Blackwell Publishing

Cladistics 23 (2007) 1–21

Cladistics

10.1111/j.1096-0031.2006.00126.x

A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms

Damon P. Little* and Dennis Wm. Stevenson

Lewis B. and Dorothy Cullman Program for Molecular Systematic Studies, The New York Botanical Garden, Bronx, New York 10458-5126, USA

Accepted 28 June 2006

Abstract

In order to use DNA sequences for specimen identification (e.g., barcoding, fingerprinting) an algorithm to compare query sequences with a reference database is needed. Precision and accuracy of query sequence identification was estimated for hierarchical clustering (parsimony and neighbor joining), similarity methods (BLAST, BLAT and megaBLAST), combined clustering/similarity methods (BLAST/parsimony and BLAST/neighbor joining), diagnostic methods (DNA–BAR and DOME ID), and a new method (ATIM). We offer two novel alignment-free algorithmic solutions (DOME ID and ATIM) to identify query sequences for the purposes of DNA barcoding. Publicly available gymnosperm nrITS 2 and plastid *matK* sequences were used as test data sets. On the test data sets, almost all of the methods were able to accurately identify sequences to genus; however, no method was able to accurately identify query sequences to species at a frequency that would be considered useful for routine specimen identification (42–71% unambiguously correct). Clustering methods performed the worst (perhaps due to alignment issues). Similarity methods, ATIM, DNA–BAR, and DOME ID all performed at approximately the same level. Given the relative precision of the algorithms (median = 67% unambiguous), the low accuracy of species-level identification observed could be ascribed to the lack of correspondence between patterns of allelic similarity and species delimitations. Application of DNA barcoding to sequences of CITES listed cycads (Cycadopsida) provides an example of the potential application of DNA barcoding to enforcement of conservation laws.

Recently, there have been several widely circulated proposals to use DNA sequences to identify specimens across all of life (Floyd et al., 2002; Hebert et al., 2003) in a manner analogous to the commercially ubiquitous Universal Product Codes (UPC; Savir and Laurer, 1975). The proposal has been branded ''DNA barcoding'' in honor of the most common graphic manifestation of UPC. Identification of samples based on their DNA sequences, instead of diagnostic morphological features, is not a new idea—a great number of researchers have used sequence-based ''DNA fingerprinting'' to identify specimens for some time now (e.g., Fox et al., 1977; Niesters et al., 1993; Li et al., 1995; Poinar et al., 1998; Amato et al., 1999; Doukakis et al., 1999; Jackson et al., 1999; Zaidi et al., 1999; Fell et al., 2000; Hofreiter et al., 2000; Kõljalg et al., 2000; Kristiansen et al., 2001; Wells and Sperling, 2001; Wells et al., 2001a,b; Brown et al., 2002; Vrålstad et al., 2002). The use of DNA fingerprinting or other biochemical markers for identification is particularly common for microbial studies perhaps due to the general lack of microbial morphological diversity. DNA fingerprinting has also been applied to more morphologically diverse taxa in instances where the only available material is of inadequate quality to make a morphology-based determination (e.g., Poinar et al., 1998; Jackson et al., 1999; Hofreiter et al., 2000; Wells and Sperling, 2001; Wells et al., 2001a,b).

DNA barcoding has immense conservation application, particularly in the enforcement of bans on trafficking of protected species. Barcoding could allow for relatively rapid ($\approx$ 12 h) and accurate identification by non-specialists of listed taxa at customs checkpoints. In addition, the identification of morphologically deficient or incomplete specimens can also be undertaken

*Corresponding author:
E-mail address: dlittle@nybg.org

(e.g., powders). For example, all 305 species of Cycadopsida are protected by the Convention on International Trade in Endangered Species (CITES; http://www.cites.org/eng/app/appendices.shtml) resulting in regulation of international transport of all specimens. With the exception of *Cycas beddomei* Dyer, all species in six of the 11 Cycadopsida genera are listed in CITES appendix II. The remaining five genera and *C. beddomei* are listed in appendix I. The ability of customs inspectors to discriminate between specimens protected by appendix I and those protected by appendix II is critical as trade in appendix II species is allowed under some circumstances. Currently, the monosaccharide profiles of mucilage (Stevenson and Gigliano, 1989) are used by customs inspectors to determine if a given specimen of Cycadopsida is protected under appendix I or II.

Problems with pair-wise distances

The use of DNA sequences for arbitrary and capricious species delimitation is widely criticized (Tautz et al., 2002; Lipscomb et al., 2003; Sperling, 2003; Moritz and Cicero, 2004; Will and Rubinoff, 2004; DeSalle et al., 2005; Monaghan et al., 2005; Vences et al., 2005b; Meier et al., in press). Ferguson (2002) provides an extensive explanation of why distances are inappropriate for species circumscription from a biosystematics perspective. However, the distance-based approach is favored by some proponents (e.g., Floyd et al., 2002; Hebert et al., 2003; Blaxter, 2004; Lambert et al., 2005) whom advocate a strict divergence threshold above which sequences are considered to belong to separate species and below which the sequences are considered conspecific. To illustrate the perils of using percent divergence, to circumscribe species, a hypothetical situation in which it is not possible to circumscribe one or more species that simultaneously incorporate all sequences with less than 5% divergence while at the same time excluding sequences that diverge 5% or more, is shown in Fig. 1. Pair-wise distances suggest either: (i) one species consisting of A + B + D and a second species consisting of only C; (ii) one species consisting of A + B + C and a second species consisting of only D; (iii) one species consisting only of A and a second species consisting of B + C + D; or (iv) one species consisting of A + C + D and a second species consisting of only B. The use of patristic distances does not resolve these contradictory circumscriptions. Patristic distances suggest either: (i′) one species consisting of A + B + D and a second species consisting of only C; (ii′) one species consisting of A + B and a second species consisting of C + D; or (iii′) one species consisting of A + C + D and a second species consisting of only B. An objective criterion to distinguish among these possible circumscriptions does not exist. Either all possible circumscriptions are correct or they are all incorrect. In this example neither pair-wise nor patristic distances can be used to consistently circumscribe species at the 5% level.

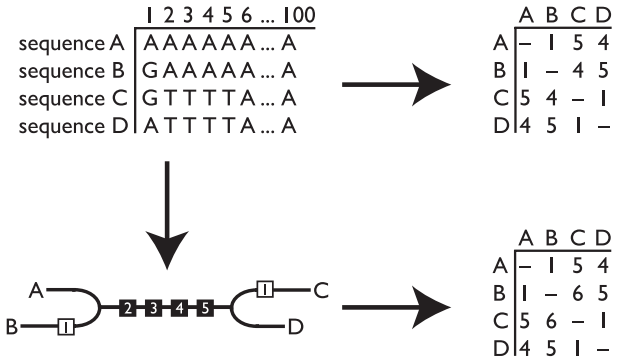

Fig. 1. The effect of enchainment on the "percent divergence" species concept. A hypothetical situation in which it is not possible to circumscribe one or more species that simultaneously incorporate all sequences with less than 5% divergence while at the same time not having 5% or more divergence within a species. Pair-wise Manhattan distances are shown to the right of the sequences, the unrooted most parsimonious tree (one of the two possible optimizations) along with the patristic distances derived from the tree (and optimization) are shown below.

When confronted with data similar to that in Fig. 1, circumscription methods such as PAQ (Baccam et al., 2001) and TaxonDNA (Meier et al., in press) can be used to generate stable groupings of similar sequences. Unfortunately, these algorithms must make sacrifices to achieve stability: PAQ produces overlapping groups that conform to the threshold while TaxonDNA violates the threshold when necessary to produce non-overlapping groups. To remedy the situation illustrated in Fig. 1, Blaxter et al. (2005) advocate the use of a modified CLOBB algorithm (Parkinson et al., 2002) in combination with randomization of input order. Although randomization could eventually produce all possible circumscriptions that are compatible with the data, randomization in-and-of itself does not provide an objective criterion to choose between alternate circumscriptions, rather it merely identifies such situations. Blaxter et al. indicate that the production of unstable groupings is a desirable feature of their algorithm, rather than a failing—this is not the case.

As percent divergence cannot reliably be used to diagnose species (Fig. 1; Ferguson, 2002) and non-arbitrary reproducible algorithms for species delimitation are available [e.g., population aggregation analysis (Nixon and Wheeler, 1990; Davis and Nixon, 1992), exact F_{ST} (Raymond and Rousset, 1995)] there is no logical reason for the use of arbitrary delimitations.

Lack of molecular variation

The low level of molecular differentiation among some species (e.g., relatively recently diverged species) in effect limits the potential application of DNA barcoding to a subset all of the species currently recognized (Lipscomb et al., 2003; Sperling, 2003; Tautz et al., 2003; Moritz and Cicero, 2004; Will and Rubinoff, 2004; DeSalle et al., 2005; Vences et al., 2005b; Meier et al., in press). Advocates of DNA barcoding who wish to use a strict percent divergence threshold, dismiss this criticism as irrelevant as they would not recognize these entities as distinct species even though biological species *sensu* Mayr (1957) or phylogenetic species *sensu* Nixon and Wheeler (1990) could, in theory, differ only by a single nucleotide change.

Selection of the barcode locus

Selection of the barcode locus is controversial (Mallet and Willmott, 2003; Tautz et al., 2003; Blaxter, 2004; Thalmann et al., 2004; DeSalle et al., 2005; Kress et al., 2005; Vences et al., 2005b). Criticisms of proposed loci include the lack of universal primers, non-ubiquitous loci, alignment issues, multiple copies per individual, and lack of appropriate levels of variation within some taxonomic groups. It is unlikely that any one locus will be sufficient for identifications across all of life (Tautz et al., 2003), but limiting the number of loci while at the same time maximizing the number of species possessing easily detectable loci (i.e., the locus is present and can be polymerase chain reaction-amplified with "universal" primers) is of great importance.

The mitochondrially encoded gene cytochrome c oxidase subunit I (*COI*) has been the locus of choice for most of the recent exemplar studies (Wells and Sperling, 2001; Wells et al., 2001a,b; Hebert et al., 2003, 2004a,b; Hogg and Hebert, 2004; Whiteman et al., 2004; Ball et al., 2005; Barrett and Hebert, 2005; Lambert et al., 2005; Lorenz et al., 2005; Meyer and Paulay, 2005; Monaghan et al., 2005; Saunders, 2005; Smith et al., 2005; Steinke et al., 2005; Vences et al., 2005a,b; Ward et al., 2005; Hajibabaei et al., 2006; Smith et al., 2006). However, portions of the nuclear ribosomal DNA (LSU, SSU or ITS) have been extensively used for barcoding (Fox et al., 1977; Niesters et al., 1993; Li et al., 1995; Poinar et al., 1998; Amato et al., 1999; Jackson et al., 1999; Fell et al., 2000; Hofreiter et al., 2000; Kõljalg et al., 2000, 2005; Floyd et al., 2002; Vrålstad et al., 2002; Rosling et al., 2003; Tedersoo et al., 2003; De Ley et al., 2005; Kopchinskiy et al., 2005; Kress et al., 2005; Markmann and Tautz, 2005; Monaghan et al., 2005; Steinke et al., 2005; Vences et al., 2005a,b; Smith et al., 2006). Regions of plastid DNA have also been used for plant barcoding (Poinar et al., 1998; Hofreiter et al., 2000; Chase et al., 2005; Kress et al., 2005). From the limited data available for land plants, it appears that *COI* is an inappropriate choice for identification of most species in this kingdom (Cho et al., 2004).

The use of mitochondrially encoded regions has been criticized because of the prevalence of mitochondrial DNA insertions into the nuclear genome of some animals (Thalmann et al., 2004). The use of other loci has drawn criticism due to the possibility of ancestral polymorphisms, lineage sorting, incomplete concerted evolution, and orthology/paralogy concerns (Mallet and Willmott, 2003; Ball et al., 2005; Chase et al., 2005; Kress et al., 2005). These criticisms conflate phylogeny reconstruction with specimen identification. Although hierarchical clustering algorithms (e.g., neighbor joining, parsimony, maximum likelihood, etc.) could be used to identify query sequences, the aim is not to produce a phylogeny rather it is to provide an identification. Unlike phylogenetic methods, diagnostic techniques are not constrained to operate only with characteristics presumed to be homologous—characteristics known to be non-homologous can be used and are often more informative than strictly homologous characteristics (e.g., opposite leaves have more than one evolutionary origin within angiosperms, but this characteristic is often among the first used in diagnostic keys). Thus, the number of copies of the barcoding locus, their physical location in the genome, and the topology of the resulting gene tree are irrelevant to the identification process provided that the reference database contains all of the detectable copies found in a particular taxon.

Identification using multidimensional scaling

The use of multidimensional scaling for identification of query sequences is problematic. Circumstances likely to occur in real data sets that will result in ambiguous identification or no identification have been discovered (Will and Rubinoff, 2004). It is unlikely that genetic distance will be uniform among species at the barcode locus (Meyer and Paulay, 2005; Meier et al., in press). In addition, the occurrence of shared haplotypes at the barcode locus due to ancestral polymorphism will probably be observed in some instances (Funk and Omland, 2003; Meyer and Paulay, 2005). As a direct consequence of these two properties, if the query sequence is placed in a region of the multidimensional scaling plot where two species overlap or if the query sequence falls outside of the circumscription for any known species, no identification can be made (Will and Rubinoff, 2004). Given these problems, multidimensional scaling has not been widely used for DNA barcoding, instead most studies use some form of hierarchical clustering or less commonly a similarity method.

Identification using hierarchical clustering

Hierarchical clustering methods (e.g., neighbor joining, parsimony, maximum likelihood) are used to identify query sequences by first aligning the query sequence to a set of reference sequences and then calculating or searching for a topology or set of topologies. Once a hierarchy has been constructed, group membership variables (Farris, 1974) must be optimized on to the hierarchy in order to determine the identity of the query sequence. This step is often accomplished intuitively rather than explicitly—leading to disagreements about the success or failure of a particular identification (e.g., Will and Rubinoff, 2004 and Meier et al., in press, versus Hebert et al., 2003).

If a hierarchical clustering method is used for query sequence identification, there are at least three ways in which an ambiguous identification can result.

1 If the query sequence is resolved as sister to a known group, an unambiguous identification cannot be made (Fig. 2a; Will and Rubinoff, 2004).

2 If the gene tree of the barcode locus does not match the classification such that a group of sequences is para- or polyphyletic (*sensu* Farris, 1974) in the gene tree and the query sequence is resolved in a position between the non-monophyletic group and another group, an unambiguous identification cannot be made (Fig. 2b). Whether or not the group of sequences represents a "non-monophyletic" species or a para- or polyphyletic higher taxon is irrelevant as the gene tree is the only source of data used to make an identification. It would be undesirable to change the species circumscription so that species are made "monophyletic" as this does not appear to reflect biological reality (Crisp and Chandler, 1996; Funk and Omland, 2003). In addition, it would be counter-productive to change the classification if a higher taxon was in fact monophyletic in the underlying species tree and the barcode gene tree simply does not reflect this pattern (e.g., due to lineage sorting).

3 An unresolved tree (Fig. 2c) can also produce an ambiguous identification.

The frequency of ambiguous identifications could be reduced by the consistent use of either a "fast" (a.k.a. ACCTRAN) or "slow" (a.k.a. DELTRAN) optimization rather than using an unambiguous optimization (the intersection between fast and slow)—this is an arbitrary and empirically unsupported decision that does not actually eliminate the underlying cause of ambiguity.

Methods that are implemented in such a way as to return only a single suboptimal tree (e.g., most implementations of neighbor joining) can exasperate the problems outlined above as well as result in outright misidentification particularly when an ambiguous identification is the only conclusion that can be supported by the data. Variation in alignment can also result in incorrect identification—even when the query sequence is identical to one in the reference database.

Identification using similarity methods

BLAST and related similarity methods compare the query sequence with the sequences in an unaligned reference database using a pair-wise partial alignment or

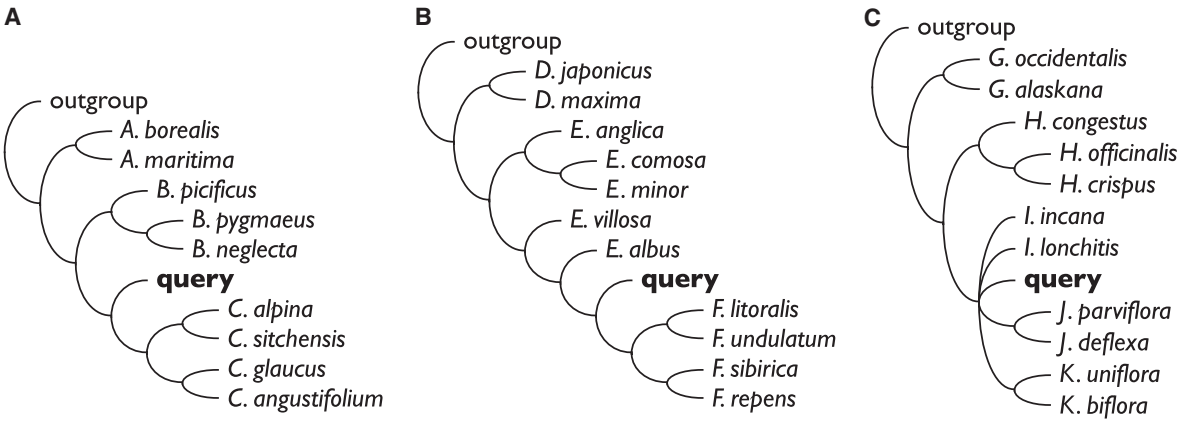

Fig. 2. Hypothetical phylogenetic (or phenetic) trees demonstrating situations in which ambiguous identification of query sequences will result (assuming unambiguous optimization of group membership variables). (A) A case in which the query sequence is sister to genus C and therefore cannot be assigned to either genus B or C (identification to species cannot be made either; Will and Rubinoff, 2004). (B) A case in which the gene tree does not match the current classification resulting in genus E appearing to be paraphyletic on the gene tree. The query sequence cannot be unambiguously assigned to either genus E or F. In addition, no identification to species can be made. (C) A case in which the query sequence cannot be identified to species due to lack of resolution (caused by either lack of data or conflicting data).

nearly exact matches of short nucleotide strings (e.g., 11 nucleotides). A similarity score is computed from the portion of the query aligned to the reference sequence. The reference sequence(s) with the highest similarity score is(are) usually presented along with some indication of the probability of drawing such a match given the properties of the reference sequences and the length of the match (e.g., *E*-value). The ability of similarity methods (e.g., BLAST, BLAT, FASTA, megaBLAST) to alleviate the problems associated with multidimensional scaling and hierarchical clustering has not been extensively discussed in the context of sequence identification for DNA barcoding (Prendini, 2005)—although BLAST (Altschul et al., 1990) or FASTA (Pearson and Lipman, 1988) have been used in some cases (e.g., Poinar et al., 1998; Wells et al., 2001b; Rosling et al., 2003; Tedersoo et al., 2003; De Ley et al., 2005; Kõljalg et al., 2005; Kopchinskiy et al., 2005; Kress et al., 2005; Markmann and Tautz, 2005; Vences et al., 2005a,b).

The various manifestations of BLAST are known to provide inconsistent (Anderson and Brass, 1998; Woodwark et al., 2001) and even incorrect (Agarwal and States, 1998; Anderson and Brass, 1998; Koski and Golding, 2001) sequence identifications under some real world circumstances. Some query sequence misidentifications can be ascribed to an incomplete reference database (Koski and Golding, 2001) or the parameters used to construct pair-wise alignments (Anderson and Brass, 1998; Woodwark et al., 2001).

Recently, Steinke et al. (2005; TaxI) and Meier et al. (in press; TaxonDNA) have proposed methods that rely on pair-wise alignment of the query sequence to an aligned reference database. The reference sequence(s) with the smallest pair-wise distance from the query sequence is taken to be the identification. Distance between the reference and query sequence can be calculated in various ways depending upon the method. In some cases, statistical tests can be used to determine if the closest matching reference sequence(s) are sufficiently similar to be considered a valid identification [e.g., "best close match" of Meier et al. (in press)]. In contrast to BLAST and related similarity methods, TaxI and TaxonDNA rely on an aligned reference database. As a result, disparate elements of an alignment variable region are difficult to meaningfully accommodate within the same database. In addition, the consistent alignment of a large number of sequences from such a region (> 10 000) is computationally difficult.

All similarity methods can produce an ambiguous identification when a query sequence equally matches more than one sequence in the reference database (similar to the problems with multidimensional scaling discussed above).

Identification using similarity methods in combination with hierarchical clustering

Disregarding the interpretation issues outlined in Fig. 2, computational barriers prevent the widespread use of hierarchical clustering methods to provide identifications for DNA barcoding because (1) the consistent alignment of large data matrices (> 10 000 sequences) from an alignment variable region is computationally difficult, and (2) the production of trees (calculation of starting tree plus branch swapping) for such large matrices is difficult to impossible using current software and hardware. To evade current computational limits, BOLD-ID (Hajibabaei et al., 2005) uses BLAST guided by "hidden Markov models based on a global protein alignment for the *cox1* [*COI*] gene" to select 100 sequences from the reference database that best match the query sequence. The resulting sequences can then be further analyzed with a distance tree to determine the final identification. An exact description of the methods used by BOLD-ID have not been published to date.

Identifications using diagnostic methods

A standard morphological approach to species identification could be adapted to DNA barcoding. Traditionally, taxonomists have intuitively constructed diagnostic descriptions from morphological data in such a way as to provide a unique set of characteristics by which to unambiguously identify specimens. That same logic can be used to construct a diagnosis using only DNA sequences. The union of available sequences should be used to construct the species diagnosis (i.e., a list of unique sequences or sequence motifs) just as morphological diagnoses are currently based on a sample of specimens empirically determined to constitute a reasonable sample of a species' diversity.

Using the available gymnosperm nrITS 2 data we can, for example, diagnose *Zamia wallisii* A. Braun as having the distinguisher CTTGCTCCTT at nrITS 2. Our current knowledge indicates that all individuals of *Z. wallisii* have this motif at nrITS 2 and that no other gymnosperm species have such a motif at this locus. This type of diagnosis is designed to be useful without the need to align the sequences and can be constructed such that the sequences do not have to be in the same +/− orientation. DNA sequences from unknown specimens can be identified by searching for diagnostic distinguishers that are recorded in the reference database. Owing to lack of sequence variability, not all species will have unique DNA distinguishers useful for differentiating them from other species. Species without unique distinguishers could be diagnosed with unique combinations of more common distinguishers—as is often done with morphology-based diagnoses.

DNA diagnoses can be constructed by enumeration of all possible distinguishers, but as the reference database increases in size such an approach becomes computationally impractical. Methods to construct diagnoses from sequence data have been proposed recently (Borneman et al., 2001; Rash and Gusfield, 2002; DasGupta et al., 2005a,b; Gibbs et al., 2005). Although these methods originally were designed to select oligonucleotide sequences for use in a Southern hybridization-based identification system, the probe sequences can just as easily be incorporated into a system that uses DNA sequences directly. As a procedure that uses DNA sequences directly is not constrained by oligonucleotide characteristics (e.g., melting temperature) there are potentially more useful diagnostic characteristics available.

Alignment-free tree-based identification

Another possible solution to alleviate the need for sequence alignment is a hybrid between the existing hierarchical clustering and similarity methods. We propose that each sequence in the reference database could be scored for the presence/absence of all possible short sequence motifs in both orientations (+/−). After the addition of a similarly scored query sequence the matrix could be subjected to some type of cladistic analysis (e.g., parsimony). Because this coding method does not require alignment it discards information about the relative order of motifs. As a direct result, it is likely that some sequences will be scored as having a given motif that an alignment would indicate are non-homologous. This method would suffer from the interpretation problems outlined in Fig. 2 as well as over/underweighting as a result of sampling the same nucleotide change repeatedly (i.e., the standard assumption of character independence is violated). Branch support for this matrix could not be correctly calculated using standard resampling techniques (e.g., jackknife) without the use of some sort of correction factor. The trees resulting from the analysis of this matrix should not be interpreted to represent the underlying phylogeny, rather the trees represent sequence similarity, which may or may not be concordant with the phylogeny. Potentially this scoring method could result in misidentification if the query sequence differs substantially in length and therefore its set of motifs—shorter query sequences will spuriously be scored as absent for motifs that "should" be present, and longer query sequences will spuriously be scored as present for motifs that "should not" be present.

Modifications to the BLAST algorithm (Altschul et al., 1997), novel but related similarity methods [e.g., megaBLAST (Zhang et al., 2000) and BLAT (Kent, 2002)], and diagnostic methods [e.g., DNA-BAR (DasGupta et al., 2005a)] have been published recently. A comparison designed to measure precision and accuracy of identifications made by the available DNA barcoding methods was conducted.

Materials and methods

Gymnosperm nuclear ribosomal DNA internal transcribed spacer 2 sequences (nrITS 2) available in GenBank were used to construct a test data set (sequences downloaded May 12, 2005; some sequences used in the analysis also included nrITS 1 and/or a small portion of 18S and/or a small portion of 26S). A second test data set was constructed from publicly available gymnosperm sequences of plastid-encoded tRNA-Lys (*trnK*) gene, intron and embedded maturase K (*matK*) downloaded from GenBank July 29, 2005 (sequences used in the analysis included some portion of *matK* and/or *trnK* and/or surrounding *trnK* intron). Identical sequences derived from the same species were reduced to a single sequence. Sequences were excluded if they were not identified to species or if they were derived from interspecific hybrids (Appendices 1 and 2).

Four reference databases were constructed from these sequences. Two "full databases" included either 1037 nrITS 2 sequences from 413 species in 71 genera or 522 *matK* sequences from 334 species in 75 genera. Generic and species circumscriptions used here follow the taxonomy used in GenBank (intraspecific taxa were ignored). The "restricted databases" contained either 413 nrITS 2 sequences or 334 *matK* sequences—one sequence per species. If multiple sequences were available from a given species, the most complete sequence was used in the restricted database. A length/completeness score (sequence information content) was calculated from the number of unambiguous nucleotides (A, C, G or T) multiplied by 4, plus the number of twofold degenerate base symbols (K, M, R, S, W or Y) multiplied by 2, plus the number of threefold degenerate base symbols (B, D, H or V). If multiple sequences derived from a given species had an identical sequence information content score, then an arbitrary selection among the highest scoring sequences was made.

Pair-wise divergence between reference sequences was calculated by aligning each pair of sequences with MUSCLE version 3.52 (Edgar, 2004) using default parameters. The Manhattan metric was used to calculate percent divergence with sites containing gaps and/or full polymorphisms (i.e., N) in one or both of the sequences excluded from consideration. Subset polymorphisms in the sequences (i.e., B, D, H, K, M, R, S, V, W and Y) were counted as a single change if there was not an intersection between the two sequences at that position.

Hierarchical clustering

Multiple sequence alignments of the full and the restricted reference databases were constructed using the default parameters in MUSCLE. These alignments were used to measure precision by querying all nrITS 2 or *matK* sequences against the appropriate full reference database while accuracy was estimated by querying all nrITS 2 or *matK* sequences against the appropriate restricted reference database (assuming consistent taxonomic identification). Each query sequence was aligned to the appropriate multiple sequence alignment using the "-profile" option in MUSCLE. The resulting output was converted to a format appropriate for TNT version 1.0 (Goloboff et al., 2004). A 200 iteration parsimony ratchet tree search (Nixon, 1999) was conducted in TNT holding a single tree. Ten percent of the informative characters were reweighted using a probability of 5 for up-weighting and a probability of 5 for down-weighting with other parameters set to the default ("xi; rs0; col3; ho200; rat:it200upf5dow5numsubsX; mu = rep1ho1rat;" where X = 10% of the number of informative characters). A fast tree search—calculation of a Wagner tree followed by SPR swapping while holding one tree—was also conducted with TNT ("xi; rs0; col3; ho1; mu = rep1ho1spr;"). Only one tree was held for each of the parsimony analyses to make results more comparable those of neighbor joining. The output from MUSCLE was also converted to an input file for the PHYLIP package version 3.63 (Felsenstein, 2004). The program dnadist was used to calculate a distance matrix using the Jukes–Cantor model of nucleotide substitution (Jukes and Cantor, 1969) with one substitution rate category (dnadist crashed when asked to apply other substitution models to the data). The resulting distance matrix was used to construct a tree using the method of Saitou and Nei (1987) as implemented in the program neighbor using the default parameters. The placement of the query sequence, in the trees derived from TNT and neighbor, was scored such that the least inclusive clade containing the query was taken to be the intended identification thereby eliminating the difficulties demonstrated above (Fig. 2) by arbitrarily selecting a "fast" (ACCTRAN) optimization of the group membership variables. The identification was taken to include all sequences in the clade.

Similarity methods

Precision of similarity methods was estimated by querying all nrITS 2 or *matK* sequences against the appropriate full reference databases and accuracy was estimated by querying all sequences against the restricted reference database. Queries used the BLASTn algorithm as implemented in blastall version 2.2.10 (Altschul et al., 1997), BLAT version 32 (2005 January 31; Kent, 2002), and megaBLAST version 2.2.10 (Zhang et al., 2000). When necessary, it was specified that the query sequence and the reference databases were composed of DNA sequences, that up to 100 sequences were returned for each query, and that the output be in the "traditional blast" style, otherwise the default parameters were used for all similarity programs.

Combination methods

The top 100 BLAST hits generated by each query sequence (output derived as described for the "similarity methods" above) were aligned using the default parameters in MUSCLE. The resulting matrix of 101 sequences was analyzed in TNT and neighbor and then scored for precision and accuracy as described for the "hierarchical clustering" methods above.

Diagnostic methods

In order to generate a presence/absence matrix of distinguishers (sequence strings), the source code of "degenbar", which implements the DNA–BAR method of DasGupta et al. (2005a) was modified to accept a maximum of 10 000 sequences. An input file for degenbar was constructed with "NumTargets" and "NumSources" set to the total number of sequences in the input file, "Redundancy" set to 10, "l_min" set to 10, "l_max" set to 50, "MinCandidTemp" set to 0, "MaxCandidTemp" set to 100, "MinCandidGC" set to 0, "MaxCandidGC" set to 100, "SaltConc" set to 50, "DNAConc" set to 50, "MaxCommSubstrWt" set to 100, and "MinMaskLength" set to 1. Each sequence and its reverse complement (separated by 50 "N" symbols) was input. The presence/absence matrix of distinguishers in the degenbar output was used as a reference database. Each query sequence was scored for the presence or absence of each distinguisher (10–50 nucleotide in length). The reference sequence(s) with the greatest number of matching presence/absence scores was(were) taken to be the identification. Precision was estimated by querying all nrITS 2 or *matK* sequences against the appropriate matrix of distinguishers derived from a full reference database and accuracy was estimated by querying all sequences against a matrix of distinguishers derived from a restricted reference database.

A diagnostic DNA barcode database for DOME ID (Diagnostic Oligo Motifs for Explicit IDentification) was constructed:

1 All sequence strings of 10 nucleotides offset by five nucleotides were extracted from the reference sequences. (Both overlapping and non-overlapping strings were used in preliminary trials. The use of overlapping strings resulted in a greater proportion of taxa with diagnostics distinguishers, but at the cost of greater computational time. Preliminary trials also indicated that a string

length of 10 nucleotides resulted in a high proportion of taxa with diagnostics distinguishers with a minimum computational burden.)

2 Each string was classified as diagnostic or non-diagnostic by searching among the nrITS 2 or *matK* reference sequences for the string and its reverse complement. A string was considered diagnostic if it and its reverse complement occurred only in sequences belonging to a single species.

3 Diagnostic strings were inserted into the diagnostic barcode database. Non-diagnostic strings were ignored. If no diagnostic distinguishers could be found for a particular taxon it was not possible to identify query sequences from that taxon—this feature eliminates some false or ambiguous identifications at the expense of taxonomic depth.

Precision of DOME ID was estimated by querying the nrITS 2 or *matK* sequences in the restricted reference database against the barcode database derived from those same sequences. The full reference database could not be used to measure precision because it was constructed to reflect a species in its known entirety rather than the individual sequences belonging to the species—because the restricted reference database includes only one sequence per species it could be used to measure precision. Accuracy was estimated by querying all nrITS 2 or *matK* sequences against the appropriate restricted reference database.

Queries to the DOME ID reference database were conducted using the following algorithm:

1 All contiguous overlapping sequence strings of 10 nucleotides were extracted from the query sequence in both orientations. (Completely overlapping strings where used so that query sequences need not be in the same register as the reference sequences.)

2 Each string extracted from the query sequence was checked against the barcode database.

3 If a query sequence matched more than one taxon in the barcode database, the sequence with the greatest number of sequence string matches was taken to be the identification. In the case of ties, the identification was ambiguous. If none of the query strings matched anything in the reference database, the sequence was considered unidentified.

Alignment-free tree-based identification

ATIM was tested by scoring each sequence in the reference database for the presence/absence of all 1 048 576 possible motifs of 10 nucleotides in length—both orientations of each reference sequence was examined. Presumed most parsimonious trees were obtained for each matrix using 100–1000 parsimony ratchets in TNT (100–200 iterations each, one to 20 tree(s) held per ratchet, 10% of the informative characters re-weighted each iteration, a probability of five for up-weighting, and a probability of five for down-weighting; "xi; rs0; col3; ho20000; rat:it200upf5dow5numsubsX; mu = rep100ho1rat;" where X = 10% of the number of informative characters). Query sequences were scored in the same manner as the reference sequences and appended to the matrix. The resulting matrix was opened in TNT with the "nstates 8;" command in order to reduce amount of RAM required. The strict consensus ("nel*;") of the presumed most parsimonious trees derived from the reference matrix was read into memory and used as a skeleton tree for positive constraints ("force/& 0; k0; const = ;"). A single replicate TBR tree search (holding 20 trees) was used to find the optimal placement of the query sequence ("xi; rs0; col3; ho21; mu = rep1ho20;"). The strict consensus was scored using the same criteria used for the "hierarchical clustering" methods (described above). Precision and accuracy were estimated by querying all sequences against matrices derived from the appropriate full and restricted reference databases.

Estimate of relative analysis time

Time trials were conducted on a 3.06 GHz Intel Pentium® 4 with 1 GB of RAM running Ubuntu Linux 5.04 (Hoary Hedgehog). PERL scripts were used to track the time required for program execution (blastall, BLAT, a PERL script that interprets the degenbar output, dnadist, a PERL script that implements DOME ID using MySQL 4.1.10a, megaBLAST, MUSCLE, neighbor and TNT). The time required to format query sequence input and convert output from one executable to input for another executable was not included in execution time calculations.[1] Also, the one-time formatting of the various reference databases was excluded from execution time calculations (i.e., alignment of all sequences in the reference database for clustering methods, creating the presence/absence matrix of distinguishers for DNA–BAR, generating the diagnostic DNA barcode database for DOME ID, and scoring the presence/absence of all possible 10 nucleotide motifs for the ATIM reference database). The full *matK* reference database was used in conjunction with all *matK* sequences from 10 arbitrarily selected genera (*Agathis*, *Bowenia*, *Cedrus*, *Chigua*, *Pseudotaxus*, *Stangeria*, *Taxodium*, *Thuja*, *Tsuga* and *Welwitschia*) for a total of 29 query sequences.

Results

The median percent pair-wise divergence (Manhattan metric), the percentage of zero divergence comparisons,

[1]If a particular procedure were to be widely used for DNA barcoding, a single program should be created for that purpose thereby eliminating these script kluges that allow interoperability.

and the number of parsimony informative characters did not vary substantially between the full and restricted reference databases for a given locus (Table 1). Aligned matrices were noticeably larger than the median unaligned length (Table 1). The number of redundant sequences (completely identical to at least one other sequence—including any implied indels) decreased post-alignment (Table 1). The presence of these redundant sequences contributed to ambiguous identifications for all methods.

Hierarchical clustering

Within data sets, the clustering methods performed at approximately the same level of precision and accuracy; however, none of the clustering methods were able to accurately identify query sequences to species at a high frequency—especially when ambiguous identifications are considered (median = 50% correct; Tables 2 and 3). Even after accounting for ambiguities, accurate identification to genus occurred at a high frequency (median = 97% correct; Tables 2 and 3). Including ambiguities, precision of clustering methods was not particularly high (median = 18% correct; Tables 2 and 3). Presumably unambiguous accuracy and precision would fall even further for all clustering methods if more than one tree was returned per search. Despite their similar performance, computer execution time varied greatly among the clustering methods. Execution time was highest for the parsimony ratchet (≈ 71 times longer than the SPR tree search), followed by neighbor joining (including the calculation of the pair-wise distance matrix; ≈ 9 times longer than the SPR tree search), and the SPR tree search in TNT (Table 4).

Similarity methods

The performances of BLAST, BLAT, and mega-BLAST were very similar in most cases with most of the variation occurring between, rather than among, data sets (Tables 2 and 3). Of the similarity methods evaluated here, all were relatively precise (94–99% correct; Tables 2 and 3) provided that ambiguity—due to multiple reference sequences producing the same similarity score—was not considered. Precision fell noticeably (median = 75% correct; Tables 2 and 3) when ambiguity was accounted for. For nrITS 2 approximately 5% of the precision lost to ambiguity can be related to the presence of 51 redundant sequences in the restricted reference database—the figure is approximately 14% for *matK* (75 redundant sequences). Although the similarity methods were more accurate than the existing clustering methods, none of the similarity methods were able to accurately identify query sequences to species (Tables 2 and 3) at a high frequency—especially when ambiguous identifications

Table 1
Properties of the nrITS 2 and *matK* data sets. Percent pair-wise divergence (Manhattan metric), pair-wise zero divergence comparisons, unaligned and aligned sequence length, redundant sequences, and parsimony informative characters. Where appropriate, data are presented in the form of median (interquartile range)

	nrITS 2		matK	
	Full	Restricted	Full	Restricted
Pair-wise divergence	30.99% (26.53–34.48%)	29.39% (25.75–33.30%)	20.39% (5.95–23.30%)	21.38% (8.13–23.89%)
Zero divergence comparisons	0.09%	0.21%	0.54%	0.42%
Unaligned length	137 (108–250) bp	196 (115–260) bp	1561 (1412–1661) bp	1601 (1530–1661) bp
Aligned length	8733 bp	6778 bp	3975 bp	3906 bp
Unaligned redundant sequences	51	18	75	47
Aligned redundant sequences	41	11	71	43
Parsimony informative characters	3720	3145	2364	2144

Table 2
Precision and accuracy of query sequence identification methods as estimated with gymnosperm nrITS 2 sequences (Appendix 1). Numbers outside parentheses represent the percentage of correct identifications including ambiguous identifications. Numbers inside parentheses represent percentage of unambiguous correct identifications. The best performance is indicated in **bold face**

Method	Precision	Accuracy to genus	Accuracy to species
Clustering methods			
parsimony ratchet	58% (13%)	98% (95%)	67% (46%)
SPR search	60% (11%)	98% (96%)	69% (47%)
neighbor joining	65% (8%)	97% (91%)	68% (42%)
Similarity methods			
BLAST	94% (81%)	**100% (100%)**	67% (63%)
BLAT	94% (82%)	99% (99%)	66% (62%)
megaBLAST	94% (80%)	95% (95%)	72% (68%)
Combination methods			
BLAST/ parsimony ratchet	86% (74%)	99% (98%)	78% (67%)
BLAST/SPR	87% (73%)	100% (99%)	79% (67%)
BLAST/ neighbor joining	93% (71%)	99% (97%)	80% (64%)
Diagnostic methods			
DNA–BAR	**98% (89%)**	86% (86%)	65% (62%)
DOME ID	80% (80%)*	86% (84%)†	67% (66%)‡
ATIM	100% (83%)	99% (98%)	**83% (71%)**

*100% (100%) excluding unidentified sequences.
†98% (96%) excluding unidentified sequences.
‡76% (75%) excluding unidentified sequences.

Table 3
Precision and accuracy of query sequence identification methods as estimated with gymnosperm *matK* sequences (Appendix 2). Numbers outside parentheses represent the percentage of correct identifications including ambiguous identifications. Numbers inside parentheses represent percentage of unambiguous correct identifications. The best performance is indicated in **bold face**

Method	Precision	Accuracy to genus	Accuracy to species
Clustering methods			
parsimony ratchet	71% (41%)	100% (99%)	77% (60%)
SPR search	70% (41%)	99% (98%)	78% (58%)
neighbor joining	44% (23%)	99% (97%)	75% (52%)
Similarity methods			
BLAST	99% (67%)	**100% (100%)**	**84% (68%)**
BLAT	99% (69%)	99% (99%)	82% (67%)
megaBLAST	99% (61%)	100% (99%)	84% (64%)
Combination methods			
BLAST/ parsimony ratchet	77% (55%)	100% (99%)	80% (60%)
BLAST/SPR	76% (53%)	100% (99%)	78% (61%)
BLAST/ neighbor joining	95% (56%)	100% (99%)	86% (56%)
Diagnostic methods			
DNA–BAR	**100% (79%)**	96% (96%)	73% (62%)
DOME ID	60% (60%)*	53% (53%)†	50% (50%)‡
ATIM	100% (67%)	98% (97%)	87% (53%)

*100% (100%) excluding unidentified sequences.
†97% (97%) excluding unidentified sequences.
‡90% (90%) excluding unidentified sequences.

Table 4
Computer time (in seconds) required for each identification method estimated from a sample of 29 *matK* sequences run against the full reference database (522 sequences; run on a 3.06 GHz Intel Pentium® 4 with 1 GB of RAM). Data have a resolution of 1 s

Method	Time (s)		
	25th percentile	Median	75th percentile
Clustering methods			
parsimony ratchet	841	855	864
SPR search	11	12	12
neighbor joining	110	111	112
Similarity methods			
BLAST	1	1	2
BLAT	1	1	2
megaBLAST	> 0	1	1
Combination methods			
BLAST/parsimony ratchet	186	219	278
BLAST/SPR	170	196	262
BLAST/neighbor joining	171	198	264
Diagnostic methods			
DNA–BAR	1	1	1
DOME ID	16	16	22
ATIM	132	133	134

are considered (median = 66% correct; Tables 2 and 3). Given the relative precision of the similarity methods a minimum of ≈37% accuracy (≈ 32% including ambiguity) was expected for nrITS 2 and ≈63% accuracy (≈ 43% including ambiguity) for *matK*—the various similarity methods were able exceeded this by 29–36% for nrITS 2 and 19–25% for *matK*. Computer execution time for the similarity methods was uniformly fast (2 s maximum; Table 4).

Combination methods

The performance of combination methods was very similar in most cases with most of the variation occurring between data sets (Tables 2 and 3). When ambiguity was accounted for, precision and accuracy of the combination methods was between that of the poorly performing clustering methods and the somewhat better performing similarity methods (Tables 2 and 3). However, none of the combination methods was able to perform better than the similarity methods. The computer time required for execution did not differ appreciably between the different combination methods, but was noticeably longer (≈ 200 times) than the better performing similarity methods (Table 4).

Diagnostic methods

The DNA–BAR algorithm found 1997 distinguishers (sequence strings) to differentiate between the sequences in the full nrITS 2 reference database—unique combinations of distinguishers could not be found for 54 (5%)

of the sequences. For the full *matK* reference database, 808 distinguishers were found to differentiate between the sequences—the set of distinguishers for 75 (14%) of the sequences were redundant. Only 49 of the 1997 nrITS 2 distinguishers were restricted to a single sequence. Eight of the 808 *matK* distinguishers were similarly restricted. Sequences in the restricted nrITS 2 database were differentiated by 813 distinguishers—distinctions between 19 (5%) of sequences could not be made. There were 27 unique distinguishers. Sequences in the restricted *matK* database were differentiated by 582 distinguishers—distinctions between 47 (14%) of sequences could not be made. There were 10 unique distinguishers. Unambiguous precision of DNA–BAR was higher than any other method (79–89% unambiguously correct; Tables 2 and 3), but the frequency of correct species-level identification was the same or lower than the worst performing similarity method (62% unambiguous; Tables 2 and 3). Given that the DNA–BAR algorithm was unable to find distinguishers for some of the sequences, approximately 5% ambiguity was expected for the nrITS 2 data set and 14% was expected for the *matK* data set—more ambiguity was observed during the estimation of precision, but less ambiguity was observed during the estimation of accuracy.

Execution time of the PERL script used to compare the query sequence to the presence/absence matrix of distinguishers for DNA–BAR was relatively fast (median = 1 s)—equivalent to the time required to conduct similarity searches (Table 4). The use of a compiled rather than interpreted programming language (e.g., C) to compare the query with the reference database would undoubtedly further improve the performance of the DNA–BAR algorithm. The DNA–BAR analysis time does not, however, include time required to generate the presence/absence matrix of distinguishers, which is rather computationally costly (DasGupta et al., 2005b) when compared with the simple formatting and conversion required for the similarity methods.

Unique 10 nucleotide distinguishers were present for 330 (80%) of the species in the nrITS 2 restricted reference database. For the *matK* restricted reference database, unique distinguishers were found for 200 (60%) of the species. DOME ID precision was relatively high (100% correct) provided unidentifiable sequences were not considered; however, precision fell noticeably (median = 70% correct) when unidentifiable sequences were counted (Tables 2 and 3). If ambiguous identifications and species without unique distinguishers are excluded from consideration, DOME ID produced notable accuracy (median = 83% correct; Tables 2 and 3); however, when all species are included and ambiguity is accounted for, median accuracy fell to 58% correct—somewhat worse than median of the similarity methods (66% correct), but better than the median of the clustering methods (50% correct; Tables 2 and 3). DOME ID had the smallest amount of ambiguity for species-level identifications. Execution time for DOME–ID searches was longer than that of DNA–BAR searches (Table 4).

Alignment-free tree-based identification

Analysis of the matrix derived from the nrITS 2 full reference database resulted in 18 presumed most parsimonious trees 399 920 steps long (informative only). The nrITS 2 restricted reference database produced 700 presumed most parsimonious trees 166 155 steps long (informative only). Analysis of the matrix derived from the *matK* full reference database produced 40 presumed most parsimonious trees 225 331 steps long (informative only). The *matK* restricted reference database produced 1437 presumed most parsimonious trees 208 404 steps long (informative only).

ATIM precision was relatively high, but not remarkable (Tables 2 and 3). ATIM accuracy to genus was relatively high (median = 98%; Tables 2 and 3), although other methods (e.g., BLAST and BLAT) had higher accuracy to genus. After accounting for ambiguity ATIM had the greatest accuracy to species for the nrITS 2 data set—this was not so with the *matK* data set (Tables 2 and 3). Although ATIM analysis time was ≈ 11 times longer than an SPR tree search, it was not as slow as the slowest clustering method (parsimony ratchet) and provided substantially greater precision and accuracy (Table 4).

Discussion

The method used here to calculate pair-wise divergence is not directly comparable with the methods used in other publications, but in general it appears that the pair-wise divergence in gymnosperm nrITS 2 and *matK* sequences is much greater (median of full reference databases = 30.99% and 20.39%, respectively) than many of the data sources used in previous studies. For example, divergence (Kimura-2-parameter model) "averaging 6.8% for congeneric taxa and higher for more distantly related taxa" was reported for lepidopteran *COI* sequences (Hebert et al., 2003). In a separate study, Hebert et al. (2004a) report an average of 2.76% divergence (Kimura-2-parameter model) for 137 distinctive *COI* sequences from Hesperiidae (Lepidoptera; after removing 479 redundant sequences from consideration). From a sample of 150 mayfly *COI* sequences representing 80 species in 30 genera, Ball et al. (2005) report that congeneric sequences were 3.3–24.8% different (Kimura-2-parameter model). Barrett and Hebert (2005) report mean Kimura-2-parameter pair-wise divergence of 16.4% at *COI* for congeneric species

belonging to 30 spider genera. Vences et al. (2005b) observed 1–16.5% uncorrected p-divergence at 16S (mode = 7–9%) between species of Madagascan frogs. Interspecific comparisons of 1333 Diptera *COI* sequences resulted in uncorrected p-divergence values of 0–17.5% (Meier et al., in press). Ward et al. (2005) observed a mean pair-wise divergence (Kimura-2-parameter) of 23.27% at *COI* when comparing 754 sequences from 207 species of fish. Sixteen species of red algae represented by 101 *COI* sequences were 4.5–13.6% divergent (maximum likelihood general time reversible model) within genera (Saunders, 2005). In angiosperms, 0–7.1% uncorrected p-divergence at nrITS and 0–5.2% uncorrected p-divergence at the *trnH-psbA* spacer was reported between sympatric species pairs (Kress et al., 2005).

The aligned length of the nrITS 2 full reference database is more than 63 times the median unaligned sequence length. A similar pattern is found in the restricted reference database (more than 34 times the median unaligned sequence length). The proportion of parsimony informative characters is also rather high: $\approx$43% for the full reference database and $\approx$46% for the restricted reference database. These two characteristics in combination with the decrease in redundant sequences after alignment indicate that the nrITS 2 alignment is rather ambiguous, inconsistent, and relatively untrustworthy. In comparison with nrITS 2, the *matK* alignment appears to be much more reliable: the aligned length of the full and restricted reference databases was between 2.4 and 2.5 times longer than the median unaligned length. The proportion of informative characters was, however, rather high: 55–59%. The poor quality of the alignment is likely responsible, in part, for the relatively poor performance of hierarchical clustering methods. Neighbor joining may be particularly victim to a poor quality alignment as dnadist (used to calculate the distance matrix for neighbor joining) was forced in many instances to use negative distance because there was "no overlap" between sequences in the nrITS 2 alignment. These effects highlight the difficulty in using length variable loci for DNA barcoding and indicate that identification methods that do not require a multiple sequence alignment should be used in preference to those that do, when length variable loci are used as a data source.

Precision and accuracy

This paper represents the first account to explicitly measure both accuracy and precision in the context of DNA barcoding. Many "DNA barcoding" studies are limited to reports of molecular diversity—within and between species—and draw conclusions about the amount of variation that can found at particular genomic regions for particular taxa (e.g., Hebert et al., 2004b; Kress et al., 2005) or reevaluate species circumscription (e.g., Hebert et al., 2004a; Lambert et al., 2005; Smith et al., 2006).

There are several papers that attempt to demonstrate the functionality and usefulness of DNA barcoding: e.g., Hogg and Hebert (2004) report that they were able to "successfully discriminate" between all of the species of Collembola from the Canadian arctic in their data set. Although Hogg and Hebert report that found consistent sequence variation between species, they did not actually test the ability of their identification method (neighbor joining and multidimensional scaling output are presented in the publication) to correctly identify the query sequences.

Hebert et al. (2003) estimated the accuracy of species-level identification by comparing 150 lepidopteran *COI* query sequences with a reference database of 201 sequences[2] belonging to 147 genera. Neighbor joining was used to identify query sequences. All of the 150 *COI* query sequences were placed sister to a sequence belonging to the correct species in their analysis. It should be noted that the reference database did not include multiple sequences per species and included few genera with more than one species represented (due to the geographically restricted nature of their sample) so that correct identification at the generic level would automatically result in correct identification of species a minimum of 56% of the time (112 of 201). Hebert et al. (2003) do not indicate which of the 201 species in the reference database were used for the test. If we presume that 112 of the 150 query sequences belonged to genera represented by only one species, accuracy was critically tested for only 38 of the queries (more extensive criticism of this study has been put forth: Sperling, 2003; Will and Rubinoff, 2004; Meier et al., in press).

Using a reference database of 80 mayfly sequences representing 80 species in 30 genera, Ball et al. (2005) were able to correctly identify 99% of 70 query sequences (representing 32 species) using the neighbor-joining algorithm.

Barrett and Hebert (2005) measured DNA barcoding accuracy using a reference database of 203 species represented by one sequence each. Although there were apparently 124 sequences that could have been used to measure accuracy, only 75 of them were used as query sequences. Barrett and Hebert (2005) observed 100% identification success using the neighbor-joining algorithm for those 75 test sequences. Insufficient documentation of sequences used (e.g., GenBank accession codes) prevents a detailed reanalysis of this study. Prendini (2005) presents some critiques of DNA barcoding using Barrett and Hebert's study as an example.

Vences et al. (2005b) attempted to identify tadpoles with a reference database of approximately 1000 16S

[2]Hebert et al. (2003) report 200 sequences in the text, but the electronic appendix contains 201 GenBank accession codes.

sequences. In the vast majority of cases (77%), query sequences derived from tadpoles were matched, by BLAST, to sequences derived from adults collected at the same locality as the tadpoles. Vences et al. (2005b) did not present data to demonstrate that those identifications were in fact taxonomically correct. Overall, the authors were apparently displeased with the quality of the identifications: "Certainly, DNA barcoding is unable to provide a fully reliable species identification in amphibians, especially if reference sequences do not cover the entire genetic variability and geographic distribution of a species." In a follow-up study, Vences et al. (2005a) successfully identified 21 of 22 (95%) *COI* sequences by using BLAST with a reference database of 1563 vertebrate *COI* sequences.

Using 6741 *rbcL* and 33 508 nrITS Euphyllophyta (vascular plant) sequences from GenBank Chase et al. (2005) measured accuracy of identification to genus and species using BLAST. For *rbcL*, 96.31% of the queries returned the correct species among the top ranked BLAST scores and 99.77% of the queries returned the correct genus among the top ranked scores. Depending upon the combination of nrITS 1, 5.8S and nrITS 2 used as the query 67.24–99.53% of the queries returned the correct species among the top ranked BLAST scores and 63.96–100% of the queries returned the correct genus among the top ranked scores. The number of queries resulting in an ambiguous assignment (due to equivalent BLAST scores) cannot be discerned from the published data (Chase et al., 2005, their Table 2) and as a consequence the data from Chase et al. (2005) can only be compared with the numbers outside parentheses in Tables 2 and 3 (representing the percentage of correct identifications including ambiguous identifications). In addition, the method used by Chase et al. (2005) to measure accuracy is not the same as the one described here (a restricted reference database), so direct comparisons are even more difficult.

Kress et al. (2005) report that BLAST searches of GenBank using nrITS and *trnH-psbA* spacer "returned correct identities at both the gene and species level". The authors do not report how many sequences were queried, to what extent the incompleteness of the reference database may have effected their results, or how it is possible to obtain correct identifications at the species level if there are no corresponding reference sequences in GenBank.[3]

[3]The sequences generated by Kress et al. (2005) were not incorporated into the BLAST database prior to publication. At the time of publication there were nrITS and *trnH-psbA* spacer sequences, for some of the same species that Kress et al. sampled, available in GenBank (submitted by other researchers), but the vast majority of species were absent from GenBank (57% absent for nrITS and 93% absent for *trnH-psbA* spacer)—therefore BLAST could not have correctly matched the query sequences in most instances. At best Kress et al. tested DNA barcoding for 49 of 194 species.

Meyer and Paulay (2005) tested the accuracy of DNA barcoding with a data set of 2026 *COI* sequences from 263 species of marine Cypraeidae (cowries). A restricted reference database of one sequence per species was constructed and 1000 randomly selected sequences not in the reference database were individually queried using hierarchical clustering (parsimony and neighbor joining). Of the queried sequences, 79–80% were unambiguously correctly identified.

Steinke et al. (2005) used two sets of aquatic animals (snails and fish) and *COI* and/or 16S rDNA sequences to test DNA barcoding. Using TaxI all 28 fish 16S rDNA query sequences were correctly identified to species and 18 of the 20 (90%) snail 16S rDNA query sequences were correctly identified to species (all were correctly identified to genus). Sixteen of the 20 (80%) snail *COI* query sequences were correctly identified to species—frequency of correct species identification was increased to 100% for the snail data sets by changing the method used to calculate distances.

Meier et al. (in press) used hierarchical clustering (neighbor joining, parsimony and Bayesian analysis) as well as the TaxonDNA algorithm to test DNA barcoding accuracy using 1333 Diptera *COI* sequences. Depending on the scoring mechanism used, hierarchical clustering correctly identified 21.8–62.8% of the sequences. For that same data set, a maximum of 68.1% of the sequences were correctly identified to species using the "best match" analysis in TaxonDNA.

Owing to the different methods and data sets used, our results may not be directly comparable with the studies cited above. It appears that our results strongly contrast with the findings of Hebert et al. (2003), Ball et al. (2005), Barrett and Hebert (2005), Chase et al. (2005), Kress et al. (2005), Steinke et al. (2005) and Vences et al. (2005a) while the results reported here are more or less in agreement with the findings of Meyer and Paulay (2005), Vences et al. (2005b) and Meier et al. (in press). One possible explanation for this discrepancy is sample composition: a world-wide sample of species was used for this study, whereas Hebert et al. (2003), Barrett and Hebert (2005), Kress et al. (2005), Steinke et al. (2005) and Vences et al. (2005a) used geographically restricted samples—such that most of the species in their samples are evolutionarily distant to one another, and therefore less likely to be identical or share haplotypes at the barcode locus. Sample size and the use of multiple sequences per species may also be a factor. In particular the identification success of the neighbor-joining algorithm reported by Hebert et al. (2003); Barrett and Hebert (2005); Ball et al. (2005) could not be replicated here—perhaps due to the alignment difficulties (detailed above) that were not present in previous studies. Ambiguous alignment cannot fully explain this discrepancy, because the test data

sets used by Meyer and Paulay (2005) and Meier et al. (in press) were unambiguously aligned and the resulting neighbor-joining identifications were correct at a frequency similar to the results reported here.

Caveats

It is clear that the taxonomy used in GenBank could lead to some difficulties in the estimation of DNA barcoding accuracy. Measurements of precision are unaffected by the taxonomy used, because precision was calculated such that it reflects only the ability of an algorithm to identify the sequence in the reference database that exactly matches the query sequence—thus the name attached to a given sequence is irrelevant. Our analysis of DNA barcoding accuracy—like all similar studies—necessarily assumes consistent species identification. As sequences were contributed to GenBank by a variety of researchers there is no assurance of correct identification. Without any objective mechanism to classify identifications as reliable we were forced to assume that all identifications were correct. Although accuracy as estimated here represents something of a worst case scenario (only one sequence used to represent a species), it is a realistic scenario—at least in the near future. As more sequences become available we expect the quality of query sequence identification to increase.

Recommended identification methods

All methods of query sequence identification fail when confronted with identical reference sequences belonging to more than one terminal—logically there is no way to identify correctly and unambiguously a query sequence in the presence of shared haplotypes. Because phylogenetic relationships are not incorporated into methods relying on similarity or diagnostic characters (e.g., BLAST, BLAT, megaBLAST, DNA–BAR, DOME ID) these methods fail in cases when sequences from one species are more similar to sequences from another species (discussed in reference to multidimensional scaling above). Although cladistic methods can recover hierarchical relationships between sequences and thereby overcome homoplasy and to some extent indistinct species boundaries, these methods fail as a result of alignment inconsistencies. ATIM is able to reduce but not eliminate the failure rate directly attributable to alignment issues while still retaining some of the advantages of conventional cladistic methods. Although ATIM had the best performance of any method tested here for the nrITS 2 data set, that performance was not replicated for the *matK* data set. Even at its best, ATIM was not able to accurately identify query sequences to species at a frequency that would be considered useful in practice.

Because none of the methods perform particularly well, the amount of computer time needed to conduct an analysis becomes an important consideration. Although converting raw sequence data to an appropriate reference database format is nearly instantaneous for BLAST, BLAT and megaBLAST, all of the other methods entail a computationally costly conversion process. A sufficiently large (e.g., 10 000–1 000 000 sequences) reference database, such as one envisioned by the proponents of DNA barcoding, could not easily be analyzed with any of the clustering methods, ATIM, DNA–BAR, or DOME ID. If any of the methods requiring extensive processing of the reference database were to be used, some form of compartmentalization would be necessary: a preliminary analysis using a representative sample would first be conducted and upon the identification of a computationally workable subset of the reference database more thorough analyses of that subset could then be conducted. Given that the similarity methods do not perform much worse—and in some cases perform better—it seems that simply using one of the similarity methods would be the preferred alternative.

Given the relatively precise nature of the algorithms (median = 67% unambiguous) one would expect accuracy to be approximately the same—assuming that there are no additional complications (e.g., shared haplotypes, haplotypes of one species more similar to those of a different species, consistent taxonomic identification of the sequences). As the observed accuracy of species-level identification is much lower, we must assume that it is not due to the inability of the algorithms to match a query sequence to the most similar sequence in the reference database, rather it is a failure of the various matching criteria to correspond with the lack of detectable interbreeding that has traditionally been used explicitly or implicitly to delimit species. This is not to say that the traditional species delimitations are in someway incorrect, but to say that the "closest" sequence suggested by the various identification algorithms is in many instances irrelevant to species as delimited by systematists.

Application of DNA barcoding to conservation

It appears that DNA barcodes could be used to identify Cycadopsida specimens for the purposes of CITES enforcement instead of monosaccharide profiles. The potential advantages of DNA barcodes are that specimens could be identified to species rather than just to genus (in many cases) and a wider range of biological materials could be used to make the determination. In addition, the equipment and expertise for DNA analysis is available in most forensic laboratories, whereas the equipment to observe and interpret monosaccharide

profiles is much less common. With the exception of DNA–BAR and DOME ID all of the methods tested here can identify nrITS 2 and *matK* sequences to genus with high accuracy (greater than 90% unambiguously correct) and some methods (e.g., BLAST, BLAT, ATIM) have an error rate of 2% or less. Using the available nrITS 2 sequences, seven of the genera can be diagnosed as having unique barcode(s) or unique combinations of distinguishers (Table 5). Using *matK* sequences, eight cycad genera can be differentiated. Using a combination of nrITS 2 and *matK*, all but one genus (*Zamia*) can be positively determined.

Conclusions

From our data it appears that DNA barcoding could be used to identify specimens to genus with a minimal amount of error ($\approx$ 0–5%). In situations where geographic ranges can be used as an additional elimination factor (e.g., regional diversity projects) DNA barcoding may be useful at the species level. Without such elimination factors, DNA barcoding does not appear to be particularly useful for species-level identification—none of the methods examined was able to identify query sequences to species at a frequency that would be considered useful in practice. Our data suggest that the most reliable identifications would be made using a reference database in which virtually all haplotypes in all species are represented. Under such circumstances, the most precise algorithms would return a correct identification in most cases (including ambiguous identifications when called for). As a result, the application of DNA barcoding techniques to identification questions should be limited to instances in which the investigators can reasonably assume that the reference database being used contains sufficient depth of sampling and sequence variation to produce reasonably accurate identifications.

Acknowledgments

We thank K. Cameron and an anonymous reviewer for providing constructive criticism on previous incarnations of this manuscript. We thank R. Meier for many useful comments and for kindly providing his manuscript to us prior to publication. C. Chaboo and T. Dikow generously provided copies of bibliographic material.

Table 5
Diagnostic nrITS 2 and *matK* distinguishers for genera of Cycadopsida (distinguishers are diagnostic provided that the specimen is assumed to be a member of the Cycadopsida). The +/− orientation of the barcode sequence is arbitrary

Taxon	CITES appendix	nrITS 2 barcode(s)	*matK* barcode(s)
Bowenia	II	CGTCCGTGTC, GTGCACCCGC, CCCTCGGCCG, GATATGCCAA, CGGGGCATGC, CGGCGTGCCA, CGGGAAACGG	(TTGAATAGGA and AACATATTAT), (TTGAATAGGA and CATATTATCA)
Ceratozamia	I	(CATCCGCGCC and CCAGTGCGAG)	AAAATAAAAG, (ACTCTTTTTT and ACATAAAAGT)
Chigua	I	not positively differentiable from other Cycadopsida	TATTATTTTT, (ATATTTTATC and ATTTCCTTTT), (ATATTTTATC and AAAAGGAAAT), (ATTTCCTTTT and TTCATCCGGA)
Cycas beddomei	I	no data	no data
All other *Cycas*	II	not positively differentiable from other Cycadopsida	AGTTTCCTAA, TCATTTTCAT, GAATAGTTTC
Dioon	II	(GGCTAAAATG and GGACGGCCAA), (CATTTTAGCC and GGACGGCCAA)	not positively differentiable from other Cycadopsida
Encephalartos	I	CGCCTCCCCT	not positively differentiable from other Cycadopsida
Lepidozamia	II	((GTGCTCGGGC and TCTCGCACTG) and not CGCCTCCCCT)	(TCTAGAGGAA and AATGATTTTG)
Macrozamia	II	CGTGTTTCTG, GCGTGTTTCT	CGATCGTTGG, (GATATCCTCGATCGATTC and TTTTTTTTCG), (TTTTTTTTCG and GATAAAAGAT), (TTTTTTTTCG and CGATTCGATCGAATTTGG)
Microcycas	I	not positively differentiable from other Cycadopsida	GATAAAACAT, GAGAATTAAT, (ACTCTTTTTT and TACAGTGGAT), (ACTCTTTTTT, GATAAAACAT)
Stangeria	I	TGGTCGTCCG, GTCGTCCGTG, CGTCTGCGTC	CTTTTTATATCC, (AAAGTATCTGG and ATTCATCCGG)
Zamia	II	not positively differentiable from other Cycadopsida	not positively differentiable from other Cycadopsida

References

Agarwal, P., States, D.J., 1998. Comparative accuracy of methods for protein sequence similarity search. Bioinformatics 14, 40–47.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Amato, G., Egan, M.G., Schaller, G.B., Baker, R.H., Rosenbaum, H.C., Robichaud, W.G., DeSalle, R., 1999. Rediscovery of Roosevelt's barking deer (*Muntiacus rooseveltorum*). J. Mammal 80, 639–643.

Anderson, I., Brass, A., 1998. Searching DNA databases for similarities to DNA sequences: When is a match significant? Bioinformatics 14, 349–356.

Baccam, P., Thompson, R.J., Fedrigo, O., Carpenter, S., Cornette, J.L., 2001. PAQ: partition analysis of quasispecies. Bioinformatics 17, 16–22.

Ball, S.L., Hebert, P.D.N., Burian, S.K., Webb, J.M., 2005. Biological identification of mayflies (Ephemeroptera) using DNA barcodes. J. North Am. Benthol Soc. 24, 508–524.

Barrett, R.D.H., Hebert, P.D.N., 2005. Identifying spiders through DNA barcodes. Can J. Zool. 83, 481–491.

Blaxter, M.L., 2004. The promise of a DNA taxonomy. Phil. Trans. R. Soc. London B Biol. Sci. 359, 669–679.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., Eyualem, A., 2005. Defining operational taxonomic units using DNA barcode data. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1935–1943.

Borneman, J., Chroback, M., Vedovoa, G.D., Figueroa, A., Jiang, T., 2001. Probe selection algorithms with applications in the analysis of microbial communities. Bioinformatics 17, S39–S48.

Brown, S., McLaughlin, W., Jerez, I.T., Brown, J.K., 2002. Identification and distribution of *Bemisia tabaci* (Gennadius) (Homoptera: Aleyrodidae) haplotypes in Jamaica. Trop Agric. 79, 140–149.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.P., Haidar, N., Savolainen, V., 2005. Land plants and DNA barcodes: short-term and long-term goals. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1889–1895.

Cho, Y., Mower, J.P., Qiu, Y.L., Palmer, J.D., 2004. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc. Natl Acad. Sci. USA 101, 17741–17746.

Crisp, M.D., Chandler, G.T., 1996. Paraphyletic species. Telopea 6, 813–844.

DasGupta, B., Konwar, K.M., Mandoiu, I.I., Shvartsman, A.A., 2005a. DNA–BAR: distinguisher selection for DNA barcoding. Bioinformatics 21, 3424–3426.

DasGupta, B., Konwar, K.M., Mandoiu, I.I., Shvartsman, A.A., 2005b. Highly scalable algorithms for robust string barcoding. Int. J. Bioinformatics Res. Appl. 1, 145–161.

Davis, J.I., Nixon, K.C., 1992. Populations, genetic variation, and the delimitation of phylogenetic species. Syst. Biol. 41, 421–435.

De Ley, P., De Ley, I.T., Morris, K., Abebe, E., Mundo-Ocampo, M., Yoder, M., Heras, J., Waumann, D., Rocha-Olivares, A., Burr, A.H.J., Baldwin, J.G., Thomas, W.K., 2005. An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1945–1958.

DeSalle, R., Egan, M.G., Siddall, M., 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1905–1916.

Doukakis, P., Birstein, V.J., Ruban, G.I., DeSalle, R., 1999. Molecular genetic analysis among subspecies of two Eurasian sturgeon species, *Acipenser baerii* and *A. stellatus*. Mol. Ecol. 8, S117–S127.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Farris, J.S., 1974. Formal definitions of paraphyly and polyphyly. Syst. Zool. 23, 548–554.

Fell, J.W., Boekhout, T., Fonseca, A., Scorzetti, G., Statzell-Tallman, A., 2000. Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis. Int. J. Syst. Evol. Microbiol. 50, 1351–1371.

Felsenstein, J., 2004. PHYLIP. Computer program distributed by the author http://evolution.genetics.washington.edu/phylip.html.

Ferguson, J.W.H., 2002. On the use of genetic divergence for identifying species. Biol. J. Linnean Soc. 75, 509–516.

Floyd, R., Abebe, E., Papert, A., Blaxter, M., 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. 11, 839–850.

Fox, G.E., Magrum, L.J., Balch, W.E., Wolf, R.S., Woese, C.R., 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc. Natl Acad. Sci. USA 74, 4537–4541.

Funk, D.J., Omland, K.E., 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Syst. 34, 397–423.

Gibbs, M.J., Armstrong, J.S., Gibbs, A.J., 2005. Individual sequences in large sets of gene sequences may be distinguished efficiently by combinations of shared subsequences. BMC Bioinformatics 6, 90.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2004. TNT. Computer program distributed by the authors http://www.zmuc.dk/public/phylogeny/TNT

Hajibabaei, M., deWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M., Hebert, P.D.N., 2005. Critical factors for assembling a high volume of DNA barcodes. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1959–1967.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W., Hebert, P.D.N., 2006. DNA barcodes distinguish species of tropical Lepidoptera. Proc. Natl Acad. Sci. USA 103, 968–971.

Hebert, P.D.N., Cywinska, A., Ball, S.L., deWaard, J.R., 2003. Biological identifications through DNA barcodes. Proc. R. Soc. London, Biol. Sci. Series B 270, 313–321.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004a. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Natl Acad. Sci. USA 101, 14812–14817.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., Francis, C.M., 2004b. Identification of birds through DNA barcodes. PLoS Biol. 2, e312.

Hofreiter, M., Poinar, H.N., Spaulding, W.G., Bauer, K., Martin, P.S., Possnert, G., Pääbo, S., 2000. A molecular analysis of ground sloth diet through the last glaciation. Mol. Ecol. 9, 1975–1984.

Hogg, I.D., Hebert, P.D.N., 2004. Biological identification of springtails (Hexapoda: Collembola) from the Canadian arctic, using mitochondrial DNA barcodes. Can J. Zool. 82, 749–754.

Jackson, R.B., Moore, L.A., Hoffmann, W.A., Pockman, W.T., Linder, C.R., 1999. Ecosystem rooting depth determined with caves and DNA. Proc. Natl Acad. Sci. USA 96, 11387–11392.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.

Kõljalg, U., Dahlberg, A., Taylor, A.F.S., Larsson, E., Hallenberg, N., Stenlid, J., Larsson, K.H., Fransson, P.M., Kårén, O., Jonsson, L., 2000. Diversity and abundance of resupinate thelephoroid fungi as ectomycorrhizal symbionts in Swedish boreal forests. Mol. Ecol. 9, 1985–1996.

Kõljalg, U., Larsson, K.H., Abarenkov, K., Nilsson, R.H., Alexander, I.J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A.F.S., Tedersoo, L., Vrålstad, T., Ursing, B.M., 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol. 166, 1063–1068.

Kopchinskiy, A., Komon, M., Kubicek, C.P., Druzhinina, I.S., 2005. *Trico*BLAST: a multilocus database for *Trichoderma* and *Hypocrea* identifications. Mycol. Res. News 109, 657–660.

Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., Janzen, D.H., 2005. Use of DNA barcodes to identify flowering plants. Proc. Natl Acad. Sci. USA 102, 8369–8374.

Kristiansen, K.A., Taylor, D.L., Kjøller, R., Rasmussen, H.N., Rosendahl, S., 2001. Identification of mycorrhizal fungi from single pelotons of *Dactylorhiza majalis* (Orchidaceae) using single-strand conformation polymorphism and mitochondrial ribosomal large subunit DNA sequences. Mol. Ecol. 10, 2089–2093.

Lambert, D.M., Baker, A., Huynen, L., Haddrath, O., Hebert, P.D.N., Millar, C.D., 2005. Is a large-scale DNA-based inventory of ancient life possible? J. Hered. 96, 1–6.

Li, J., Wirtz, R.A., McConkey, G.A., Sattabongkot, J., Waters, A.P., Rogers, M.J., McCutchan, T.F., 1995. *Plasmodium*: genus-conserved primers for species identification and quantitation. Exp Parasitol. 81, 182–190.

Lipscomb, D., Platnick, N., Wheeler, Q., 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol. Evol. 18, 65–66.

Lorenz, J.G., Jackson, W.E., Beck, J.C., Hanner, R., 2005. The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1869–1877.

Mallet, J., Willmott, K., 2003. Taxonomy: renaissance or Tower of Babel? Trends Ecol. Evol. 18, 57–59.

Markmann, M., Tautz, D., 2005. Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1917–1924.

Mayr, E., 1957. Species concepts and definitions. In: Mayr, E. (Ed.), The Species Problem. American Association for the Advancement of Science, Washington, DC, pp. 1–23.

Meier, R., Shiyang, K., Vaidya, G., Ng, P.K.L., 2006. DNA barcoding and taxonomy in Diptera: a tail of high intraspecific variability and low identification success. Systematic Biol. 55, 715–728.

Meyer, C.P., Paulay, G., 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3, e422.

Monaghan, M.T., Balke, M., Gregory, T.R., Vogler, A.P., 2005. DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1925–1933.

Moritz, C., Cicero, C., 2004. DNA barcoding: promise and pitfalls. PLoS Biol. 2, e279.

Niesters, H.G.M., Goessens, W.H.F., Meis, J.F.M.G., Quint, W.G.V., 1993. Rapid, polymerase chain reaction-based identification assays for *Candida* species. J. Clin. Microbiol. 31, 904–910.

Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Nixon, K.C., Wheeler, Q.D., 1990. An amplification of the phylogenetic species concept. Cladistics 6, 211–223.

Parkinson, J., Guiliano, D.B., Blaxter, M., 2002. Making sense of EST sequence by CLOBBing them. BMC Bioinformatics 3, 31.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl Acad. Sci. USA 85, 2444–2448.

Poinar, H.N., Hofreiter, M., Spaulding, W.G., Martin, P.S., Stankiewicz, B.A., Bland, H., Evershed, R.P., Possnert, G., Pääbo, S., 1998. Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. Science 281, 402–406.

Prendini, L., 2005. Comment on 'Identifying spiders through DNA barcodes'. Can J. Zool. 83, 498–504.

Rash, S., Gusfield, D., 2002. String barcoding: uncovering optimal virus signatures. In: Proceedings of the Sixth Annual International Conference on Computational Biology, Association for Computing Machinery. ACM Press, New York, pp. 254–261.

Raymond, M., Rousset, F., 1995. An exact test for population differentiation. Evolution 49, 1280–1283.

Rosling, A., Landeweert, R., Lindahl, B.D., Larsson, K.H., Kuyper, T.W., Taylor, A.F.S., Finlay, R.D., 2003. Vertical distribution of ectomycorrhizal fungal taxa in a podzol soil profile. New Phytol. 159, 775–783.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Saunders, G.W., 2005. Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1879–1888.

Savir, D., Laurer, G.J., 1975. The characteristics and decodability of the Universal Product Code symbol. IBM Systems J. 14, 16–34.

Smith, M.A., Fisher, B.L., Hebert, P.D.N., 2005. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1825–1834.

Smith, M.A., Woodley, N.E., Janzen, D.H., Hallwachs, W., Hebert, P.D.N., 2006. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proc. Natl Acad. Sci. USA 103, 3657–3662.

Sperling, F., 2003. DNA barcoding: deus ex machina. Newsl. Biol. Survey Can (Terrestrial Arthropods), 22, 50–53.

Steinke, D., Vences, M., Salzburger, W., Meyer, A., 2005. TaxI: a software tool for DNA barcoding using distance methods. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1975–1980.

Stevenson, D.W., Gigliano, G.S., 1989. The systematic value of the monosaccharide composition and distribution pattern of cycad mucilages. Biochem. Syst. Ecol. 17, 185–190.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., Vogler, A.P., 2002. DNA points the way ahead in taxonomy. Nature 418, 479.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., Vogler, A.P., 2003. A plea for DNA taxonomy. Trends Ecol. Evol 18, 70–74.

Tedersoo, L., Kõljalg, U., Hallenberg, N., Larsson, K.H., 2003. Fine scale distribution of ectomycorrhizal fungi and roots across substrate layers including coarse woody debris in a mixed forest. New Phytol. 159, 153–165.

Thalmann, O., Hebler, J., Poinar, H.N., Pääbo, S., Vigilant, L., 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol. Ecol. 13, 321–335.

Vences, M., Thomas, M., Bonett, R.M., Vieites, D.R., 2005a. Deciphering amphibian diversity through DNA barcoding: chances and challenges. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1859–1868.

Vences, M., Thomas, M., der Meijden, A., Chiari, Y., Vieites, D.R., 2005b. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. Frontiers Zool. 2, 5.

Vrålstad, T., Myhre, E., Schumacher, T., 2002. Molecular diversity and phylogenetic affinities of symbiotic root-associated ascomycetes of the Helotiales in burnt and metal polluted habitats. New Phytol. 155, 131–148.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R., Hebert, P.D.N., 2005. DNA barcoding Australia's fish species. Phil. Trans. R. Soc. London B Biol. Sci. 360, 1847–1857.

Wells, J.D., Sperling, F.A.H., 2001. DNA-based identification of forensically important Chrysomyinae (Diptera: Calliphoridae). Forensic Sci. Int. 120, 110–115.

Wells, J.D., Introna, F., Jr, Di Vella, G., Campobasso, C.P., Hayes, J., Sperling, F.A.H., 2001a. Human and insect mitochondrial DNA analysis from maggots. J. Forensic Sci. 46, 685–687.

Wells, J.D., Pape, T., Sperling, F.A.H., 2001b. DNA-based identification and molecular systematics of forensically important Sarcophagidae (Diptera). J. Forensic Sci. 46, 1098–1102.

Whiteman, N.K., Santiago-Alarcon, D., Johnson, K.P., Parker, P.G., 2004. Differences in straggling rates between two genera of dove lice (Insecta: Phthiraptera) reinforce population genetic and cophylogenetic patterns. Int. J. Parasitol. 34, 1113–1119.

Will, K.W., Rubinoff, D., 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20, 47–55.

Woodwark, K.C., Hubbard, S.J., Oliver, S.G., 2001. Sequence search algorithms for single pass sequence identification: does one size fit all? Comp. Func Genomics 2, 4–9.

Zaidi, R.H., Jaal, Z., Hawkes, N.J., Hemingway, J., Symondson, W.O.C., 1999. Can multiple-copy sequences of prey DNA be detected amongst the gut contents of invertebrate predators? Mol. Ecol. 8, 2081–2087.

Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7, 203–214.

Appendix 1

GenBank accession codes for the internal transcribed spacer 2 of the nuclear ribosomal DNA (nrITS 2) used in this study. Accession codes in **boldface** were used to construct the restricted reference database. The reverse complements of AF024627, AY430069, AY430070, AY430071, AY430073, AY430074, AY430076 and AY430078 were used for analyses.

AF283015, **AF283013**, AF283012, AF283011, AF283010, **AF283014**, **AY083057**, **AY083058**, **AY178417**, **AY178415**, AY178416, **AY845216**, **AY083061**, **AY083062**, **AY083031**, **AF259285**, **U50742**, **AF387538**, **AF387539**, **AF387537**, **AF531227**, **AF531228**, **AY178411**, **AY178404**, **AY178423**, **AY178425**, **AY178399**, **AY178398**, AY178397, **AY178400**, AY178414, AY178410, AY178405, AY178424, **AY178421**, **AY178412**, **AY178429**, **AY178406**, AY178408, AY178407, AY178409, AY178413, **AY178422**, **AY178426**, AY178428, AY178427, **AY380854**, AY150685, AY150684, AY150683, AY150682, AY150681, AY836782, **AY150695**, AY150694, AY150693, AY150691, AF287248, **AY380855**, AY150690, AY150689, AY150686, AF287249, AY150687, **AF036978**, **AF407279**, **AF407304**, **AF407282**, **AF407284**, **AF407285**, **AF407286**, **AF407290**, AF407288, AJ287351, **AF407291**, **AF531240**, AF407305, AJ287352, **AF407306**, **AF407307**, **AF407293**, **AF531241**, AF407294, AF407299, AJ287353, **AF407295**, **AF407301**, **AF407302**, **AF407309**, **AY380856**, AY211258, AY211257, **AY211254**, AY380857, **AY380858**, AY836791, AY283433, AY836792, **AY380859**, AY380860, AY211253, AY211252, AY211251, AY211250, **AY211255**, AY211256, AY380861, **AY380863**, AY380862, AY283429, AY283428, AY283430, AY211249, AY211248, **AF531243**, AJ287354, **AF387522**, **AF387523**, **U77960**, U77957, U77958, U77961, U77962, U77959, U77955, U60747, U77954, U77956, **AY380864**, **U60752**, **AY836793**, **U60748**, **AY380865**, **AY380866**, **AY380867**, AY497210, **AY380868**, **U60753**, **U60750**, **AY380869**, **U60749**, AY497209, **U61265**, **U60751**, **AF394436**, **AF531222**, AF394439, **AF531225**, **AF394438**, **AF394435**, **AF531223**, X91870, AF394434, **AF407283**, AF394441, **AF394437**, **AF394442**, **AF394440**, **AF394443**, **AY083055**, **AY083056**, **AY083054**, **AY083053**, **AY083052**, **AF531229**, AF407289, AJ287355, **AJ287356**, **AJ287357**, **AF531230**, **AF394427**, **AF407303**, AF394424, **AF394383**, **AF531237**, AF394416, **AF394412**, **AF394411**, **AF394388**, **AF394379**, **AF394393**, **AF394394**, **AF394395**, **AF394374**, **AF394396**, **AF394399**, **AF394401**, **AF394385**, AF394384, **AF531236**, AF394375, **AF394407**, AF394405, AF394406, **AY335262**, **AF394404**, **AF394417**, **AF394378**, **AF394387**, **AF394402**, **AF394377**, **AF394376**, **AF394418**, **AF394415**, **AF394430**, **AF394419**, AF394420, **AF394421**, **AF394389**, **AF394400**, **AF394429**, **AF394392**, AF394391, **AF394432**, **AF394386**, **AF394398**, AF394397, **AF394428**, **AF394425**, **AF394390**, **AF394403**, **AF394426**, **AF394413**, **AF394414**, **AF394431**, **AF394423**, AF394422, **AF394408**, **AF394380**, **AF394382**, **AF394410**, AF394409, **AF394433**, **AY755774**, **AY755773**, AY755772, **AY755744**, **AY755757**, **AY755771**, **AY755750**, **AY755777**, **AY755754**, AY755767, **AY755776**, **AY755769**, AY755761, **AY755751**, AY755740, AY755770, AY394065, **AY755775**, **AY755752**, **AY755743**, AY755748, **AY755745**, AY755760, AY755753, AY394067, **AY755758**, AY755741, AY394062, **AY755766**, AY755739, AY394068, **AY755778**, **AY755756**, AY755742, AY755755, AY394069, **AY755746**, AY394066, **AY755764**, **AY755779**, **AY755763**, **AY394064**, **AY755765**, **AY755747**, AY755749, AY394063, **AY755759**, **AY755762**, **AY755768**, **AF531224**, **AY083060**, **AY083059**, **AY836777**, AY178418, **AY380870**, AY211260, AY211259, **Y16892**, Y16380, **AF387525**, **AY449546**, AY449543, AY449544, AY449545, **AY449551**, AY449548, AY449547, AY449549, AY449550, **AY449554**, AY449556, AY449552, AY449557, AY449555, AY449553, **AY449561**, AY449560, AY449558, AY449559, AY449562, **AY449576**, AY449565, AY449569, AY449568, AY449566, AY449564, AY449563, AY449574, AY449572, AY449570, AY449567, AY449575, AY449571, AY449577, AY449573, **AY449583**, AY449579, AY449581, AY449582, AY449580, AY449578, **AY449586**, AY449584, AY449585, **AY445623**, **AY449600**, AY449599, AY449598, AY449597, **AY449601**, AY449602, **AY449603**, AY449604, **AY449608**, AY449609, U50741, **AY449610**, AY449611, **AY449616**, AY449614, AY449613, AY449612, AY449617, AY449615, **AY083044**, **AY083045**, **AY083046**, **AF174627**, AH009237, AH009238, AF231975, AF231978, AF176413, AF231980, AF231976, **AY046525**, AY046526, **AY836794**, **AY283435**, **AY380871**, **AY521466**, AY521465, AY521464, **AY836795**, **AY046516**, AY380872, **AY046521**, AY046522, **AY046518**, AY046517, **AY046520**, AY046519, AY521468, AY521472, AY521470, AY521469, AY521467, AY046524, AY046523, AY521471,

AY380873, **AY836797**, **AY836796**, **AY283434**, **AY083048**, **AF041343**, AY523434, AY523433, AY523432, AY523431, AY523435, **AF538066**, AF538065, AF538064, AF415239, AY523437, AY523441, AY523448, AY523440, AY523439, AY523438, AY523449, AF415237, AY523436, **AF041349**, AY523417, AY603161, AY603160, AY603159, AY523416, AY523415, AY523414, AY523418, AY523413, **AF538068**, **AF041344**, AY523445, AY523447, AY523446, AY523443, AY523444, AY523442, **AF041348**, AY523462, AY523461, AY523459, AY188549, AY188548, AY523460, **AF041346**, **AF538067**, AY523422, AY523421, AY523423, AY523419, AY523420, **AF041347**, AY523457, AY523455, AY523458, AY523456, AY523454, **AF538061**, AY188535, AF538062, AF538060, AY188546, AY188532, AY188531, AY188552, AY188547, AY188545, AY188544, AY188543, AY188542, AY188541, AY188530, AY188540, AY188529, AY188528, AY188527, AY188526, AF415238, AY188539, AY188538, AY188536, AY188537, AY188553, AY188550, AY188533, AY188534, AY188551, **AF041345**, AY603178, AY603176, AY603173, AY523453, AY603181, AY603180, AY603174, AY523450, AY603179, AY603177, AY603175, AY523452, AY523451, **AY603165**, AY523428, AY603171, AY603168, AY523430, AY603164, AY523424, AF538063, AY603170, AY603169, AY603167, AY603166, AY603162, AY523429, AY523425, AY603172, AY523427, AY603163, AY523426, **AY083035**, **AY083036**, **AF531238**, AF394373, **AF531239**, AF407296, **AF531232**, **AF531233**, **AF531235**, **AF531231**, **AF531234**, **AY083049**, **AF387527**, AF387526, AF387528, AF387529, AF387530, **AY380874**, AY836779, **AY083050**, **AF531242**, AF407281, AJ287358, **AY083051**, **AY083063**, **AY178420**, **AY083047**, **AY442162**, AY442161, AY442158, AY442160, AY442159, **AY442167**, AY083033, **AY442157**, AY442156, AY083032, **AY442163**, **AY442164**, AY442165, AY083034, **AJ243167**, **AF136621**, AF136618, AF136620, AF136619, **AF024627**, **AF136616**, AF136614, AF136617, **U24251**, **AF136611**, AF136613, AF136610, AF136612, **AF036983**, AY430073, **AF037000**, **AF036980**, AY430072, **AF037020**, **AF036981**, **AF344000**, **AF036992**, **AF037022**, **AF343992**, **AF344002**, AF036987, AY430076, **AF343983**, AF343994, AF036997, **AF036988**, **AF037014**, U23956, **AF037013**, **AF343988**, AF036995, **AF343981**, **AF037012**, **AF037010**, **AF367378**, AF037016, **AF343993**, **AY430070**, **AF344001**, AY430075, **AF036993**, **AF037007**, **AF037008**, **AF036999**, **AF343987**, **AF037004**, **AY430067**, **AF305061**, **AF036985**, AY430071, **AF343985**, **AF036990**, **AF037017**, **AF037026**, **AF305063**, **AF343995**, AF036994, **AF037006**, **AF343986**, AF343984, **AF037009**, **AF343998**, AF343999, AF037001, **AF343982**, **AF200522**, AF305062, **AF036984**, **AF037019**, **AF036989**, **AF037024**, **AF343996**, AF036998, **X87936**, **AF037011**, **AF036986**, AY430066, **AF343990**, AF343991, **AF343980**, AF343989, **AF037002**, **AF037021**, **AF036996**, **AY430077**, **AF473562**, **AF036982**, AY430069, **AF037003**, **AF200523**, AF367379, **AF200524**, **AF037018**, **AF037025**, **AF037005**, **AF037015**, **AF036991**, AY430074, **AY430078**, **AY380875**, AY836780, **AY083067**, **AY845209**, **AY845215**, **AY083065**, **AY083066**, **AY845208**, **AY083064**, **AY083040**, **AY083043**, **AY083042**, **AY083041**, **AY083039**, **AF041352**, **AF041354**, **AF041353**, **AF041350**, **AF041351**, **AY083037**, **AF387521**, **AF387520**, **AF531226**, AF407300, **AY083038**, **AY916940**, AY917057, AY917052, AY917043, AY917013, AY917012, AY916962, AY916958, AY916904, AY917065, AY917054, AY917038, AY917023, AY917010, AY917006, AY916991, AY916980, AY916978, AY916969, AY916964, AY916957, AY916956, AY916948, AY916946, AY916941, AY916937, AY916935, AY916928, AY916920, AY916891, AY916887, AY916878, AY916872, AY916869, AY916868, AY916853, AY916836, AY916827, AY916826, AY895147, AY917063, AY917047, AY917041, AY917040, AY917037, AY917034, AY917020, AY917019, AY917017, AY917001, AY916979, AY916943, AY916942, AY916938, AY916936, AY916930, AY916926, AY916916, AY916909, AY916908, AY916907, AY916900, AY916899, AY916894, AY916879, AY916877, AY916860, AY916854, AY916845, AY916842, AY917022, AY916952, AY916949, AY916919, AY916912, AY916905, AY916903, AY916902, AY916896, AY916888, AY916886, AY916885, AY916884, AY916875, AY916871, AY916863, AY916861, AY917051, AY916981, AY916975, AY916897, AY916893, AY916890, AY916862, AY917033, AY917030, AY917009, AY916989, AY916987, AY916974, AY916892, AY916880, AY916870, AY916849, AY916831, AY917004, AY916994, AY916986, AY916955, AY916910, AY916851, AY916847, AY916834, AY917062, AY917060, AY917056, AY917049, AY917048, AY917036, AY917026, AY917025, AY917018, AY917015, AY917003, AY916999, AY916997, AY916996, AY916995, AY916993, AY916988, AY916985, AY916984, AY916982, AY916961, AY916953, AY916947, AY916939, AY916933, AY916932, AY916931, AY916929, AY916923, AY916922, AY916914, AY916889, AY916883, AY916876, AY916874, AY916844, AY916843, AY916833, AY916830, AY916829, AY916828, AY916825, AY916823, AY895146, AY917064, AY917055, AY917007, AY916998, AY916992, AY916971, AY916970, AY916968, AY916925, AY916906, AY916837, AY917059, AY916990, AY916873, AY916838, AY917061, AY917058, AY916973, AY916882, AY917046, AY917045, AY917042, AY917035, AY917031, AY917029, AY917027, AY917024, AY917016, AY917002, AY917000, AY916977, AY916966, AY916965, AY916954, AY916915, AY916913, AY916901, AY916895, AY916867, AY916864, AY916859, AY916835, AY916951, AY916866, AY916865, AY917039, AY917014, AY916911, AY916881, AY916855, AY916846, AY916832, AY917053, AY917011, AY916960, AY916934, AY916857, AY916850, AY916840, **AF387524**, AY917084, AY917077, AY917097, AY917096, AY917087, AY917086, AY917073, AY917072, AY917068, AY917093, AY917089, AY917088, AY917080, AY917074, AY917069, AY917081, AY917066, AY917071, AY917070, AY917085, AY917092, AY917094, AY917090, AY917083, AY917082, AY917079, AY917067, AY917091, AY917095, AY917078, **AF387534**, AF387535, AF387536, AF387531, AF387533, **AF259294**, AF259287, **AF259295**, **AF259296**, AF259298, **AF259292**, AF259291, AF259300, **AF259290**, AF259289, **AF259286**, **AF259299**,

AF259293, **AY836781**, AY380876, **AY846281**, AY283432, AY846280, **AY380851**, AY846284, AY283431, **AY380852**, AY846282, AY846283, AY836783, AY846279, **AY846286**, AY836784, **AY846285**, **AY380853**, AY846278, AY836785, **AF259274**, AF259275, **AF259283**, **AF259278**, AF259276, AF259277, **AF259282**, **AF259279**, AF259280, AF259281, **AF259271**, AF259273, AF259272, **AF259284**, **AY570231**, **U50740**, **AY178419**, **AY380877**, AY836787, AY836790, AY836789, AY836788, AY836786, **AJ287324**, **AB106626**, AJ287325, **AJ287326**, **AJ287327**, **AJ287338**, **AJ287328**, **AF531244**, **AJ287364**, **AJ287329**, **AF407292**, AJ287330, **AJ287331**, **AJ287332**, **AJ287339**, **AF531245**, **AJ287333**, **AJ287334**, **AB106625**, AJ287363, AJ287362, **AJ287335**, **AJ287336**, **AJ287337**, **AJ287341**, **AJ287342**, **AB106624**, AJ287343, **AJ287359**, **AJ287344**, **AJ287323**, **AJ287345**, **AJ287346**, **AF531246**, AJ287347, **AJ287360**, **AJ287348**, **AJ287349**, **AJ287340**, **AJ287350**, **AJ287361**.

Appendix 2

GenBank accession codes for the maturase K (*matK*; plastid encoded) sequences used in this study. Accession codes in **boldface** were used to construct the restricted reference database. Complete plastid genome sequences (NC_004677, NC_001631) were reduced to include only *matK* and flanking spacer regions. The reverse complements of NC_001631, NC_004677, and X57098 were used for analyses.

AF456365, AF295026, **AB029657**, **AB029658**, **AF143436**, AB029659, **AB029660**, **AF143441**, AB029661, **AB029662**, **AB029663**, **AB029664**, **AB029665**, **AB029666**, **AB019864**, **AB029667**, **AB029668**, **AB029669**, **AF152175**, **AF457111**, **AB023975**, **AF456371**, **AF456372**, **AF228106**, AF152219, **AB023977**, **AB023981**, **AF456373**, AF543723, **AF456374**, **AF456375**, **AB030131**, **AB030129**, AF152176, **AB030130**, **AF152177**, **AB023979**, AF456378, **AB076221**, AF410173, **AB076223**, **AF152180**, **AB023982**, AF152178, **AF152179**, **AF143435**, AB019867, **AF143431**, **AF295025**, AF456366, **AF228109**, **AB030138**, **AB023986**, **AF457108**, **AF228110**, AB023988, **AF410172**, **AB076195**, **AB076198**, AF279794, **AB076200**, **AY380840**, **AF152181**, **AF152182**, **AB030133**, AY380842, AY380841, AF152183, **AB030132**, **AY380844**, AY380843, **AB076224**, **AB030117**, **AB030116**, AB023984, AF152184, **AB030126**, **AB030125**, AF152185, **AY380845**, AF152188, **AF152186**, **AY380846**, AF152191, **AY380847**, AF152189, **AY497216**, AY380848, AF152190, **AY380849**, **AF152192**, **AY497215**, AY497214, **AF152187**, **AF410160**, **AF410164**, **AF410157**, **AB116585**, **AF410162**, **AB076235**, AB116591, **AB076236**, AF410155, **AF143440**, AF410158, **AB076238**, AB116590, **AB116583**, **AF410161**, **AF410165**, **AB116587**, **AB076231**, AF410159, AF279795, **AF410156**, **AB116584**, **AB076233**, AB116589, **AF410163**, **AF457112**, **AB076193**, AF279796, **AF410166**, **AB076228**, AF279797, **AB076230**, **AF152193**, **AF279798**, **AB076203**, AF279799, **AB076207**, **AF410169**, **AB076210**, **AY492008**, **AY492009**, **AY492010**, **AY492011**, **AY492012**, **AY492013**, **AY492014**, **AY492015**, **AY492016**, **AY492017**, **AY492019**, **AY492020**, **AY492021**, **AY492022**, **AY492023**, **AF279805**, AY492024, **AY492025**, **AY492026**, **AY492027**, **AY492028**, **AF152194**, **AF152195**, **AF279806**, DQ069584, AF456370, AF543736, **AB030118**, AF152196, **AY449631**, **AY449626**, **AY449620**, AY449621, AF542561, **AY449625**, **AY449623**, **AY492029**, **AF280994**, **AY449622**, **AF280995**, **AY449624**, **AY449629**, **AY449628**, **AF457117**, **AF152197**, **AF152198**, **AF152199**, **AB030136**, **AF295027**, AB161020, AB019865, **AF143430**, **AY391401**, **AB019863**, **AF143433**, **AY391403**, **AF295028**, **AF295029**, **AY391402**, **AF457114**, **AB076212**, **AB076213**, AF410167, AF279800, **AF152202**, **AF152200**, **AF152201**, **AB076215**, AF279801, **AF410168**, **AB076218**, **AB076220**, AF279802, **AB030122**, AF152203, **AF152204**, **AB076194**, AF410171, **AF228112**, AB023990, **AF152205**, **AF143437**, AF295030, **AF152206**, **AF280997**, AY442146, **AY442147**, **AY442148**, **AY442149**, **AF456376**, AY442150, **AY289610**, AB161012, AB019862, **AY035202**, **AY035197**, **AY035198**, **AY035196**, **AF133926**, AF133924, AF133923, **AY035201**, **AF133920**, AF133922, AF133919, **AY035194**, **AY035200**, **AY035195**, AF456367, **AF133916**, AF133917, AF133915, AF133918, **AY035204**, **AY035203**, **AY035199**, AF143429, **AY035193**, **AF152207**, **AY497261**, **AY115795**, AB019842, **AF143428**, AB161002, AB019841, **AB080933**, AB063497, **AY497257**, **AY115799**, AB019843, **AF143427**, AB080922, AB063499, **AY497262**, **AB161018**, AB019857, **AY115800**, AB019845, **AB084494**, AB019860, **AB080942**, AB080940, AY497280, AB063517, **AY115773**, **AB161011**, AB160985, AB019832, **AY115784**, AY115785, AY313928, **AY497256**, **AB161003**, AY497265, **AB080921**, AY497266, AB063501, **AB161004**, **AB097785**, AY724751, **AB080938**, AB063502, **AY115776**, **AB097779**, **AB084497**, AB019851, **AY497277**, **AY115780**, AY115777, **AB080925**, AY497274, AB063520, **AY497276**, **AB080936**, AY724746, AY947429, **AY115766**, AY115765, AY313929, **AB080931**, AY724747, **AB080927**, AY497275, **AB161005**, **AY497258**, **AY115801**, AB019844, **AY497282**, **AB081089**, AB019856, **AB161019**, AY497267, **AB161006**, AB019858, **AB080943**, AB063518, **AB161007**, AB019849, **AB080926**, AY497271, **AY115779**, AY115778, **AY115770**, **AB161008**, AY497289, AB019850, **NC_004677**, AB161009, AB019834, **AB019831**, **AB019839**, **AY497260**, **AB097784**, **AB081085**, AY497279, AB063512, **AY115797**, AY115796, AF456368, AY313930, **AB097780**, **AY497278**, **AB080939**, AB063503, **AB081088**, AB019852, **AY115790**, AY313931, **AB161010**, **AY497287**, AB019848, **AY115768**, AY313932, **AY497269**, **AY497259**, AB019836, **AF295031**, AY497263, **AB081087**, AB063504, **AB080935**, AB063519, **AY115793**, AY313933, **AB084498**, AB019854, **AY497281**, **AB081084**, AB063514, **AB080937**, AY724748, AB063505, **AB081086**, AB019837, **AB080944**, AY497284, AB063513,

AY497254, AB019840, **AB084493**, AB019855, **AY115788**, AY115787, AY115786, AY313934, **AB084496**, AB019859, **AB080924**, AY497270, AB063506, **AY497283**, **AY497268**, **AB161013**, AB019833, **AB080932**, AB063507, **AY115771**, AY313935, **AB080934**, AY497286, AB063515, **AY115775**, AY313936, **AB080945**, AY497288, AB063516, **AB080929**, AY724749, AB063508, **AB084495**, AY724752, AB019861, **AY115791**, AY313937, **AY497272**, **AB080930**, AY724753, **AB161014**, AY497264, **AY115802**, AF473563, **AY497255**, AB019835, **AB097781**, AB084492, AB019846, X57098, **AB161015**, AB019847, **AB080928**, AY724750, AB063509, **AB161016**, **AB097783**, AY497285, **NC_001631**, **AY497273**, **AB080920**, AB063510, **AB097782**, **AB097778**, **AB080923**, AY947427, AB063511, **AY734482**, AB019838, **AB161017**, AB019853, **AF152208**, **AF457113**, **AF228111**, **AF457115**, **AF143432**, AB019866, **AF228105**, AF456379, AB023992, **AF143439**, **AF457116**, **AB023994**, AF457107, **AB030123**, AF152209, **AB030124**, AF152210, **AB076201**, AF410174, AF279803, **AB030127**, AB023999, AF152211, **AB030128**, **AB030121**, AF152212, **AB030119**, **AF457109**, AB023996, **AF228103**, **AF228104**, **AB024001**, **AF152213**, **AF152214**, **AF152216**, **AB030135**, AF152215, **AB030134**, AF152217, **AB023998**, **AF228107**, **AF228108**, **AB030137**, AB024003, **AF457110**, **AF143438**, **AF143434**, AF456369, **AF280996**, AY492030, AF542562, **AF152218**, **AF456377**, **AY380850**, **AB076567**, **AB076192**, **AF279804**, **AF410170**, AF542563, **AB076187**, **AB076189**.