MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 1231–1238 doi: 10.1111/1755-0998.12262

Identification of species in the angiosperm family Apiaceae using DNA barcodes

JINXIN LIU,*†[1] LINCHUN SHI,‡[1] JIANPING HAN,‡ GENG LI,*† HENG LU,*† JINGYI HOU,*† XIAOTENG ZHOU,*† FANYUN MENG*† and STEPHEN R. DOWNIE§

*Beijing Area Major Laboratory of Protection and Utilization of Traditional Chinese Medicine, Beijing Normal University, Beijing 100875, China, †State Key Laboratory of Dao-di Herbs, Academy of Chinese Medical Sciences, Beijing 100700, China, ‡Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China, §Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801-3707, USA

Abstract

Apiaceae (Umbelliferae) is a large angiosperm family that includes many medicinally important species. The ability to identify these species and their adulterants is important, yet difficult to do so because of their subtle fruit morphological differences and often lack of diagnostic features in preserved specimens. Moreover, dried roots are often the official medical organs, making visual identification to species almost impossible. DNA barcoding has been proposed as a powerful taxonomic tool for species identification. The Consortium for the Barcode of Life (CBOL) Plant Working Group has recommended the combination of *rbcL+matK* as the core plant barcode. Recently, the China Plant BOL Group proposed that the nuclear ribosomal DNA internal transcribed spacer (ITS), as well as a subset of this marker (ITS2), be incorporated alongside *rbcL+matK* into the core barcode for seed plants, particularly angiosperms. In this study, we assess the effectiveness of these four markers plus *psbA-trnH* as Apiaceae barcodes. A total of 6032 sequences representing 1957 species in 385 diverse genera were sampled, of which 211 sequences from 50 individuals (representing seven species) were newly obtained. Of these five markers, ITS and ITS2 showed superior results in intra- and interspecific divergence and DNA barcoding gap assessments. For the matched data set (173 samples representing 45 species in five genera), the ITS locus had the highest identification efficiency (73.3%), yet ITS2 also performed relatively well with 66.7% identification efficiency. The identification efficiency increased to 82.2% when using an ITS+*psbA-trnH* marker combination (ITS2+*psbA-trnH* was 80%), which was significantly higher than that of *rbcL+matK* (40%). For the full sample data set (3052 ITS sequences, 3732 ITS2 sequences, 1011 *psbA-trnH* sequences, 567 *matK* sequences and 566 *rbcL* sequences), ITS, ITS2, *psbA-trnH*, *matK* and *rbcL* had 70.0%, 64.3%, 49.5%, 38.6% and 32.1% discrimination abilities, respectively. These results confirm that ITS or its subset ITS2 be incorporated into the core barcode for Apiaceae and that the combination of ITS/ITS2+*psbA-trnH* has much potential value as a powerful, standard DNA barcode for Apiaceae identification.

Keywords: Apiaceae, DNA barcoding, identification, internal transcribed spacer, internal transcribed spacer2, *psbA-trnH*

Received 19 June 2013; revision received 19 March 2014; accepted 4 April 2014

Introduction

The flowering plant family Apiaceae (Umbelliferae) comprises approximately 450 genera and 3700 species (Pimenov & Leonov 1993). It is widely distributed in the temperature zones of both northern and southern hemispheres and exhibits a great diversity in Central Asia. The aromatic nature of these plants, both in their foliage and fruits, has led to their common use as foods and spices, and their distinctive chemistry is also reflected in their toxicity and widespread medical use (Heywood 1993). Many species are of ecological importance; several others are grown as ornamentals. Traditionally, the discrimination of Apiaceae species has relied on a diverse array of subtle fruit morphological and anatomical differences, such as the degree and direction of fruit compression, the shape and elaborations of the mericarp ribs, the width of the mericarp commissure, and the number of secretory ducts in the mericarp furrows (She *et al.* 2005). Plants not at the right stage of maturation, however, or preserved or fragmentary specimens

Correspondence: Fanyun Meng and Stephen R. Downie, Fax: 86-10-58807656 and 217-244-7246; E-mails: mfy@bnu.edu.cn and sdownie@life.illinois.edu

[1]These authors contributed equally to this work.

without diagnostic features are usually very difficult if not impossible to identify (Downie *et al.* 2002). To cope with these difficulties, several chloroplast genes (*rbcL*, *matK*), introns (*rpl16*, *rps16*, *rpoC1*) and intergenic spacers (e.g. *psbA-trnH*), as well as the nuclear ribosomal DNA internal transcribed spacer (ITS) region, have been employed to elucidate generic and species-level boundaries (reviewed in Downie *et al.* 2001, 2010; Degtjareva *et al.* 2013). Of all the above loci, the ITS region is evolving most rapidly (Downie *et al.* 2001) and, at present, these sequences comprise the most comprehensive database for Apiaceae phylogenetic study (Downie *et al.* 2010).

There are 100 genera (10 endemic) and 614 species (340 endemic) of Apiaceae native to China (She *et al.* 2005). The indigenous species include many medicinal plants, such as *Angelica sinensis* (Oliv.) Diels, *Ligusticum chuanxiong* S.H. Qiu, Y.Q. Zeng, K.Y. Pan, Y.C. Tang & J.M. Xu, *L. jeholense* (Nakai & Kitag.) Nakai & Kitag. and *L. sinense* Oliv. (National Pharmacopoeia Committee 2010). Most of these medicinal species' official organs are their dried roots or dried ripened fruits. These commonly used medicinal products and their adulterants are frequently found in modern traditional medicinal markets. For example, Gao Ben has been used for over a thousand years to expel wind and cold. According to the National Pharmacopoeia Committee (2010), Gao Ben's botanical origins are *L. sinense* and *L. jeholense* whose dried roots are the official organs. Adulterants from other Apiaceae species have entered the markets, such as *Peucedanum terebinthaceum* (Fisch. ex Trevir.) Turcz. and *Sinodielsia yunnanensis* H. Wolff; all of these species have been misused as Gao Ben because of their morphological similarities to *L. sinense* and *L. jeholense*. The difficulties in identifying both living materials of these plants and the medicinal products derived from their dried roots and other organs result in an instability of the medicinal materials market and jeopardize safety. To overcome these problems, a new accurate identification method for these plants is urgently needed.

DNA barcoding is a practical technique for species identification. It uses short, universal DNA regions to identify species and has been widely applied in species recognition, exploration and conservation (Hebert *et al.* 2003; Lahaye *et al.* 2008b). Although a portion of the mitochondrial gene encoding cytochrome c oxidase subunit 1 (*CO1*) has been used successfully to identify most animal species (Hebert *et al.* 2004; Ward *et al.* 2005; Hajibabaei *et al.* 2006, 2007), this marker is not useful as a barcode in plants because of its low substitution rate (Kress *et al.* 2005; Hollingsworth 2011). The rates of sequence substitutions in plant mitochondrial genomes are 50- to 100-fold lower than those in mammalian mitochondrial genomes (Wolfe *et al.* 1987; Palmer & Herbon 1988; Cho *et al.* 2004). To date, coding (e.g. *matK*, *rbcL*, *rpoB*, *accD* and *rpoC1*) and noncoding intergenic spacer (e.g. *atpF-atpH*, *psbA-trnH*, *psbK-psbI*, *trnL-trnF*) regions from the chloroplast genome, as well as the nuclear ribosomal DNA marker ITS, or a subset of this marker ITS2, have been tested for universal and identification efficiency as alternative barcoding regions (Kress *et al.* 2005; Chase *et al.* 2007; Chiou *et al.* 2007; Lahaye *et al.* 2008a,b; CBOL Plant Working Group 2009; Chen *et al.* 2010; China Plant BOL Group 2011; Shi *et al.* 2011). The Consortium for the Barcode of Life (CBOL) Plant Working Group (2009) officially proposed that chloroplast markers *rbcL* and *matK* serve as the core barcodes for plant species identification. Additional markers or marker combinations still need to be assessed, however, as this two-locus combination only has a 72% identification efficiency at the species level (CBOL Plant Working Group 2009). Researchers have previously used the *psbA-trnH* barcode to identify species of medicinal pteridophytes and members of the orchid genus *Dendrobium* (Yao *et al.* 2009; Ma *et al.* 2010). ITS2 has been selected as a standard barcode to identify medical plants (Chen *et al.* 2010), and both ITS and ITS2 have been incorporated into the core barcode for angiosperms (China Plant BOL Group 2011). The ITS and ITS2 regions have been used to identify species of the families Rutaceae, Asteraceae and Rosaceae (Gao *et al.* 2010; Luo *et al.* 2010; Pang *et al.* 2010), as well as cultivars of the medically important umbellifer species *Angelica anomala* Avé-Lall. (He *et al.* 2012), and all have the potential to be as powerful as the mitochondrial gene *CO1* in discriminating species (Yao *et al.* 2010).

In this study, we used a DNA barcode technique to discriminate species in the angiosperm family Apiaceae. We examined the effectiveness of five loci (ITS, ITS2, *psbA-trnH*, *matK* and *rbcL*) as barcodes by testing them on sequences obtained from over half the species recognized in the family, with an emphasis on those species widely used in traditional Chinese medicine.

Materials and methods

Plant materials

A total of 6032 sequences belonging to 1957 species from 385 diverse genera were used to evaluate the five candidate DNA barcodes (Table S1, Supporting information). These sequences comprised 3052 ITS sequences (representing 1378 species), 3732 ITS2 sequences (some of which were excised from the aforementioned ITS sequences, and representing 1859 species), 1011 *psbA-trnH* sequences (representing 407 species), 567 *matK* sequences (representing 215 species) and 566 *rbcL* sequences (representing 239 species). A total of 211 of these sequences were newly obtained from 50 samples

representing seven species used in traditional Chinese medicine. These medicinal plants were collected from 12 provinces of China (Table S2, Supporting information), and all were authenticated by Prof. Yu-guang Zheng of Hebei Medical University. Voucher samples were deposited in the herbarium at Beijing Normal University. A matched data set (Table S3, Supporting information) comprising 173 samples (representing 45 species from five genera) was constructed to ensure that the levels of sequence variation and species discrimination were compatible (CBOL Plant Working Group 2009; China Plant BOL Group 2011). In this data set, each genus was represented by at least two species, each species had at least two individuals and each sample was successfully sequenced for all five markers.

In an effort to ensure correct species identification, particularly for those ITS sequences downloaded from GenBank, we refer to the study of Downie *et al.* (2010) where they included most Apiaceae ITS sequences available in GenBank. The simultaneous analysis of these sequences in their phylogenetic study permitted misidentifications and other problematic sequences to be revealed. We do realize, however, that we cannot directly confirm Apiaceae identifications in GenBank, as examining all voucher specimens would be a time-consuming task. The sequences we have used were taken mostly from published papers and their sources identified by specialists working on the group. While these plants are notoriously difficult to identify, these individuals are best suited to identify them.

DNA extraction, amplification and data processing

Total genomic DNA was extracted from 30 mg of root tissue that was dried in silica gel. DNA extractions were performed using the Plant Universal Genomic DNA kit (Tiangen Biotech Beijing Co., China). Primer pairs S2F/S3R for ITS2 (Chiou *et al.* 2007), 1F/724R and rbcLa_f/rbcLa_rev for *rbcL* (Fay *et al.* 1997; Kress & Erickson 2007; CBOL Plant Working Group 2009), 3F_KIM/1R_KIM (Ki-Joong Kim, School of Life Sciences and Biotechnology, Korea University, Seoul, Korea, unpublished) and 390F/1326R (Hilu & Liang 1997) for *matK*, ITS4/ITS5 for ITS (White *et al.* 1990), and trnH/psbA for *psbA-trnH* (Sang *et al.* 1997) were used for PCR amplifications of each of the five markers. Processing of the sequences was performed according to the workflow described by Yao *et al.* (2010) and Pang *et al.* (2012). The ITS2 regions were annotated and delimited using a hidden Markov model (HMM)-based method (Keller *et al.* 2009). All ITS sequences were composed of ITS1, 5.8S rRNA and ITS2 regions, except for 135 sequences that lacked data from the intervening 5.8S rRNA gene. We developed and implemented a hairpin-associated programme to find and reinvert inversions in *psbA-trnH* prior to DNA barcode analyses to reduce the effect of homoplasy (Borsch & Quandt 2009; Whitlock *et al.* 2010). All inter- and intraspecific inversions were further confirmed through visual inspection of the alignments and estimating secondary structure (Zuker 2003).

Sequence variation

The DNA sequences were aligned with MUSCLE vers. 3.8 (Edgar 2004) and Kimura 2-parameter (K2P) distances (Kimura 1980) were calculated using PAUP* version 4b10 (Swofford 2003). As described by Chen *et al.* (2010), three parameters (average intraspecific distance, theta and average coalescent depth) were calculated to compare the intrasequence variability of the five markers. Three additional parameters (average interspecific distance, average theta prime and smallest interspecific distance) were computed to evaluate the intersequence divergence of the markers. For each species, we compared the minimum interspecific divergence with the maximum intraspecific divergence to evaluate whether there were any DNA barcoding gaps (Meyer & Paulay 2005). A dot above the 1:1 slope means that there is a barcoding gap for this species for a specific marker, whereas a dot below the 1:1 slope implies no barcoding gap (Collins & Cruickshank 2013).

Species discrimination

To estimate a species discrimination ability, four different methods (Blast, Distance, PWG Distance and Blast+P) were applied to all single markers and all possible 2- to 4-marker combinations. For the Blast analysis, all of the sequences from a specified marker were used as query sequences with an E value of $<1 \times 10^{-5}$ (China Plant BOL Group 2011). The NCBI BLAST program (vers. 2.27) was then used to query the reference database with each sequence, checking whether the best hit was from a conspecific species (the query sequences themselves were excluded). A sample was considered successfully identified if its best hits only included conspecific individuals. For the Distance analysis (China Plant BOL Group 2011), sequences from the same genera were aligned using MUSCLE and their pairwise P distances were calculated using PAUP*. We considered a species to be successfully identified if its minimum interspecific distance was larger than its maximum intraspecific distance. For the PWG Distance analysis, only unambiguous base substitutions were counted based on pairwise alignments (CBOL Plant Working Group 2009). When using the Blast+P method (Pang *et al.* 2012), we initially used Blast to test whether a

sample could be successfully identified by a specified marker. If not, we then extracted sequences of the closest Blast hit and employed a Distance method to see whether this sample could be distinguished from its closest samples from different species. Throughout all subsequent analyses, this method was employed for comparative purposes. To directly compare the relative species discrimination efficiency of the five markers and marker combinations, we first focused on the matched data set. The species discrimination assessments were then expanded to a data set containing all of the samples, including samples where one or more sequences were not successfully sequenced; for this data set, species discrimination abilities of multimarker combinations were not calculated. As done by the CBOL Plant Working Group (2009), species that were represented by only one sequence were not counted in the discrimination success statistics, but were included to serve as potential sources of discrimination failure.

Results

Genetic divergence within and between species

The results of the six parameters used to estimate genetic divergences of the five loci in the matched data set are presented in Table 1. ITS2 had the highest interspecific divergence, followed by ITS. *PsbA-trnH* was at an intermediate level of variation, but much higher than either *rbcL* or *matK*. These results indicate that ITS, ITS2 and *psbA-trnH* exhibit a relatively higher ability for interspecific discrimination than *matK* or *rbcL*. For intraspecific divergence, *rbcL* showed the lowest level of intraspecific variation, while *psbA-trnH* exhibited the highest level, followed by ITS2.

Interspecific inversions in *psbA-trnH* were detected in 23 of 37 genera when more than two species were examined. Intraspecific inversions were less frequent, occurring in only five species. For the 1011 *psbA-trnH* sequences examined, their lengths ranged from 110 to 424 bp, with a mean length of 267 bp and a median length of 335 bp.

Barcoding gap assessment

When intraspecific variations were calculated between conspecific individuals and interspecific divergences were calculated between species, the five markers displayed significant overlap for more or less all species (Fig. 1). For ITS and ITS2, the majority of species had DNA barcoding gaps and showed relatively better performances than any of the other three markers. For *rbcL* and *matK*, most of the species were distributed in the lower left corner of the scatter plot. For *psbA-trnH*, some species had especially higher interspecific variation than intraspecific variation, while other species showed just the opposite.

Identification efficiency of the DNA barcodes

When comparing the identification efficiency of the five markers using two data sets and four analytical methods, Blast+P, in the majority of cases, gave a relatively higher value of species discrimination than did either Distance (based on within-genera multiple alignments) or PWG Distance (based on pairwise alignments); this value, however, was almost always identical to that obtained by Blast. For all markers except *psbA-trnH*, species discrimination of PWG Distance was slightly lower than Blast+P. Hereafter, Blast+P is adopted for the discussion of identification efficiency.

For the matched data set (Fig. 2, Fig. S1, Supporting information), ITS had the highest single-locus species-level identification efficiency (73.3%), followed by ITS2 (66.7%). *PsbA-trnH* performed lower than ITS and ITS2 with a 55.6% success rate, and discrimination success for *rbcL* and *matK* was 24.4% and 33.3%, respectively. The rate of successful species identification with two-locus combinations was highest with ITS+*psbA-trnH* (82.2%), followed closely by ITS2+*psbA-trnH* (80%). The *rbcL*+*matK* combination showed a 40% species discrimination rate. The combination of ITS and any chloroplast DNA locus reached 73.3% to 82.2% species identification success rates (any chloroplast locus and ITS2 reached 66.7–80.0%). The three-locus combinations achieved

Table 1 Comparison of inter- and intraspecific genetic divergences of five loci. Internal transcribed spacer (ITS) sequences included ITS1, 5.8S rRNA and ITS2, except for 135 sequences that lacked data from 5.8s rRNA

	ITS	ITS2	*psbA-trnH*	*rbcL*	*matK*
Average interspecific distance	0.0477 ± 0.0381	0.0619 ± 0.048	0.0291 ± 0.0276	0.0038 ± 0.0035	0.0096 ± 0.0096
Average theta prime	0.0607 ± 0.0632	0.0897 ± 0.1039	0.0335 ± 0.0326	0.005 ± 0.0042	0.0151 ± 0.0204
Smallest interspecific distance	0.015 ± 0.0331	0.0208 ± 0.0521	0.006 ± 0.0155	0.0007 ± 0.0021	0.0029 ± 0.0096
Average intraspecific distance	0.0026 ± 0.0064	0.003 ± 0.0088	0.0066 ± 0.0113	0.0008 ± 0.0019	0.0023 ± 0.0053
Theta	0.0036 ± 0.0087	0.0044 ± 0.0119	0.0069 ± 0.0093	0.0006 ± 0.0012	0.0022 ± 0.0034
Average coalescent depth	0.0051 ± 0.0102	0.0064 ± 0.0148	0.0109 ± 0.0154	0.0012 ± 0.0026	0.0038 ± 0.007

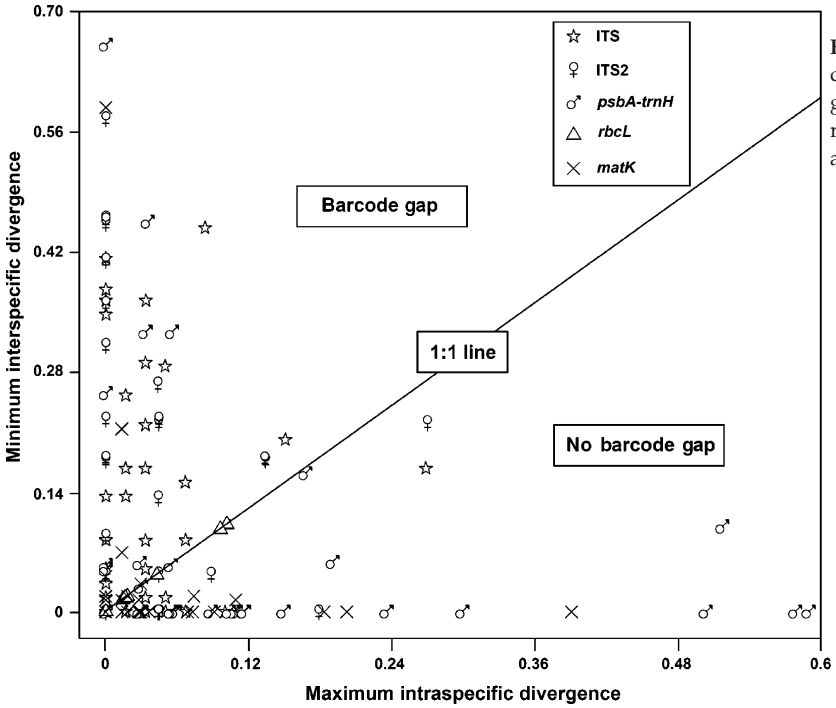

Fig. 1 Scatter plot of the minimum interspecific divergence vs. the maximum intraspecific divergence for the matched data set where each plot represents a species and each symbol represents a specific marker region.

66.7% to 84.4% discrimination rates. The combination of three chloroplast DNA markers (*psbA-trnH+rbcL+matK*) gave a 66.7% species discrimination success rate. The three-locus combinations did not significantly improve the success rates of species discrimination compared to ITS+*psbA-trnH* (at 82.2%). Considering the four barcode combinations (ITS+*psbA-trnH+rbcL+matK* and ITS2+*psbA-trnH+rbcL+matK*), the identification efficiencies were 84.4% and 82.2%, respectively.

When the species discrimination evaluation of the five markers was extended to all of the samples, similar trends in species discrimination were obtained (i.e. ITS > ITS2 > *psbA-trnH* > *matK* > *rbcL*). For ITS, a 70.0% discrimination efficiency was achieved among 3052 samples representing 1378 species from 272 genera. Among the 23 genera with more than 10 sampled species for ITS, four genera (*Chaerophyllum* [45.5%], *Ligusticum* [40%], *Heracleum* [25.6%] and *Peucedanum* [26.7%]) showed <50% discrimination efficiency. ITS2 had a 64.3% species discriminatory value among 3732 samples representing 1859 species from 361 genera. Among the 28 genera with more than 10 sampled species for ITS2, five genera (*Bupleurum* [45.5%], *Chaerophyllum* [36.4%], *Ferula* [40%], *Heracleum* [20.5%] and *Peucedanum* [22.2%]) showed <50% discrimination efficiency. For *psbA-trnH*, a 49.5% discrimination efficiency was achieved among 1011 samples representing 407 species from 80 genera. Among the five genera with more than 10 sampled species for *psbA-trnH*, only *Bupleurum* showed more than 50% discrimination efficiency. *MatK* obtained a 38.6% species discriminatory power from 567 samples representing 215 species in 68 genera. Among the four genera with more than 10 sampled species for *matK*, only *Angelica* showed more than 50% discrimination efficiency. *RbcL* had a 32.1% species identification efficiency among 566 samples representing 239 species and 109 genera. Among the three genera with more than 10 sampled species for *rbcL*, no genus showed more than 50% discrimination efficiency.

Discussion

An ideal DNA barcode should have a larger interspecific than intraspecific divergence, so that it can distinguish one species from another (Kress *et al*. 2005; Kress & Erickson 2007). Furthermore, it should have significantly separated, nonoverlapping genetic variation between its intra- and interspecific samples using a universal primer pair (Hebert *et al*. 2003; Moritz & Cicero 2004). Heretofore, a combination of two or more markers has been suggested for plant DNA barcoding. The China Plant BOL Group (2011) proposed that ITS and a subset of this marker, ITS2, be incorporated alongside *rbcL+matK* into the core barcode for angiosperms. However, in Apiaceae, the combination *rbcL+matK* can only separate 40% of the sampled species in the matched data set for the family. Moreover, *rbcL* and *matK* showed 32.1% and 38.6% species identification efficiencies, respectively, across all samples. The discriminatory power of these two loci decreases substantially when conspecific taxa are

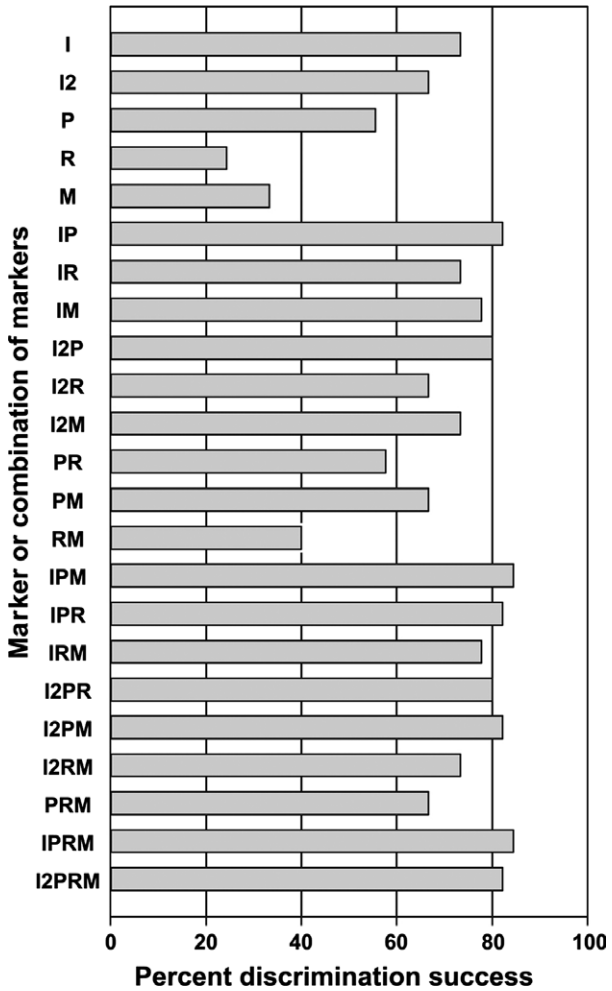

Fig. 2 A comparison of per cent discrimination success for five individual markers and two- to four-marker combinations (I, internal transcribed spacer (ITS); I2, ITS2; P, *psbA-trnH*; R, *rbcL*; M, *matK*).

considered (Chen *et al.* 2010; China Plant BOL Group 2011). These findings are consistent with the conclusion of Downie *et al.* (2001) in suggesting that among all markers commonly employed in Apiaceae phylogenetic study, the ITS region is evolving most rapidly, as evidenced by the greater percentage of sites that are potentially parsimony informative and its higher rate of sequence change.

We report herein that, of the five markers considered, ITS and a subset of this region, ITS2, had the highest discrimination abilities in Apiaceae. This result corroborates studies by Chen *et al.* (2010) and the China Plant BOL Group (2011) done on a wider sampling of taxa where both ITS and ITS2 were determined to be most suitable for DNA barcoding applications. The study of Chen *et al.* (2010) included a large number of medical plants from the Apiaceae (25 genera, 98 species). Similarly, He *et al.* (2012) reported that ITS was the best candidate to authenticate cultivars of *Angelica anomala*, an important medical plant used in traditional Chinese medicine, and to distinguish between *A. anomala* and its closest relatives.

Problems and challenges associated with using the ITS region as a universal plant barcode, such as the potential for fungal contamination, the presence of paralogous gene copies and the lack of universal primers (Hollingsworth 2011), can be mitigated or may not be warranted in the Apiaceae. Fungal contamination can be reduced considerably by proper sample preparation and, should it occur, be identified using a software pipeline that can facilitate the processing and identification of fungal ITS sequences (Nilsson *et al.* 2009). Regarding the issue of paralogous gene copies, the China Plant BOL Group (2011) reported that such paralogs were limited to only 7.4% of the 6286 individuals they investigated and, as such, may not constitute a major problem. Song *et al.* (2012), using pyrosequencing, confirmed that intragenomic multiple copies did not impact phylogenetic reconstruction and species determination in most cases. Similarly, Spalik & Downie (2007) reported that the few intraindividual polymorphisms published to date for Apiaceae do not interfere with the phylogeny estimation. Downie *et al.* (2010) have indicated that ITS sequences are readily obtainable from Apiaceae, even from herbarium specimens a century old. Even if there were taxa for which ITS could not be amplified and sequenced in its entirety, ITS2 is a useful backup because it is shorter and more universal than ITS (Chen *et al.* 2010).

PsbA-trnH is readily amplifiable using universal primers and its levels of species discrimination are usually higher than for other chloroplast regions (Kress *et al.* 2005; Kress & Erickson 2007; Chen *et al.* 2010). Indeed, based on metadata analysis, *psbA-trnH* has been shown to be a valuable marker (Pang *et al.* 2012). In our study, *psbA-trnH* had a species-level identification efficiency of 55.6% in matched data set, much higher than that of *rbcL* or *matK*. However, there are limitations to using *psbA-trnH* as a barcode, and these include its high frequency of length variation and the presence of homopolymers and inversions (CBOL Plant Working Group 2009; Whitlock *et al.* 2010; China Plant BOL Group 2011). In some taxonomic groups, these limitations have severely reduced the utility of *psbA-trnH* as a barcode (Goremykin *et al.* 2005). Its high frequency of length variation resulting from its many insertions and deletions makes it difficult to construct an accurate alignment of these sequences (Kress *et al.* 2005). Moreover, because the region can exceed 1000 bp in size, taxon-specific internal sequencing primers are necessary (Goremykin *et al.* 2005). The high frequency of *psbA-trnH* length variants in some species impacted species identification

using alignment-based methods, such as the Distance analysis, but had minimal effect on the Blast searches, as reported previously by Kress *et al.* (2005) and Kress & Erickson (2007). Moreover, indels can enhance both species identification (Kress *et al.* 2005) and phylogenetic analysis (Degtjareva *et al.* 2012) by providing additional characters. While the presence of homopolymers in *psbA-trnH* can reduce the quality of the sequence, only 6.7% of the Apiaceae *psbA-trnH* sequences examined had long homopolymer tracks, and the confounding effects of these homopolymers can be overcome using cloning approaches. Intraspecific inversions can lead to an overestimate of intraspecific variation, and interspecific inversions can confuse relationships among closely related species (Whitlock *et al.* 2010). A common mechanistic explanation for the origin of inversions is a single mutation in the terminal portion of a hairpin structure, and the most acceptable way to handle these inversions is to reverse complement them in the data matrix (Štorchová & Olson 2007; Borsch & Quandt 2009).

In summary, ITS and ITS2 showed superior results in both intra- and interspecific divergence comparisons, DNA barcoding gap assessments and identification efficiency. The identification efficiency of the ITS/ITS2+*psbA-trnH* combination was significantly higher than that of *rbcL+matK*. These results strongly suggest that ITS or its subset ITS2 be incorporated into the core barcode for Apiaceae and that the combination of ITS/ITS2+*psbA-trnH* markers is a powerful, standard DNA barcode for species identification.

Acknowledgements

We are grateful to Prof. Yu-guang Zheng of Hebei Medical University for confirmation of the Chinese medical plant species and to several anonymous reviewers for constructive comments on earlier versions of the manuscript. This work was supported by the Projects in the National Science and Technology Pillar Program (2012BAI29B02), the National Natural Science Foundation of China (81072999) and the Fundamental Research Funds for the Central Universities, all of which were granted to Fanyun Meng.

References

Borsch T, Quandt D (2009) Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Systematics and Evolution*, **282**, 169–199.

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

Chase MW, Cowan RS, Hollingsworth PM *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Chen SL, Yao H, Han JP *et al.* (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, **5**, e8613.

China Plant BOL Group (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences USA*, **108**, 19641–19646.

Chiou SJ, Yen JH, Fang CL, Chen HL, Lin TY (2007) Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. *Planta Medica*, **73**, 1421–1426.

Cho Y, Mower JP, Qiu YL, Palmer JD (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. *Proceedings of the National Academy of Sciences USA*, **101**, 17741–17746.

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, **13**, 969–975.

Degtjareva GV, Logacheva MD, Samigullin TH, Terentieva EI, Valiejo-Roman CM (2012) Organization of chloroplast *psbA-trnH* intergenic spacer in dicotyledonous angiosperms of the family Umbelliferae. *Biochemistry (Moscow)*, **77**, 1056–1064.

Degtjareva GV, Kljuykov EV, Samigullin TH, Valiejo-Roman CM, Pimenov MG (2013) ITS phylogeny of Middle Asian geophilic Umbelliferae-*Apioideae* genera with comments on their morphology and utility of *psbA-trnH* sequences. *Plant Systematics and Evolution*, **299**, 985–1010.

Downie SR, Plunkett GM, Watson MF *et al.* (2001) Tribes and clades within *Apiaceae* subfamily *Apioideae*: the contribution of molecular data. *Edinburgh Journal of Botany*, **58**, 301–330.

Downie SR, Hartman RL, Sun F, Katz-Downie DS (2002) Polyphyly of the spring-parsleys (*Cymopterus*): molecular and morphological evidence suggests complex relationships among the perennial endemic genera of western North American *Apiaceae*. *Canadian Journal of Botany*, **80**, 1295–1324.

Downie SR, Spalik K, Katz-Downie DS, Reduron JP (2010) Major clades within *Apiaceae* subfamily *Apioideae* as inferred by phylogenetic analysis of nrDNA ITS sequences. *Plant Diversity and Evolution*, **128**, 111–136.

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

Fay MF, Cameron KM, Prance GT, Lledó MD, Chase MW (1997) Familial relationships of *Rhabdodendron* (*Rhabdodendraceae*): plastid *rbcL* sequences indicate a caryophyllid placement. *Kew Bulletin*, **52**, 923–932.

Gao T, Yao H, Song JY, Zhu YJ, Liu C, Chen SL (2010) Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large *Asteraceae* family. *BMC Evolutionary Biology*, **10**, 324.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Molecular Biology and Evolution*, **22**, 1813–1822.

Hajibabaei MD, Janzen H, Burns JM, Hallwachs W, Hebert PD (2006) DNA barcodes distinguish species of tropical *Lepidoptera*. *Proceedings of the National Academy of Sciences USA*, **103**, 968–971.

Hajibabaei M, Singer GA, Clare EL, Hebert PD (2007) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, **5**, 24.

He Y, Hou P, Fan G *et al.* (2012) Authentication of *Angelica anomala* Avé-Lall cultivars through DNA barcodes. *Mitochondrial DNA*, **23**, 100–105.

Hebert PDN, Cywinska A, Ball SL (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**, 313–321.

Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.

Heywood VH (1993) *Flowering Plants of the World*. Oxford University Press, New York.

Hilu KW, Liang H (1997) The *matK* gene: sequence variation and application in plant systematics. *American Journal of Botany*, **84**, 830–839.

Hollingsworth PM (2011) Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences USA*, **108**, 19451–19452.

Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M (2009) 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, **430**, 50–57.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS ONE*, **2**, e508.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences USA*, **102**, 8369–8374.

Lahaye R, Savolainen V, Duthoit S, Maurin O, Van der Bank M (2008a) A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa). *Nature Precedings*. Available from http://hdl.handle.net/10101/npre.2008.1896.1

Lahaye R, Van der Bank M, Bogarin D *et al.* (2008b) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences USA*, **105**, 2923–2928.

Luo K, Chen SL, Chen KL *et al.* (2010) Assessment of candidate plant DNA barcodes using the *Rutaceae* family. *Science China Life Sciences*, **53**, 701–708.

Ma XY, Xie CX, Liu C *et al.* (2010) Species identification of medicinal pteridophytes by a DNA barcode marker, the chloroplast *psbA-trnH* intergenic region. *Biological and Pharmaceutical Bulletin*, **33**, 1919–1924.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.

Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*, **2**, 1529–1531.

National Pharmacopoeia Committee (2010) *Pharmacopoeia of the People's Republic of China*. China Medical Science Press, Beijing.

Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N (2009) A software pipeline for processing and identification of fungal ITS sequences. *Source Code for Biology and Medicine*, **4**, 1.

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, **28**, 87–97.

Pang XH, Song JY, Zhu YJ, Xu HX, Huang LF, Chen SL (2010) Applying plant DNA barcodes for *Rosaceae* species identification. *Cladistics*, **27**, 165–170.

Pang XH, Liu C, Shi LC *et al.* (2012) Utility of the *trnH–psbA* intergenic spacer region and its combinations as plant DNA barcodes: a meta-analysis. *PLoS ONE*, **7**, e48833.

Pimenov MG, Leonov M (1993) *The Genera of the Umbelliferae: A Nomenclator*. Royal Botanic Gardens, Kew.

Sang T, Crawford D, Stuessy T (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (*Paeoniaceae*). *American Journal of Botany*, **84**, 1120–1136.

She ML, Pu FT, Pan ZH *et al.* (2005) *Apiaceae*. Flora of China. St. Louis, Missouri, 14, 1–205.

Shi LC, Zhang J, Han JP *et al.* (2011) Testing the potential of proposed DNA barcodes for species identification of Zingiberaceae. *Journal of Systematics and Evolution*, **49**, 261–266.

Song JY, Shi LC, Li DZ *et al.* (2012) Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. *PLoS ONE*, **7**, e43971.

Spalik K, Downie SR (2007) Intercontinental disjunctions in *Cryptotaenia* (*Apiaceae*, *Oenantheae*): an appraisal using molecular data. *Journal of Biogeography*, **34**, 2039–2054.

Štorchová H, Olson M (2007) The architecture of the chloroplast *psbA-trnH* non-coding region in angiosperms. *Plant Systematics and Evolution*, **268**, 235–256.

Swofford DL (2003) PAUP* 4b10. Phylogenetic analysis using parsimony (* and other methods). Version 4.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1847–1857.

White TJ, Bruns T, Lee S, Taylor JW (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols: a guide to methods and applications, 18, 315–322.

Whitlock BA, Hale AM, Groff PA (2010) Intraspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. *PLoS ONE*, **5**, e11533.

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences USA*, **84**, 9054–9058.

Yao H, Song JY, Ma XY *et al.* (2009) Identification of *Dendrobium* species by a candidate DNA barcode sequence: the chloroplast *psbA-trnH* intergenic region. *Planta Medica*, **75**, 667–669.

Yao H, Song JY, Liu C *et al.* (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE*, **5**, e13102.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, 3406–3415.

F.M. and J.L. designed the study; J.H., G.L. and H.L. collected plant materials; J.H., G.L. and H.L. performed laboratory works; J.L. and L.S. performed the analyses; J.L., L.S. and S.R.D. wrote the paper; F.M. and S.R.D. edited the paper.

Data Accessibility

DNA sequences: Supporting Information and Dryad doi: 10.5061/dryad.pm0hn

Supporting Information

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

Fig. S1 Per cent discrimination success for five markers and all possible 2- to 4-marker combinations based on Blast, PWG Distance and Distance methods (I, ITS; I2, ITS2; P, *psbA-trnH*; R, *rbcL*; M, *matK*).

Table S1 Apiaceae sequences used to evaluate the five DNA barcodes.

Table S2 Sources of medicinal materials and their GenBank accession number.

Table S3 A matched data set consisting of five markers from the same samples.