MISS SONG-ZHI XU (Orcid ID : 0000-0002-4178-5559)

Article type : Resource Article

DNA barcoding of invasive plants in China: a resource for identifying invasive plants

Song-Zhi Xu[1], Zhen-Yu Li[1]*, Xiao-Hua Jin[1]*

[1]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, P. R. China* (xusongzhi2013@ibcas.ac.cn, lizy@ibcas.ac.cn, xiaohuajin@ibcas.ac.cn)

Keywords: DNA barcoding, invasive plant species, molecular marker, ITS/ITS2+*matK*

*Authors for correspondence: Zhen-Yu Li, lizy@ibcas.ac.cn; Xiao-Hua Jin, xiaohuajin@ibcas.ac.cn.

Running title: DNA barcoding of invasive plants

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12715

Resource article: Permanent Genetic Resources

Abstract

Invasive plants have aroused attention globally for causing ecological damage and having a negative impact on the economy and human health. However, it can be extremely challenging to rapidly and accurately identify invasive plants based on morphology because they are an assemblage of many different families and many plant materials lack sufficient diagnostic characteristics during border inspections. It is therefore urgent to evaluate candidate loci and build a reliable genetic library to prevent invasive plants from entering China. In this study, five common single markers (ITS, ITS2, *matK*, *rbcL* and *trnH-psbA*) were evaluated using 634 species (including 469 invasive plant species in China, 10 new records to China, 16 potentially invasive plant species around the world but not introduced into China yet and 139 plant species native to China) based on three different methods. Our results indicated that ITS2 displayed largest intra- and interspecific divergence (1.72% and 91.46%). Based on NJ tree method, ITS2, ITS, *matK*, *rbcL* and *trnH-psbA* provided 76.84%, 76.5%, 63.21%, 52.86% and 50.68% discrimination rates, respectively. The combination of ITS+*matK* performed best and provided 91.03% discriminatory power, followed by ITS2+*matK* (85.78%). For identifying unknown individuals, ITS+*matK* had 100% correct identification rate based on our database, followed by ITS/ITS2 (both 93.33%) and ITS2+*matK* (91.67%). Thus, we propose ITS/ITS2+*matK* as the most suitable barcode for invasive plants in China.

This study also demonstrated that DNA barcoding is an efficient tool for identifying invasive species.

Keywords: DNA barcoding, invasive plant species, molecular marker, ITS/ITS2+*matK*

Introduction

Biological invasion is one of the most severe threats to biodiversity, and it also causes significant economic damage (Mack *et al.* 2000; Mooney & Hobbs 2000; Pimentel *et al.* 2001; Pimentel *et al.* 2005). Invasive plants are the main components of invasive species, and they have had a tremendous negative impacts on the native biodiversity, economy and human health in locations across the world (Li *et al.* 2015). For example, *Eichhornia crassipes* (water hyacinth) in Dian-Chi in China has led to the local extinction of at least 10 native species of aquatic plants (Ding *et al.* 1995). Additionally, noxious agricultural weeds such as *Alternanthera philoxeroides* (alligator weed) have caused an average annual crop output decline in many parts of the world (Qiang & Cao 2000), and the pollen of *Ambrosia artemisiifolia* (common ragweed) and *A. trifida* (giant ragweed) affect human health by causing allergic rhinitis (Xia 1983). China has a rich biodiversity, with diverse climates and environmental conditions that make it extremely vulnerable to exotic species (Xie *et al.* 2001). Researches show that the number of invasive plant species in China has dramatically increased during the recent decades (Huang *et al.* 2009; Liu *et al.* 2005; Ma *et al.* 2013; Xie *et*

al. 2001), partly due to the explosive growth in trade and transportation. According to the most recent checklist, 515 invasive plant species have been documented in China and have been divided into five classes (1−5) based on their economic and ecological impacts (Ma *et al.* 2013).

Recent studies suggest that the most economical and efficient way to manage invasive species is to prevent them from entering a country in the first place (Ghahramanzadeh *et al.* 2013; Park & Potter 2013). Therefore, the pivotal issue is determining how to rapidly and accurately recognise invasive taxa during border inspection, but this remains a difficult task. Firstly, invasive plants are composed of different families, including several large families of angiosperms such as Asteraceae, Fabaceae and Poaceae, the taxonomy of which is complicated (Fu *et al.* 2016; Käss & Wink 1996; Panero & Zelnik 2014; Saarela *et al.* 2015; Soreng *et al.* 2015; Wojciechowski *et al.* 2004). Secondly, plant materials such as seeds, seedlings, vegetative tissue and fruits, often lack diagnostic morphological features that can be readily assessed during border inspections. Thirdly, the number of classical morphology-based taxonomists are becoming insufficient (Li *et al.* 2011b; Pyšek *et al.* 2013), and most border officers do not major in systematic botany and therefore cannot provide rapid and correct identification of alien specimens (Hoveka *et al.* 2016).

DNA barcoding is an effective tool for identifying species using a standard DNA region (Hebert *et al.* 2003b). Numerous studies have indicated that the mitochondrial cytochrome oxidase 1 (CO1) gene sequence can be used to successfully identify a large variety of animal

species (Blagoev *et al.* 2016; Breman *et al.* 2016; Hajibabaei *et al.* 2006; Hebert *et al.* 2003a; Mendoza *et al.* 2016; Ward *et al.* 2005). However, due to the lower substitution rate of CO1 in plants (Hollingsworth *et al.* 2011; Kress *et al.* 2005) and the complex evolutionary processes such as hybridisation and polyploidy in plants (Fazekas *et al.* 2009; Rieseberg *et al.* 2006), no universal barcode sequence is suitable for all plants. Instead, researchers have proposed several candidate barcodes or combinations including *trnH-psbA* (Kress *et al.* 2005), *matK* (Lahaye *et al.* 2008) and ITS (Li *et al.* 2011a), *matK+rbcL* (CBOL Plant Working Group 2009), *rbcL+trnH-psbA* (Kress & Erickson 2007). In addition, many studies also proposed the best performing barcodes for each family or genus (Ashfaq *et al.* 2013; Chen *et al.* 2015; Li *et al.* 2012; Liu *et al.* 2014; Xiang *et al.* 2011; Xu *et al.* 2015; Yan *et al.* 2015; Yang *et al.* 2012; Zhang *et al.* 2012), or for endangered plant species (Parveen *et al.* 2012; Sosa *et al.* 2013) or for medicinal plant species (Asahina *et al.* 2010; Chen *et al.* 2010). However, few studies have focused on invasive plants (Ghahramanzadeh *et al.* 2013; Hoveka *et al.* 2016; Thum *et al.* 2012; Van De Wiel *et al.* 2009), and the ability of DNA barcoding to identify all invasive plants is yet to be fully explored.

In this study, we sampled ~91% of invasive plant species present in China with the aims of (i) proposing a core barcode by evaluating five commonly recommended DNA regions (ITS, ITS2, *matK*, *rbcL* and *trnH-psbA*) alone and in combination using three different methods; (ii) establishing a reliable barcode dataset for identification of invasive plants in China and exotic plants that have not been introduced into China yet; (iii) testing the performance of the database with 15 unknown plant species collected from border inspection sites.

Materials and methods

Samples

A total of 7861 sequences representing 634 species (469 invasive species from 68 families in China, 10 new records to China, 16 potentially invasive plant species around the world but not introduced into China yet and 139 native species that congeneric with some invasive species in China), were obtained to evaluate the five candidate barcodes (Table S1). These sequences comprised 2491 ITS sequences, 2491 ITS2 sequences (excised from aforementioned ITS sequences), 2022 *matK* sequences, 2198 *rbcL* sequences, and 1150 *trnH-psbA* sequences. More than 80% of species were represented by at least two individuals, except for *trnH-psbA* (full details are provided in Table 1). To test the efficiency of barcode databases, 15 unknown samples were collected from border inspection sites (numbered X1–X15).

In total, 7237 sequences were downloaded from NCBI. To avoid misidentified sequences from NCBI, most downloaded sequences were from published papers or provided voucher specimens. Meanwhile, 624 sequences were newly generated from 164 individuals representing 99 species and one variety. Voucher specimens were deposited at the Herbarium of the Institute of Botany, Chinese Academy of Science, Beijing, China (PE), and collection information are summarised in Table S2. Specimens were identified by Prof. Zhen-Yu Li from the Institute of Botany, Chinese Academy of Science. To ensure correct species identification, we authenticated specimens using the following steps: (i) reviewing relative monographs and floras; (ii) comparing with online image libraries of living plants; and (iii)

comparing with type and native specimens. All newly acquired sequences have been submitted to GenBank (Table S2).

DNA extraction, PCR and sequencing

Genomic DNA was extracted from silica gel-dried plant leaves using a Plant Universal Genomic DNA kit (Tiangen Biotech Beijing Co., China). Target DNA regions (nuclear ITS and three chloroplast regions consisting of coding genes *matK* and *rbcL*, and the intergenic spacer *trnH-psbA*) were amplified in 25 µL reactions containing 10 ng (1–2 µL) template DNA, 12.5 µL 2× PCR mix (0.005 units/µL Taq DNA polymerase, 4 mM $MgCl_2$, 0.4 mM dNTPs), 0.2 µL of each primer, and 6.5–7.5 µL ddH_2O. Primer pairs for PCR and sequencing used in this study are listed in Table S3. For *matK*, a second pair of primers was necessary when the primer combination (390F and 1326R) failed during PCR amplification or sequencing. PCR products were bidirectionally sequenced by Shanghai Majorbio Bio-Pharm Technology Co., Ltd.

Data analysis

Sequencing output files were trimmed and assembled using the program ContigExpress Application 6.0 (InforMax, Inc.). Edited sequences were aligned using MAFFT (Katoh *et al.* 2002) and then manually adjusted in BioEdit v.7 (Hall 1999). We evaluated twelve barcodes

including five single loci and seven combinations (ITS/ITS2+*matK*, ITS/ITS2+*rbcL*, *matK*+*rbcL* and ITS/ITS2+*matK*+*rbcL*) using the methods described below.

(1) Genetic distance-based method: We calculated intra- and interspecific divergence using Kimura 2-parameter (K2P) distances in MEGA 6.0 (Tamura *et al.* 2013). Two different approaches were used to detect barcoding gaps, namely, histograms and scatter plots. Histograms were generated from the distribution of divergence at intervals of 0.005 distance units using the program TaxonDNA (Meier *et al.* 2006). Scatter plots were compiled using R version 3.2.5 (R Development Core Team 2014) and each dot represents a species. The genetic pairwise distance for each marker was calculated in MEGA 6.0 (Tamura *et al.* 2013) using the K2P model, and we searched the minimum interspecific distance and maximum intraspecific distance for each species using custom R script. The dot above the 1:1 slope indicated a barcoding gap (Collins & Cruickshank 2013; Liu *et al.* 2014). We counted the number of species having barcoding gaps for each marker, and differences between any two markers were analysed by chi-square tests using IBM SPSS version 19.0 (IBM Corp. 2010).

(2) Similarity-based method: To assess the accuracy of the twelve barcodes for species assignment, best match (BM) and best close match (BCM) functions in the program TaxonDNA (Meier *et al.* 2006) were implemented. For BM analysis, each query species is assigned to its best matching sequence(s). Identification was considered correct when query and BM sequences were from the same species, ambiguous when they were from both the same and different species, or incorrect when they belonged to different species. BCM-based species assignment was considered correct if a query matched all conspecific sequences within the 95% pairwise genetic threshold (Meier *et al.* 2006). Since species represented by a

single sequence are always identified incorrectly, they were omitted from BM and BCM tests.

(3) Tree-based method: To evaluate discriminatory power in a more straightforward manner, unrooted neighbour-joining (NJ) trees were constructed in MEGA 6.0 (Tamura *et al.* 2013), with pairwise deletion based on the p-distance model (Collins & Cruickshank 2013; Ma *et al.* 2013; Srivathsan & Meier 2012; Yan *et al.* 2015). Node support was calculated based on 1000 bootstrap replicates. A species was considered successfully identified only when all conspecific individuals formed a single clade with a bootstrap value ≥ 50%. To test the practical efficiency of our database, we added 15 unknown species into the corresponding database to build an NJ tree. The unknown species were considered as those which it formed a monophyly with. We then compared the results with specimens that had been morphologically identified by Prof. Zhen-Yu Li.

Results

Barcode universality and sequence characteristics

The PCR amplification and sequencing success rates were very high for *rbcL* (100%) and *trnH-psbA* (98.17%) (Table 1). For ITS, 90.24% of the sequences were successfully amplified and sequenced. Sequencing failed for 17 individuals because of polymorphic sites (double peaks) or a poly-G structure in the trace file. For *matK*, two primer pairs have been used. Longer sequences were obtained using 390F/19F primer pair. However, the success rate was low (less than 50%). Failed sequences were resequenced using the AF/8R primer pair,

and only nine individuals, mainly from Fabaceae, failed to generate readable sequences. In total, 92.07% of sequences were successfully sequenced. Sequence alignment was most reliable for *rbcL* and *matK*, followed by ITS. By contrast, *trnH-psbA* was extremely difficult to align unambiguously resulting from a high level of length variation (ranging from 120 bp to 771 bp).

The results of genetic distance indicated that ITS2 had the greatest interspecific variation (91.46%), followed by *trnH-psbA* (61.12%) and ITS (52.02%), whereas *rbcL* was the most conserved and displayed the lowest intra- and interspecific divergence (0.23% and 10.49%, respectively), and *matK* consistently exhibited medium values for intra- and interspecific distance (0.35% and 28.96%) (Table 1).

Performance of barcodes based on different analytical methods

Three different analytical methods were used to evaluate the discriminatory power of the twelve barcodes and slightly different results were obtained. The genetic distance method based on histograms did not detect distinct barcoding gaps for any marker and showed total overlap between intra- and interspecific distance distributions (Fig. S1). By contrast, results based on scatter plots did identify barcoding gaps that varied between barcodes (Fig. 1). For ITS+*matK*, 90.08% of species had distinct barcoding gaps, and this barcode was the best performing, followed by ITS2+*matK* (88.74%), ITS+*matK*+*rbcL* (88.44%) and ITS2+*matK*+*rbcL* (88.15%). Chi-squared test results indicated that these four combinations performed significantly better than five single markers ($p<0.01$, Table S4). The results of the

similarity-based method performed by TaxonDNA exhibited the highest power for species identification among three analytical methods (Table 3). Based on the BM model, ITS/ITS2+*matK* (both 96.48%) had highest abilities to identify species, followed by ITS/ITS2+*matK*+*rbcL* (both 96.32%). Base on the BCM model, the best-performing barcode was ITS2+*matK*+*rbcL* (95.46%), followed by ITS2+*matK* (94.95%). The NJ tree analysis demonstrated that ITS+*matK* had the highest resolution rates (91.03%; Fig. 2) among all single and combined barcodes, followed by ITS2+*matK* (85.78%) and ITS+*matK*+*rbcL* (85.45%) (Fig. S7–S13). Five single region barcodes exhibited relatively lower identification power, ranging from 50.68% to 76.84% (Table 2; Fig. S2–S6). Upon testing the efficiency of each database, the results of NJ tree analysis indicated that ITS+*matK* performed best among the twelve barcodes and could identify 100% (12/12) individuals correctly (Fig. S19). ITS, ITS2, ITS2+*matK* and ITS+*matK*+*rbcL* provided slightly lower discrimination rates (93.33%, 93.33%, 91.67% and 91.67%, respectively) (Table 2; Fig. S14–S25).

Discussion

Evaluation of DNA barcoding for invasive plants in China

In general, an ideal DNA barcode should exhibit high interspecific but low intraspecific divergence, showing distinct DNA barcoding gaps (Kress & Erickson 2007; Kress *et al.* 2005). Most prior studies evaluated barcoding gaps using histograms. However, the absence of barcoding gaps has been reported in many studies (Ashfaq *et al.* 2013; Li *et al.* 2012; Liu *et al.* 2014; Yan *et al.* 2015; Yang *et al.* 2012). In this study, two different approaches were used

to detect barcoding gaps. Similarly, histogram analysis also failed to detect barcoding gaps for any of the markers (Fig. S1). By contrast, when scatter plots were used (Liu *et al.* 2014; Mendoza *et al.* 2016), all barcodes exhibited barcoding gaps to varying degrees and ITS+*matK* performed best (Fig. 1). This obvious difference between the two approaches may come from the variation between intra- and interspecific differences in different taxa (Barrett & Hebert 2005) and intraspecific differences are larger than interspecific differences in some cases (Chen *et al.* 2015). Based on these results, we recommended using scatter plots instead of histograms to evaluate the presence of barcoding gaps.

At the single barcode level, ITS/ITS2 provided highest discrimination abilities based on three analytical methods (Table 2, Table 3), which indicated ITS/ITS2 should be incorporated into the core barcode. In prior studies, *trnH-psbA* has been proposed as a core barcode (Kress & Erickson 2007; Kress *et al.* 2005). However, the potential drawbacks of *trnH-psbA* also have been reported in many researches including high frequency of length variation and the presence of inversions and insertions (CBOL Plant Working Group 2009; Fazekas *et al.* 2008; Whitlock *et al.* 2010). In this study, it was omitted from combination with other single markers for two reasons: 1) Although *trnH-psbA* showed higher intra- and interspecific divergences (0.85% and 61.12%) than *matK* and *rbcL*, our results showed *trnH-psbA* provided similar discrimination rates with *matK* based on three different methods; 2) One of the aims of this study was to build a reliable and complete library, but it was only represented by relatively few sequences (1150) and species (365).

Several combinations have been previously proposed as core barcodes for all land plants (CBOL Plant Working Group 2009; Chase *et al.* 2007; Kress & Erickson 2007). In the present study, we combined four loci (ITS/ITS2, *matK* and *rbcL*) into seven multi-region barcodes, of which *matK*+*rbcL* was the most authoritative barcode for land plants proposed by the Consortium for the Barcode of Life (CBOL) (CBOL Plant Working Group 2009). However, this combination displayed a relatively low discrimination rate in this study. Due to a lack of sufficient variable sites, the combination of *matK*+*rbcL* was incapable of identifying related species has been reported in previous studies (Chen *et al.* 2015; Liu *et al.* 2014; Yang *et al.* 2012; Zhang *et al.* 2012). Therefore, one rapidly evolving locus appears to be essential as part of plant barcodes (Chen *et al.* 2015; Kress & Erickson 2007), and ITS/ITS2 serves this function (Li *et al.* 2011a). The ITS+*matK* combination achieved 91.03% discrimination efficiency based on NJ tree analysis, followed by ITS2+*matK* (85.78%). Most species that were not correctly identified belong to *Amaranthus* (Amaranthaceae), Poaceae or Asteraceae (Fig. 2, Fig. S7 and Fig. S8). This may result from hybridisation and polyploidization, which can cause morphological intermediates and lead to incorrect identification (Yan *et al.* 2015). Hybridisation is very common in *Amaranthus* (Alves *et al.* 2014; León-Romero *et al.* 2012) and chromosome geographical races have been reported in Poaceae, such as *Paspalum dilatatum* (2n = 40, 50, 60) (Burson 1985; Dandin & Chennaveeraiah 1983) and *P. notatum* (2n = 20, 30, 40, 60) (Espinoza *et al.* 2002; Pozzobon & Valls 1997). A precondition for using DNA barcoding to identify species is that species are monophyletic (Hebert *et al.* 2003b). However, lots of studies have demonstrated that many species from Asteraceae are paraphyletic or polyphyletic (Godoy *et al.* 2017; Rieseberg & Brouillet 1994), which is likely

to be another reason for the low discrimination rates. In addition to these reasons, another possible factor could be improper taxonomic treatment or cryptic speciation (Zhang *et al.* 2012). As for the combination of ITS/ITS2+*rbcL*, the discriminatory power improved very little compared with ITS/ITS2 alone. Finally, *rbcL* was combined with ITS/ITS2+*matK*, but the species resolutions were not increased (88.44% and 88.15%, respectively) based on the NJ tree method. Overall, *rbcL* made a minimal contribution to improving the resolution power (Yang *et al.* 2012), and *matK* is more appropriate for supplementing ITS/ITS2.

The performance of DNA barcoding as a potential resource for identifying invasive plants around the world

In this study, 16 potentially invasive plant species that have not been introduced into China yet, were also included in this study and eight of them are listed as 100 of the world's worst invasive alien species (Lowe *et al.* 2000), for instance, *Schinus terebinthifolia* (Brazilian pepper tree), which has no congener in China. The sequences of *S. terebinthifolia* have been included into the database and result of NJ tree showed that five individuals have formed a monophyly with a 99% node support, which indicated the dataset could be an effective way to prevent world's worst invasive alien species from entering a country in the first place.

Furthermore, 139 plant species native to China have been sampled, and some of them are potential to be or have been invasive species in other countries. For instance, *Croton tiglium*, a toxic plant, was introduced into America and became an invasive species (Chen & Zheng 1987). *Myriophyllum spicatum* and *M. verticillatum* are popular aquarium plants and widely

distributed in Eurasia, considered as invasive in all or part of North America (Thum *et al.* 2012). Under these situations, our dataset could be a useful resource for all countries other than China.

Conclusions

When a species is introduced into a new eco-system, it may cause permanent damage to the ecology, or lead to the endangerment or even extinction of indigenous species, or inflict damage to economy. Though the full impact is often unpredictable, correct identification will lay a solid foundation for further action. This study demonstrated that DNA barcoding is an efficient tool for identifying invasive species and ITS/ITS2+*matK* are the most suitable barcode for invasive plants in China. This study represents the first step towards building a worldwide resource for identifying invasive plants and our long-term goal is to expand and complete the existing library.

Acknowledgements

We sincerely thank You-Sheng Chen, Wen-Li Chen, Tian-Gang Gao, De-Yuan Hong and Shu-Ren Zhang, Institute of Botany, Chinese Academy of Science and David E. Boufford of Harvard University for the help with specimen identification. We also thank Dr. Mu-Fan Geng for the help with data analysis and laboratory work. Thanks also to Herbarium of the Institute of Botany, Chinese Academy of Science, Beijing, China (PE), Chinese Academy of

Inspection and Quarantine, and Entry-Exit Inspection and Quarantine Bureau of Jiangsu and Shandong province for their help of collecting specimens. This study was supported by STS Program of Chinese Academy of Sciences (ZSSD-018).

References

Alves S, Tiago L, Chauveau O, Eggers L, Souza-Chies TT (2014) Species discrimination in *Sisyrinchium* (Iridaceae): assessment of DNA barcodes in a taxonomically challenging genus. *Molecular Ecology Resources* **14**, 324-335.

Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M (2010) Identification of medicinal *Dendrobium* species by phylogenetic analyses using *matK* and *rbcL* sequences. *Journal of Natural Medicines* **64**, 133-138.

Ashfaq M, Asif M, Anjum Z, Zafar Y (2013) Evaluating the capacity of plant DNA barcodes to discriminate species of cotton (*Gossypium*: Malvaceae). *Molecular Ecology Resources* **13**, 573-582.

Barrett RD, Hebert PD (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* **83**, 481-491.

Blagoev GA, deWaard JR, Ratnasingham S, *et al.* (2016) Untangling taxonomy: a DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources* **16**, 325-341.

Breman FC, Loix S, Jordaens K, Snoeks J, Van Steenberge M (2016) Testing the potential of DNA barcoding in vertebrate radiations: the case of the littoral cichlids (Pisces, Perciformes, Cichlidae) from Lake Tanganyika. *Molecular Ecology Resources* **16**, 1455-1464.

Burson BL (1985) Cytology of *Paspalum chacoence* and *P. durifolium* and their relationship to *P. dilatatum*. *Botanical Gazette* **146**, 124-129.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* **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 J, Zhao J, Erickson DL, Xia N, Kress WJ (2015) Testing DNA barcodes in closely related species of *Curcuma* (Zingiberaceae) from Myanmar and China. *Molecular Ecology Resources* **15**, 337-348.

Chen JS, Zheng S (1987) Poisonous plants of China. Science Press, Beijing:.

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

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

Dandin S, Chennaveeraiah M (1983) Chromosome number and meiotic behaviour in interpretation of basic chromosome number in the genus *Paspalum*. *Journal of Cytology and Genetics* **18**, 26-33.

Ding JQ, Wang R, Fan ZN (1995) Occurrence, harms and control of *Eichhornia crassipes* (Mart.) Solms in China. *Chinese Journal of Weed Science* **9**, 49-52.

Espinoza F, Pessino SC, Quarin CL, Valle EM (2002) Effect of pollination timing on the rate of apomictic reproduction revealed by RAPD markers in *Paspalum notatum*. *Annals of Botany* **89**, 165-170.

Fazekas AJ, Burgess KS, Kesanakurti PR, *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* **3**, e2802.

Fazekas AJ, Kesanakurti PR, Burgess KS, *et al.* (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* **9**, 130-139.

Fu ZX, Jiao BH, Nie B, Zhang GJ, Gao TG (2016) A comprehensive generic‐level phylogeny of the sunflower family: Implications for the systematics of Chinese Asteraceae. *Journal of Systematics and Evolution* **54**, 416-437.

Ghahramanzadeh R, Esselink G, Kodde L, *et al.* (2013) Efficient distinction of invasive aquatic plant species from non-invasive related species using DNA barcoding. *Molecular Ecology Resources* **13**, 21-31.

Godoy SMD, Silva JFMD, Paula GBND, *et al.* (2017) Phylogenetic relationships of Brazilian *Mikania* species (Asteraceae, Eupatorieae) based on multilocus DNA markers. *Botanical Journal of the Linnean Society* **184**, 326-346.

Hajibabaei M, Singer GA, Hickey DA (2006) Benchmarking DNA barcodes: an assessment using available primate sequences. *Genome* **49**, 851-854.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98.

Hebert PD, Ratnasingham S, de Waard JR (2003a) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, S96-S99.

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

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and Using a Plant DNA Barcode. *PLoS One* **6**, e19254.

Hoveka L, Van der Bank M, Boatwright J, Bezeng B, Yessoufou K (2016) The noncoding *trnH-psbA* spacer, as an effective DNA barcode for aquatic freshwater plants, reveals prohibited invasive species in aquarium trade in South Africa. *South African Journal of Botany* **102**, 208-216.

Huang QQ, Wu JM, Bai YY, Zhou L, Wang GX (2009) Identifying the most noxious invasive plants in China: role of geographical origin, life form and means of introduction. *Biodiversity and Conservation* **18**, 305-316.

IBM Corp. (2010) *IBM SPSS Statistics for windows, Version 19.0* IBM Corp., Armonk, NY.

Käss E, Wink M (1996) Molecular evolution of the Leguminosae: phylogeny of the three subfamilies based on *rbcL*-sequences. *Biochemical Systematics and Ecology* **24**, 365-378.

Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066.

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 of the United States of America* **102**, 8369-8374.

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

León-Romero Y, Mejía O, Soto-Galera E (2012) DNA barcoding reveals taxonomic conflicts in the Herichthys bartoni species group (Pisces: Cichlidae). *Molecular Ecology Resources* **12**, 1021-1026.

Li DZ, Gao LM, Li HT, *et al.* (2011a) 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 of the United States of America* **108**, 19641-19646.

Li DZ, Liu JQ, Chen ZD, *et al.* (2011b) Plant DNA barcoding in China. *Journal of Systematics and Evolution* **49**, 165-168.

Li HQ, Chen JY, Wang S, Xiong SZ (2012) Evaluation of six candidate DNA barcoding loci in *Ficus* (Moraceae) of China. *Molecular Ecology Resources* **12**, 783-790.

Li ZY, Fan XH, Boufford DE (2015) Naturalized and Invasive Plants in China. In: *Plants of China: A Companion to the Flora of China*, p. 397.

Liu J, Liang SC, Liu FH, Wang RQ, Dong M (2005) Invasive alien plant species in China: regional distribution patterns. *Diversity and Distributions* **11**, 341-347.

Liu JX, Shi LC, Han JP, *et al.* (2014) Identification of species in the angiosperm family Apiaceae using DNA barcodes. *Molecular Ecology Resources* **14**, 1231-1238.

Lowe S, Browne M, Boudjelas S, De Poorter M (2000) *100 of the world's worst invasive alien species: a selection from the global invasive species database.* The Invasive Species Specialist Group (ISSG), Auckland.

Ma JS, Yan XL, Shou HY (2013) The checklist of the Chinese invasive plants. Higher Education Press, Beijing.

Mack RN, Simberloff D, Mark Lonsdale W, *et al.* (2000) Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Applications* **10**, 689-710.

Meier R, Shiyang K, Vaidya G, Ng PK (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* **55**, 715-728.

Mendoza ÁM, Torres MF, Paz A, *et al.* (2016) Cryptic diversity revealed by DNA barcoding in Colombian illegally traded bird species. *Molecular Ecology Resources* **16**, 862-873.

Mooney HA, Hobbs RJ (2000) *Invasive species in a changing world.* Island Press, Washington.

Panero J, Zelnik M (2014) *Human dimension and interior space: a source book of design reference standards.* Watson-Guptill, New York.

Park DS, Potter D (2013) A test of Darwin's naturalization hypothesis in the thistle tribe shows that close relatives make bad neighbors. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17915-17920.

Parveen I, Singh HK, Raghuvanshi S, Pradhan UC, Babbar SB (2012) DNA barcoding of endangered Indian *Paphiopedilum* species. *Molecular Ecology Resources* **12**, 82-90.

Pimentel D, McNair S, Janecka J, *et al.* (2001) Economic and environmental threats of alien plant, animal, and microbe invasions. *Agriculture, Ecosystems & Environment* **84**, 1-20.

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics* **52**, 273-288.

Pozzobon MT, Valls JFM (1997) Chromosome number in germplasm accessions of *Paspalum notatum* (Gramineae). *Brazilian Journal of Genetics* **20**, 29-34.

Pyšek P, Hulme PE, Meyerson LA, *et al.* (2013) Hitting the right target: taxonomic challenges for, and of, plant invasions. *AoB Plants* **5**, plt042.

Qiang S, Cao XZ (2000) Survey and analysis of exotic weeds in China. *Journal of Plant Resources and Environment* **9**, 34-38.

R Development Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

Rieseberg LH, Brouillet L (1994) Are many plant species paraphyletic? *Taxon*, 21-32.

Rieseberg LH, Wood TE, Baack EJ (2006) The nature of plant species. *Nature* **440**, 524-527.

Saarela JM, Wysocki WP, Barrett CF, *et al.* (2015) Plastid phylogenomics of the cool-season grass subfamily: clarification of relationships among early-diverging tribes. *AoB Plants* **7**, plv046.

Soreng RJ, Peterson PM, Romaschenko K, *et al.* (2015) A worldwide phylogenetic classification of the Poaceae (Gramineae). *Journal of Systematics and Evolution* **53**, 117-137.

Sosa V, Mejía-Saules T, Cuéllar MA, Vovides AP (2013) DNA barcoding in endangered Mesoamerican groups of plants. *The Botanical Review* **79**, 469-482.

Srivathsan A, Meier R (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics* **28**, 190-194.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725-2729.

Thum RA, Mercer AT, Wcisel DJ (2012) Loopholes in the regulation of invasive species: genetic identifications identify mislabeling of prohibited aquarium plants. *Biological Invasions* **14**, 929-937.

Van De Wiel C, Van Der Schoot J, Van Valkenburg J, Duistermaat H, Smulders M (2009) DNA barcoding discriminates the noxious invasive plant species, floating pennywort (*Hydrocotyle ranunculoides* L.f.), from non-invasive relatives. *Molecular Ecology Resources* **9**, 1086-1091.

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

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

Wojciechowski MF, Lavin M, Sanderson MJ (2004) A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* **91**, 1846-1862.

Xia FY (1983) Ragweed pollen allergy. *Plant Quarantine* **6**, 16-19.

Xiang XG, Hu H, Wang W, Jin XH (2011) DNA barcoding of the recently evolved genus *Holcoglossum* (Orchidaceae: Aeridinae): a test of DNA barcode candidates. *Molecular Ecology Resources* **11**, 1012-1021.

Xie Y, Li ZY, Gregg WP, Li DM (2001) Invasive species in China—an overview. *Biodiversity & Conservation* **10**, 1317-1341.

Xu SZ, Li DZ, Li JW, *et al.* (2015) Evaluation of the DNA barcodes in *Dendrobium* (Orchidaceae) from mainland Asia. *PLoS One* **10**, e0115168.

Yan LJ, Liu J, Möller M, *et al.* (2015) DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya–Hengduan Mountains. *Molecular Ecology Resources* **15**, 932-944.

Yang JB, Wang YP, Moeller M, Gao LM, Wu D (2012) Applying plant DNA barcodes to identify species of *Parnassia* (Parnassiaceae). *Molecular Ecology Resources* **12**, 267-275.

Zhang CY, Wang FY, Yan HF, *et al.* (2012) Testing DNA barcoding in closely related groups of *Lysimachia* L.(Myrsinaceae). *Molecular Ecology Resources* **12**, 98-108.

Author contributions

X.-H.J., S.-Z.X. and Z.-Y.L. conceived and designed the study; Z.-Y.L., X.-H.J. and S.-Z.X. collected plant materials; S.-Z.X. performed the laboratory works, analysed data and wrote the manuscript. Z.-Y.L. and X.-H.J. revised and finalized the manuscript.

Data Accessibility

GenBank accession numbers for nucleotide sequences: see Table S1 (supporting information).

DNA sequences and tree files: Dryad doi:10.5061/dryad.4t13k

Table captions

Table 1 Characteristics of the five single markers and seven combinations evaluated in this study.

Table 2 Identification success rates obtained using distance and NJ tree methods for the five single markers and seven combinations.

Table 3 Identification success rates based on the similarity method using 'best match' and 'best close match' models in the TaxonDNA program.

Figure legends

Fig. 1 Scatter plots of the maximum intraspecific K2P distance versus minimum interspecific K2P distance for five single markers and seven combinations (I, internal transcribed spacer (ITS); I2, ITS2; M, *matK*; R, *rbcL*; T, *trnH-psbA*).

Fig. 2 Neighbour-joining tree based on the combination of ITS+*matK*. Coloured clades (except black) represent species that were correctly identified, and different families are colour-coded. Black clades represent species that were not identified successfully. Further details are included in supplementary Fig. S7.

Supporting Information

Fig. S1 Frequency distribution of intraspecific and interspecific K2P distance for five single markers and seven combinations.

Fig. S2 Neighbour-joining tree based on ITS. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S3 Neighbour-joining tree based on ITS2. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S4 Neighbour-joining tree based on *matK*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S5 Neighbour-joining tree based on *rbcL*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S6 Neighbour-joining tree based on *trnH-psbA*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S7 Neighbour-joining tree based on ITS+*matK*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S8 Neighbour-joining tree based on ITS2+*matK*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S9 Neighbour-joining tree based on ITS+*rbcL*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S10 Neighbour-joining tree based on ITS2+*rbcL*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S11 Neighbour-joining tree based on *matK*+*rbcL*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S12 Neighbour-joining tree based on ITS+*matK*+*rbcL*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S13 Neighbour-joining tree based on ITS2+*matK*+*rbcL*. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S14 Neighbour-joining tree of ITS including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S15 Neighbour-joining tree of ITS2 including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S16 Neighbour-joining tree of *matK* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S17 Neighbour-joining tree of *rbcL* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S18 Neighbour-joining tree of *trnH-psbA* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S19 Neighbour-joining tree of ITS+*matK* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S20 Neighbour-joining tree of ITS2+*matK* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S21 Neighbour-joining tree of ITS+*rbcL* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S22 Neighbour-joining tree of ITS2+*rbcL* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S23 Neighbour-joining tree of *matK*+*rbcL* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S24 Neighbour-joining tree of ITS+*matK*+*rbcL* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Fig. S25 Neighbour-joining tree of ITS2+*matK*+*rbcL* including fifteen unknown species. Bootstrap values (>50%) are shown above the relevant branches.

Table S1 GenBank accession numbers of five markers for all samples used in this study.

Table S2 Sample details with voucher information and GenBank accession numbers of four markers newly sequenced in this study.

Table S3 Primers used for amplification and sequencing of four single markers in this study.

Table S4 Chi-square tests of percentage of species having distinct barcoding gaps between any two markers.

Table 1 Characteristics of the five single markers and seven combinations evaluated in this study

DNA region	Success rate of PCR and sequencing	No. of sampled species (individuals)	sequences with valid conspecifics	species with valid conspecifics	Mean intraspecific distance	Minimum intraspecific distance	Maximum intraspecific distance	Mean interspecific distance	Minimum interspecific distance	Maximum interspecific distance
ITS	90.24%	583 (2491)	96.62%	85.59%	0.0144	0	0.312	0.5202	0	1.679
ITS2	/	583 (2491)	96.62%	85.59%	0.0172	0	0.23	0.9146	0	2.868
matK	92.07%	511 (2022)	95.49%	82.19%	0.0035	0	0.063	0.2896	0	0.793
rbcL	100%	526 (2198)	96.63%	85.93%	0.0023	0	0.059	0.1049	0	0.221
trnH-psbA	98.17%	365 (1150)	89.47%	66.84%	0.0085	0	0.0887	0.6112	0	1.651
ITS+*matK*	/	457 (1389)	94.09%	82.05%	/	/	/	/	/	/
ITS2+*matK*	/	457 (1389)	94.09%	82.05%	/	/	/	/	/	/
ITS+*rbcL*	/	476 (1485)	94.88%	84.03%	/	/	/	/	/	/
ITS2+*rbcL*	/	476 (1485)	94.88%	84.03%	/	/	/	/	/	/
matK+*rbcL*	/	449 (1546)	94.82%	82.18%	/	/	/	/	/	/
ITS+*matK*+*rbcL*	/	433 (1255)	93.14%	80.13%	/	/	/	/	/	/
ITS2+*matK*+*rbcL*	/	433 (1255)	93.14%	80.13%	/	/	/	/	/	/

Table 2 Identification success rates obtained using distance and NJ tree methods for the five single markers and seven combinations

DNA region	Distance method	NJ tree method	NJ tree method (including unknown species)
ITS	83.06%	76.50%	93.33% (14/15)
ITS2	80.65%	76.84%	93.33% (14/15)
matK	70.00%	63.21%	66.67% (8/12)
rbcL	57.96%	52.86%	53.33% (8/15)
trnH-psbA	72.95%	50.68%	66.67% (10/15)
ITS+*matK*	90.08%	91.03%	100% (12/12)
ITS2+*matK*	88.74%	85.78%	91.67% (11/12)
ITS+*rbcL*	85.25%	80.25%	53.33% (8/15)
ITS2+*rbcL*	82.25%	77.94%	53.33% (8/15)
matK+*rbcL*	75.34%	71.94%	58.33% (7/12)
ITS+*matK*+*rbcL*	88.44%	85.45%	91.67% (11/12)
ITS2+*matK*+*rbcL*	88.15%	81.06%	75% (9/12)

Table 3 Identification success rates based on the similarity method using 'best match' and 'best close match' models in the TaxonDNA program

DNA region	Best Match			Best Close Match		
	Correct	Ambiguous	Incorrect	Correct	Ambiguous	Incorrect
ITS	91.69%	5.02%	3.28%	89.73%	5.02%	2.57%
ITS2	88.36%	9.09%	2.53%	85.99%	8.97%	1.70%
matK	82.70%	13.15%	4.14%	82.44%	13.10%	3.78%
rbcL	73.68%	21.61%	4.70%	73.49%	21.61%	4.70%
trnH-psbA	84.93%	11.75%	3.30%	83.96%	11.66%	3.10%
ITS+*matK*	96.48%	1.60%	1.91%	94.41%	1.60%	1.37%
ITS2+*matK*	96.48%	2.29%	1.22%	94.95%	2.29%	0.99%
ITS+*rbcL*	93.39%	3.05%	3.54%	90.84%	3.05%	2.83%
ITS2+*rbcL*	91.48%	4.96%	3.54%	89.56%	4.96%	3.33%
matK+*rbcL*	90.58%	5.11%	4.29%	90.45%	5.11%	4.22%
ITS+*matK*+*rbcL*	96.32%	1.36%	2.30%	94.69%	1.36%	2.22%
ITS2+*matK*+*rbcL*	96.32%	1.53%	2.13%	95.46%	1.53%	1.96%

Fig. 1 Scatter plots of maximum intraspecific K2P distance versus minimum interspecific K2P distance for five single markers and seven combinations (I, internal transcribed spacer (ITS); I2, ITS2; M, *matK*; R,*rbcL*; T, *trnH-psbA*).

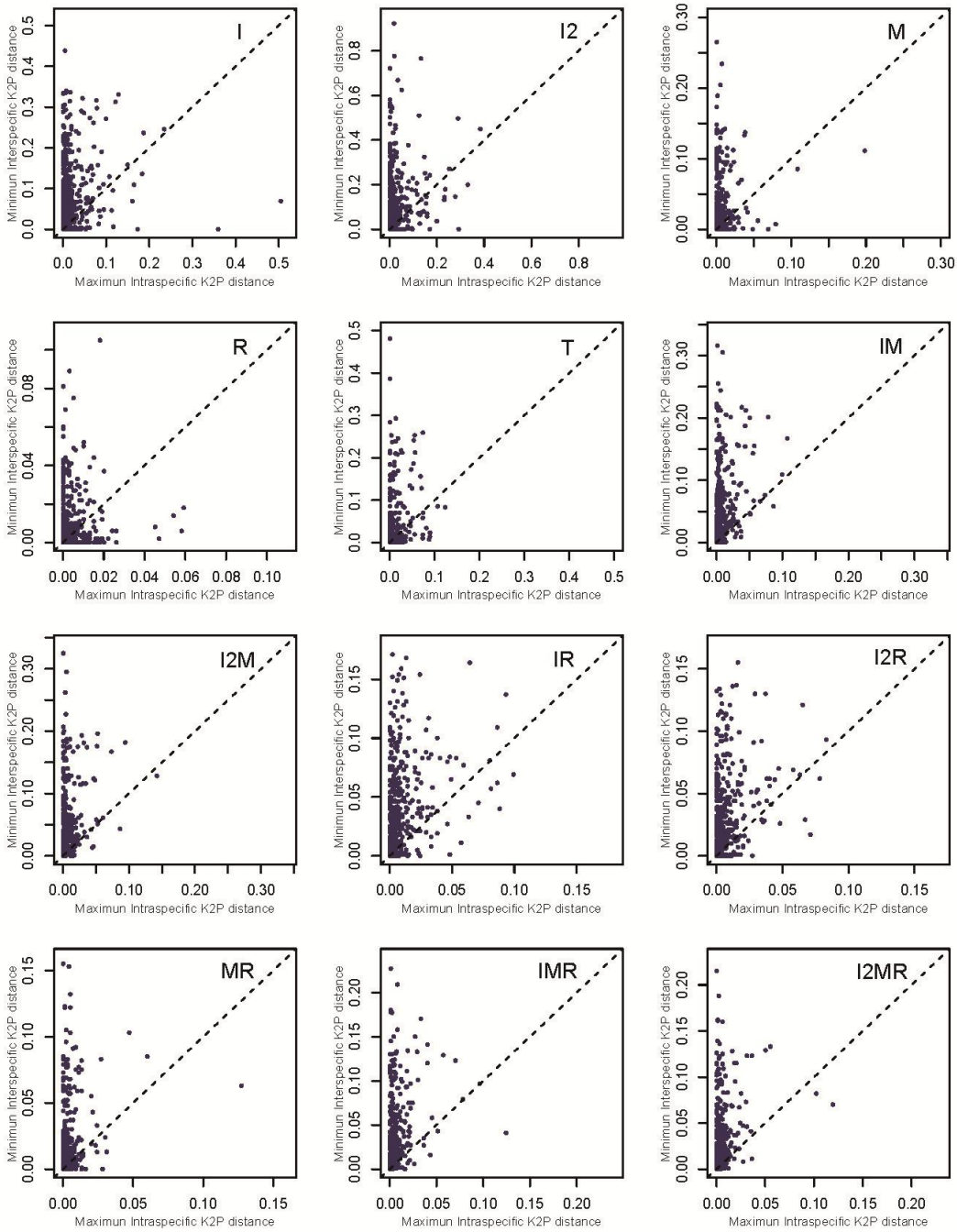

Figure 2 Neighbour-joining tree based on the combination of ITS+*matK*. Coloured clades (except black) represent species that were correctly identified, and different families are colour-coded. Black clades represent species that were not identified successfully. Further details are included in supplementary Fig. S7.