward combination of results. Chesters & Vogler (2013) introduced a method for combination of such species clusters, using graph algorithms. The method first identifies all possible links between the first two genes (a link/edge occurs where a species cluster in one gene contains a sequence ID identical to at least one sequence ID in a cluster at the second gene). The links are then decomposed by identifying the set in which most taxonomic links occur, by maximal cardinality matching, which is iterated over all loci. The method therefore permits the formation of global MOTU from different loci even where they are incongruent.

MULTI LOCUS CLUSTERING OPTIMIZATION AND DELINEATION

Hierarchical clustering requires specification of separate parameters for each locus. By clustering a reference data set under a range of parameters, settings producing the highest taxonomic congruence can be inferred and applied to delineate unidentified data. As a measure of taxonomic congruence we use the Hubert and Arabie adjusted Rand index (hereafter, the HA Rand index) (Hubert & Arabie 1985) as calculated by the Clues R module (R Development Core Team 2008; Chang *et al.* 2010). Assessing congruence of various parameters is a nonlinear problem where multiple loci are used, therefore, we perform a heuristic search in which a global optimal set of clustering parameters is sought. A flow chart illustrating the search is given in Fig. 2. Starting from random states (a random clustering threshold for each locus), a Monte Carlo chain is run in which new states are proposed (Fig. 2B), corresponding to a set of

clusters of individuals (Fig. 2C), which are subsequently linked by partite matching (Fig. 2D), then assessed for taxonomic congruence (Fig. 2E). Multiple simulated annealing searches were performed for each test data set, optimizing clustering thresholds according to taxonomic congruence of identified sequences, to delineate unidentified and newly generated data. The simulated annealing solution was implemented in Perl (MLCO.pl: multi locus clustering optimization) and is made freely available (see Data Accessibility).

We implement a cluster-based method of species diversity estimation that incorporates uncertainty in regard to threshold. The use of a single-clustering threshold is problematic, as it is likely that multiple peaks exist, which lack significant difference in optimality score. This is compounded where multiple loci are clustered as the parameter space is greatly increased. We here perform global species clustering of the unidentified sequences concurrent with that of the reference data set during the heuristic search. At each iteration of the heuristic search, the current optimal set of thresholds is used to form global MOTU of the unidentified sequences. This gives a distribution of delineations over the various parameter combinations formed during the search. Species diversity estimates are made using MOTU delineations from the stationary phase, which permits the incorporation of the uncertainty in species delineation and allows estimation of species diversity without reliance on specific thresholds.

Results

Insect sequences belonging to four gene fragments were mined from GenBank, totalling 159 221 (COI), 12 108 (CytB), 16 960 (EF1a) and 14 627 (28S). From these files, we obtained reference data sets for specific taxonomic groups. Reference data sets were obtained according to three criteria; those with a relatively high number of (i) unidentified sequences for each of the input loci, (ii) species in which most of the input loci are sequenced and (iii) named species which have multiple sequences available (i.e. intraspecific data). Based on these criteria, we filtered the taxa; Apoidae (Hymenoptera) for COI+CytB+28S, Formicinae (Hymenoptera) for COI+CytB+28S+EF1a, Satyrinae (Lepidoptera) for COI+CytB+EF1a and Chrysomelidae (Coleoptera) for COI+CytB+EF1a. The unidentified sequences numbered 364, 546, 237 and 217 for the Apoidae, Formicinae, Satyrinae and Chrysomelidae, respectively (full details of the mined reference data sets are given in Table 1). The Apoidae data set was supplemented with newly generated sequences from a regional species diversity study. The sequencing was incomplete for the majority of specimens, with all three loci obtained for 61. In total, 476 sequences were obtained (164 COI, 136 CytB, 176 28S) over *c.* 250 specimens. All of the unidentified data (both the mined unidentified sequences and the newly sequenced Apoidae) were subject to molecular species-level clustering.

Single-linkage clustering was performed at thresholds varying between 96 and 100% identity in steps of 0·1, then a heuristic search of the clustering parameter space was made in which the taxonomic congruence of the (multilocus) species clusters to the reference data was maximized. We first optimized the heuristic search parameters using the Satyrinae data set, this being the only data set problematic for convergence in preliminary analyses. The search was tuned by varying the decay parameter. This parameter effectively increases or decreases the phase of the search in which broad regions of parameter space are covered. An appropriate decay is dependent on many features of the data set, such as the number of

Table 1. Mined test data sets. Data sets were selected for various taxa, loci and number of loci, based on criteria of composition given in the text. Superfamily Apoidae contains an additional row giving data newly sequenced as part of this study. 'Unlabelled' refers to lack of species-level taxonomic annotation of sequence

					Loci:			
	COI	CytB	28S	EF1a	1	2	3	4
Superfamily Apoidae								
Total sequences	2123	417	943	NA				
Unlabelled sequences	146	50	168	NA				
Laboratory-generated sequences	164	136	176	NA				
Species with >1 sequence	320	46	87	NA				
Number of species with…					1165	301	17	NA
Subfamily Formicinae								
Total Sequences	1022	485	135	275				
Unlabelled sequences	342	74	47	83				
Species with >1 sequence	82	42	10	42				
Number of species with…					142	120	45	28
Subfamily Satyrinae								
Total Sequences	8171	428	NA	2495				
Unlabelled sequences	138	22	NA	77				
Species with >1 sequence	987	60	NA	417				
Number of species with…					736	1329	82	NA
Family Chrysomelidae								
Total Sequences	1451	472	NA	561				
Unlabelled sequences	133	42	NA	42				
Species with >1 sequence	139	32	NA	98				
Number of species with…					454	235	19	NA

variables, but visual inspection of searches under several arbitrary values gives reasonable values. Figure 3 shows the result of varying the simulated annealing temperature decay parameter over 2000 generations, with four independent runs for each setting. All runs attained identical optimal taxonomic congruence when using decay values of 0·990, 0·995 and 0·998, whereas convergence was not attained within the 2000 generations at a temperature decay of 0·9995. At lower values (0·990, 0·995), the stationary phase of individual runs was ambiguous based on visual assessment, therefore, we adopted default temperature decay of 0·998, which results in a distribution in which the stationary phase is most reliably identified.

Next, we assessed the impact of locus matching on parameter optimization. By default, the calculation of the HA Rand index was made on molecular clusters in which loci were linked by multipartite matching (e.g. as in Fig. 1c,d), primarily to give a global delineation and thus overall estimate of species diversity. The analysis was repeated on unmatched data (e.g. Fig. 1a,b), in which molecular species units were formed within loci only. Figure 4 contrasts clustering parameters for species clusters unmatched (Fig. 4a), and matched (Fig. 4b), between loci. At the stationary phase, CytB is shown switching between different thresholds whether loci are matched or not, whereas a single optimal parameter is obtained for the other two loci. The optimal parameters differ for the two genes between the linked and unlinked searches, although not in the way predicted under the simplified example in Fig. 1, indicating a more complex relationship. Where unlinked, the HA Rand index favours thresholds of 97·1 for COI and *c.* 97·5 for CytB, whereas this is increased to 99 for COI and *c.* 99·2 for CytB, where linked.

A heuristic search for optimal clustering parameters was performed on each of the four test data sets, with two independent runs from random start thresholds. All runs reached the stationary phase after *c.* 500 generations (as assessed by the HA Rand Index), and most attain a single optimal score after *c.* 1000 generations. The concurrent clustering of the reference (identified) and subject (unidentified) data over the heuristic search gave a distribution of delineations over a space of reasonable parameter combinations. This is demonstrated in Fig. 5, which gives the number of global MOTU formed from the unidentified data at each step during the search. The search phase prior to stationary was discarded, then species diversity estimated according to the delineations remaining. The MOTU delineations from the stationary phase are given in Table 2. The number of species clusters was found to be stable over this phase of the search, with an error margin of <1 MOTU in all cases. The majority of unidentified sequences are likely derived from unique species, as indicated by the number of species clusters. For example, the 342 (COI), 74 (CytB), 47

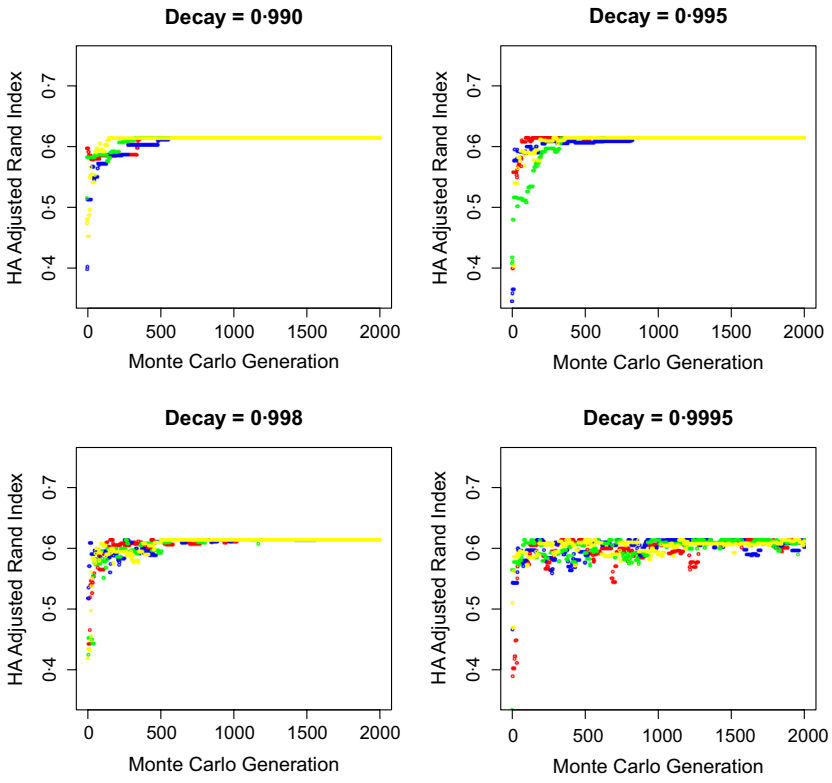

Fig. 3. Tuning the decay of the temperature parameter during simulated annealing, using the Satyrinae data set. The *x* axis gives the generation during the heuristic search for the global optimal clustering thresholds, whereas the *y* axis gives the HA Rand index of similarity between the global molecular clusters and taxonomic species. Four decay values are shown, with four independent runs (indicated by point colour in the plot) of 2000 generations performed for each. Lower decay values lead to rapid reduction in temperature over the search, and thus, a shorter period in which broad parameter space is covered.

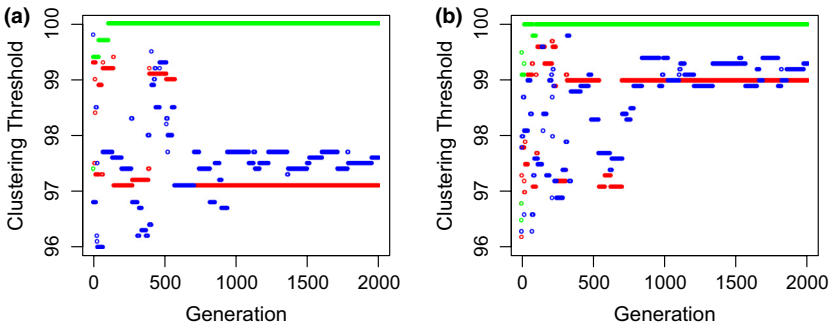

Fig. 4. Optimal clustering thresholds for COI (red), CytB (blue), EF1a (green), during a heuristic search in which the tRI is maximized. (a) results where unlinked tRI is maximized. (b) gives thresholds where tRI is calculated from species clusters matched between loci and is adopted by default in this study.

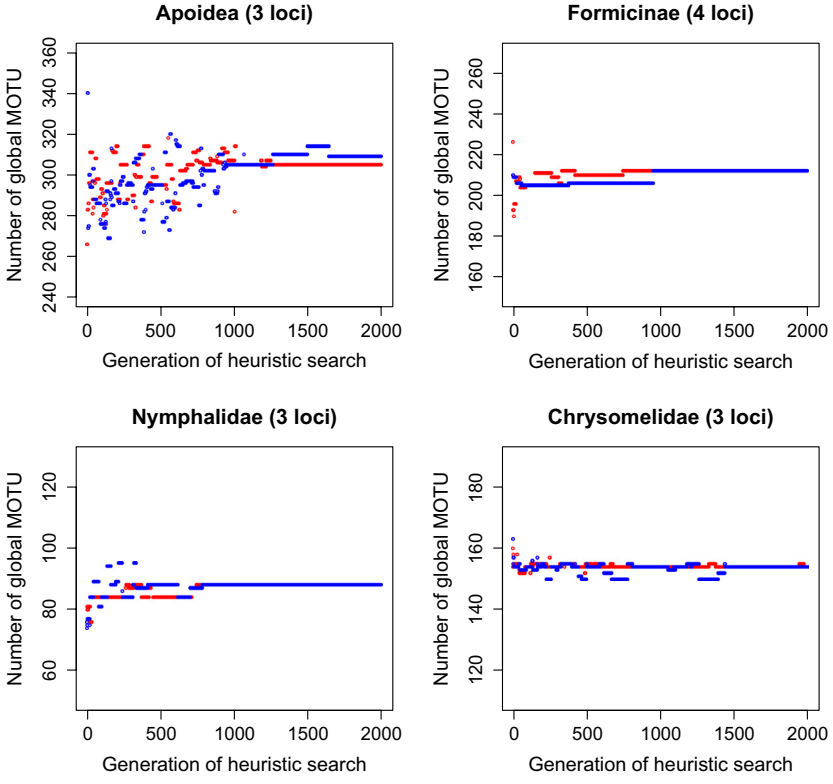

Fig. 5. Number of global MOTU delineated from the unidentified data, for the four test data sets. Points are plotted for MOTU counts produced under the parameters adopted at each generation of the Markov chain. Two independent runs shown as coloured differently.

Table 2. Delineation results for data mined from Genbank, and newly produced herein. Number of species-level clusters both for individual loci and matched (global column) loci, based on the optimal clustering parameters. Error is given according to variance in number of species clusters during the stationary phase of the heuristic search. Note dissimilarity in species estimates for individual loci is expected due to sampling variation, see Discussion on core specimens

	COI	CYTB	EF1a	28S	Global
Apoidae					
Mined clusters	172 ± 0·19	75 ± 0·4	NA	134 ± 0	306 ± 0·13
Lab clusters	54 ± 0·05	41 ± 0·16	NA	21 ± 0	70 ± 0·07
Formicinae					
Mined clusters	158 ± 0·0	59 ± 0·0	72 ± 0·0	16 ± 0·0	212 ± 0·0
Satyrinae					
Mined clusters	68 ± 0·00	13 ± 0·00	51 ± 0·00	NA	88 ± 0·00
Chrysomelidae					
Mined clusters	116 ± 0·02	31 ± 0·0	40 ± 0·01	NA	154 ± 0·05

(28S), 83 (EF1a) Formicinae sequences (Table 1) were clustered into 158, 59, 70, 16 single-locus units (Table 2). Thus, the species diversity represented in unidentified data may be quite substantial. Further, many of the global MOTU have members spanning multiple loci. In the case of Formicinae, the MOTU from the four individual loci numbers 300 in total, but where multipartite matching is performed, together these form 212 global MOTU. Again for global MOTU, the multilocus delineation appears quite stable, with an error margin <1 for all test data sets.

The global MOTU formed from the laboratory-generated Apoidae data are illustrated in Fig. 6, with COI, CytB and 28S data represented as white, blue and red segments. As is apparent, much of the global species-level diversity in the Chinese Apoidae sample is represented at the COI and CytB loci, with many MOTU containing only sequences of these loci, in contrast to the 28S locus, in which are only present in 21 MOTU, despite a greater number of sequences being used. Additionally, the majority of the global MOTU consist of representatives of more than one locus. This does not necessarily mean a high level of congruence between loci, just that the method of matching is effective in producing a high number of links between genes. Species names are assigned where unambiguous, that is, when a given MOTU contain unidentified data and no more than a single-named species, which is not also present in other MOTU. Of the 70 global MOTU from the new Apoidea data set, 15 contained named GenBank derived data, with the remaining 55 species units therefore probably not yet represented on the public data base. Although the incongruence in these 15 global MOTU was high, with taxonomic species split over multiple MOTU, and MOTU containing multiple species names. Thus, only four species names (*Megachile sidalceae, Osmia cornuta, O. taurus and Bombus ignitus*) could be assigned by association of the new and existing annotated data. Genus level assignments were given in the remaining 11.

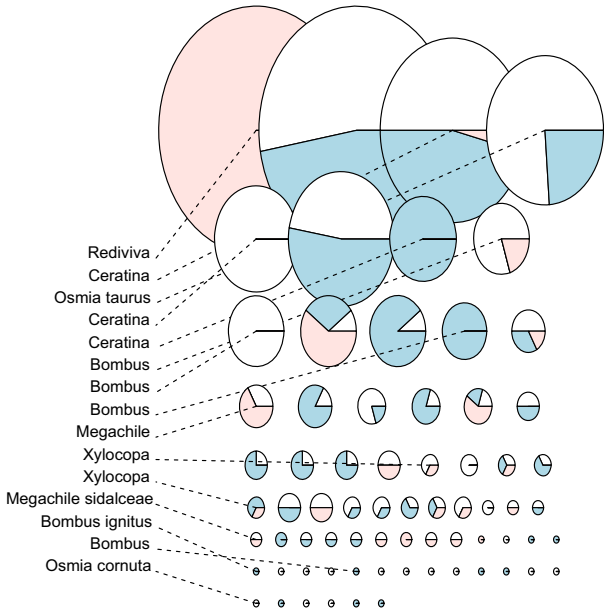

Fig. 6. The 70 global MOTU delineated from a newly generated three gene Apoidea data set. Each MOTU is represented as a pie chart, sized relative to the number of sequences composing the MOTU, and coloured according to loci represented, where blue segments = CytB, white = COI, red = 28S. MOTU are assigned taxonomic names based on association of a MOTU with species labelled GenBank data. Species names are given where unambiguous, that is, where the MOTU consists of (in addition to the newly generated data) no more than a single labelled species name, and where this species name is not found in other MOTU. Genus names are given otherwise.

Discussion

Four mined data sets and 476 new Apoidea sequences were used to demonstrate a method of species-level clustering for multilocus data, while the consolidation of species units formed from multiple loci has received little prior attention. Setaro *et al.* (2012) recently conducted clustering parameter optimization on a two-locus data set, by varying clustering parameters at one locus, to maximize congruence in MOTU with a second locus in which parameters were fixed. The optimal parameters were then applied for estimating species diversity in fungal ecosystems. While the utility of their approach is delimitation based on internal criteria only (i.e. species discovery), a limitation is that the loci remain unlinked, and thus global diversity remains unknown.

In the current work, we address the difficulties in clustering optimization of multilocus data, with a recently developed procedure for the formation of global species units. Species units separately formed by linkage clustering multiple loci are subsequently combined using an algorithm that maximizes taxonomic links, creating the set of global MOTU such as those illustrated in Fig. 6. Formation of global MOTU is performed during each iteration of a heuristic search (Fig. 5), to optimize clustering parameters on the reference data, while simultaneously delineating the unidentified data. Each locus is clustered under a starting threshold by single linkage, then clusters from each loci are matched by maximal cardinality. The quality of threshold combination is assessed by determining the similarity to taxonomic species clusters, as scored using the HA Rand Index. New parameters are proposed, in which improvements in HA Rand Index are always accepted, and reduced HA Rand Index are accepted at a rate determined by the temperature. Where performing multiple independent heuristic searches from random start thresholds allows the attainment of a maximal optimality score to be confirmed.

The proposed method lacks biological basis, perhaps characteristic of phenetics in general, which has been discussed at great length elsewhere (Ferguson 2002; Meier *et al.* 2006). The use of such algorithmic solutions sidesteps what can be extensive incongruence, and unintuitive results can be generated when forming global MOTU in this way. For example, it is difficult to immediately accord many of the MOTU presented in Fig. 6 as derived from the data in the supplementary Table S1. Specifically, by separately clustering loci into the initial species-level clusters, a given specimen that possesses multiple fragments may become subdivided among multiple MOTU, if

these loci give an incongruent signal. The splitting of individual specimens among multiple global MOTU may complicate some downstream applications, as the individuals become fuzzy entities, at the behest of fully delineating the species-level entities. The degree of such incongruence can be stated based on a set of core specimens, with 61 of which all gene fragments were sequenced. The number of species-level clusters formed from these core specimens differ, with 21 in the case of COI, 20 for CytB, and 10 for 28S. This variation could be due to complicating biological factors, or error in reconstruction of species units from a lack of information content at the species level. In the current case, this may suggest 28S is unsuitable for species-level clustering by single linkage. Still, much unidentified data are routinely generated for many varied gene regions, the incongruence is present between these genes, and so it is necessary to use procedures such as these if the data are to be utilized.

The approach described herein complements a number of multilocus species delineation methods have been developed recently that operate within a phylogenetic framework. The latter are often restrictive in a way that can be problematic for users of mined data sets. In particular, existing methods may require a priori assignment of members to putative species units (Yang & Rannala 2010; Ence & Carstens 2011), they may assume a virtually complete sampling pattern (Lim, Balke & Meier 2012) in which a representative sequence is present for all members to be delineated, or they may limited by computations to the delineation of only dozens of members (Liu *et al*. 2008; Heled & Drummond 2010; O'Meara 2010). The approach used here requires only the prior partitioning of sequences into homologs and is applicable for data sets numbering well into the thousands. The rate-limiting step of this protocol is not the heuristic search itself, but the calculation of pairwise distances, of which computational complexity increases quadratically with the number of sequences. There are means to accelerate such computations where necessary (Li, Jaroszewski & Godzik 2001; Rattei *et al*. 2010), and as pairwise alignments are independent, they are inherently parallelizable (e.g. Mathog 2003). Further, clustering based on pairwise similarity avoids the often uncertain and problematic steps of multiple sequence alignment and phylogenetic inference, which can be a source of uncertainty and error in diversity estimation (Sun *et al*. 2009). Finally, by delineating over a distribution of parameters, it sidesteps an often disputed aspect of species clustering; the selection of a specific threshold, which is frequently cited as problematic for clustering. The clustering threshold is the degree of similarity above which two sequences are regarded to be conspecifics, but a specific threshold may not be appropriate for all taxa (Meier, Zhang & Ali 2008) and multiple peaks of similar optimality may exist. Here, uncertainties with regard to threshold are accommodated in the estimate of species diversity, through using a distribution of thresholds over the heuristic search. Thus, heterogeneous multilocus data sets can be clustered at the species level in a way practical for further evolutionary analysis, while giving diversity estimates with indication of error arising from selection of clustering parameters.

Acknowledgements

The authors would like to thank Ze-Qing Niu for identification of Apoidea specimens, and editors and reviewers who advised on improving earlier drafts of this manuscript.

Data Accessibility

- DNA sequence data: Available on GenBank under accession numbers KC560250 to KC560725.
- Apoidae sample information: Full details are given in supplementary Table S1.
- Perl scripts: Implementation of multilocus clustering method described in this paper is available at http://sourceforge.net/projects/mlco/files/, and script for (single, average, complete linkage) clustering based on the output of Blast or Uclust all-against-all search is available at http://sourceforge.net/projects/singlelinkage/files/.

Funding

This work was supported by the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KSXC2-EW-B-02); the National Science Foundation of China (Grants No. 30870268, 31172048, J1210002); the Public Welfare Project from the Ministry of Agriculture of the People's Republic of China (Grant no. 201103024); and the Program of Ministry of Science and Technology of the People's Republic of China (2012FY111100) to CDZ.

References

Blaxter, M., Elsworth, B. & Daub, J. (2004) DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **271**, S189–S192.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R. & Abebe, E. (2005) Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1935–1943.

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

Campbell, B.C., Steffen-Campbell, J.D. & Werren, J.H. (1993) Phylogeny of the Nasonia species complex (Hymenoptera: Pteromalidae) inferred from an internal transcribed spacer (ITS2) and 28S rDNA sequences. *Insect Molecular Biology*, **2**, 225–237.

Chang, F., Qiu, W., Zamar, R.H., Lazarus, R. & Wang, X. (2010) Clues: an R Package for nonparametric clustering based on local shrinking. *Journal of Statistical Software*, **33**, 1–16.

Chesters, D. & Vogler, A.P. (2013) Resolving ambiguity of species limits and concatenation in multilocus sequence data for the construction of phylogenetic supermatrices. *Systematic Biology*, **62**, 456–466.

Cognato, A.I. (2006) Standard percent DNA sequence difference for insects does not predict species boundaries. *Journal of Economic Entomology*, **99**, 1037–1045.

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

Ence, D.D. & Carstens, B.C. (2011) SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, **11**, 473–480.

Ferguson, J.W.H. (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*, **75**, 509–516.

Floyd, R., Abebe, E., Papert, A. & Blaxter, M. (2002) Molecular barcodes for soil nematode identification. *Molecular Ecology*, **11**, 839–850.

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.

Gillespie, J.J., Munro, J.B., Heraty, J.M., Yoder, M.J., Owen, A.K. & Carmichael, A.E. (2005) A secondary structural model of the 28S rRNA expansion segments D2 and D3 for chalcidoid wasps (Hymenoptera: Chalcidoidea). *Molecular Biology and Evolution*, **22**, 1593–1608.

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.

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 of London. Series B: Biological Sciences, **270**, S596–S599.
Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S. & Francis, C.M. (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.
Heled, J. & Drummond, A.J. (2010) Bayesian inference of species trees from multi locus data. *Molecular Biology and Evolution*, **27**, 570–580.
Heraty, J.M., Hawks, D., Kostecki, J.S. & Carmichael, A. (2004) Phylogeny and behavior of the Gollumiellinae, a new subfamily of the ant parasitic Eucharitidae (Hymenoptera: Chalcidoidea). *Systematic Entomology*, **29**, 544–559.
Hibbett, D.S., Ohman, A., Glotzer, D., Nuhn, M., Kirk, P. & Nilsson, R.H. (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biology Reviews*, **25**, 38–47.
Hubert, L. & Arabie, P. (1985) Comparing partitions. *Journal of Classification*, **2**, 193–218.
Hugenholtz, P., Goebel, B.M. & Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, **180**, 4765–4774.
Knowles, L.L. & Carstens, B.C. (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, **56**, 887–895.
Koulianos, S. (1999) Phylogenetic relationship of the bumblebee subgenus Pyrobombus (Hymenoptera: Apidae) inferred from mitochondrial cytochromeb and cytochrome oxidase I sequence. *Annals of the Entomological Society of America*, **92**, 355–358.
Li, W., Jaroszewski, L. & Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
Lim, G.S., Balke, M. & Meier, R. (2012) Determining species boundaries in a world full of rarity: singletons, species delimitation methods. *Systematic biology*, **61**, 165–169.
Liu, L., Pearl, D.K., Brumfield, R.T. & Edwards, S.V. (2008) Estimating species trees using multiple-allele DNA sequence data. *Evolution*, **62**, 2080–2091.
Mathog, D.R. (2003) Parallel BLAST on split databases. *Bioinformatics*, **19**, 1865–1866.
Meier, R., Zhang, G. & Ali, F. (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the "Barcoding Gap" and leads to misidentification. *Systematic Biology*, **57**, 809–813.
Meier, R., Shiyang, K., Vaidya, G. & Ng, P.K.L. (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.
Nilsson, R., Ryberg, M., Abarenkov, K., Sjokvist, E. & Kristiansson, E. (2009) The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters*, **296**, 97–101.
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.
O'Meara, B.C. (2010) New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*, **59**, 59–73.
Osman, I.H. & Kelly, J.P. (Eds) (1996) *Meta-Heuristics: Theory and Applications*. Springer, Kluwer Academic Publishers, Norwell MA, USA.
Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. & Vogler, A.P. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.
Powell, J.R., Monaghan, M.T., Opik, M. & Rillig, M.C. (2011) Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology*, **20**, 655–666.
R Development Core Team (2008) *R: A Language and Environment for Statistical Computing [Computer Software and Manual]*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
Rattei, T., Tischler, P., Götz, S., Jehl, M.A., Hoser, J., Arnold, R., Conesa, A. & Mewes, H.W. (2010) SIMAP—a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic acids research*, **38**, D223–D226.
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.
Schloss, P.D. & Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, **71**, 1501–1506.
Setaro, S.D., Garnica, S., Herrera, P.I., Suárez, J.P. & Göker, M. (2012) A clustering optimization strategy to estimate species richness of Sebacinales in the tropical Andes based on molecular sequences from distinct DNA regions. *Biodiversity and Conservation*, **21**, 2269–2285.
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.
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 pyrosequences. *Nucleic Acids Research*, **37**, e76.
Vogler, A.P. & Monaghan, M.T. (2006) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.
Yang, Z. & Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9264–9269.

Received 4 February 2013; accepted 15 July 2013
Handling Editor: Robert Freckleton

Supporting Information

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

Table S1. Spreadsheet giving details of the 251 Apoidea specimens sequenced as part of this study. Genbank accession numbers are given where a locus was successively sequenced for the given specimen, with blank entries indicating the locus was not sequenced. Different primer combinations were used only with 28S, so primer names are listed for this gene.