JSE Journal of Systematics and Evolution

Journal of Systematics and Evolution 49 (3): 176–181 (2011) doi: 10.1111/j.1759-6831.2011.00134.x

Research Article

New universal *matK* primers for DNA barcoding angiosperms

[1,2]Jing YU [1]Jian-Hua XUE [1]Shi-Liang ZHOU*

[1](*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences,* Beijing 100093, China)
[2](*Graduate University of the Chinese Academy of Sciences,* Beijing 100049, China)

Abstract The chloroplast maturase K gene (*matK*) is one of the most variable coding genes of angiosperms and has been suggested to be a "barcode" for land plants. However, *matK* exhibits low amplification and sequencing rates due to low universality of currently available primers and mononucleotide repeats. To resolve these technical problems, we evaluated the entire *matK* region to find a region of 600–800 bp that is highly variable, represents the best of all *matK* regions with priming sites conservative enough to design universal primers, and avoids the mononucleotide repeats. After careful evaluation, a region in the middle was chosen and a pair of primers named *matK*472F and *matK*1248R was designed to amplify and sequence the *matK* fragment of approximately 776 bp. This region encompasses the most variable sites, represents the entire *matK* region best, and also exhibits high amplification rates and quality of sequences. The universality of this primer pair was tested using 58 species from 47 families of angiosperm plants. The primers showed a strong amplification (93.1%) and sequencing (92.6%) successes in the species tested. We propose that the new primers will solve, in part, the problems encountered when using *matK* and promote the adoption of *matK* as a DNA barcode for angiosperms.

Key words angiosperm, DNA barcoding, *matK*, primer.

DNA barcoding, a concept that has recently become popular, is characterized by using one or a few DNA fragments to identify different species (Kress et al., 2005). The mitochondrial cytochrome *c* oxidase subunit 1 (*CO1*) gene was selected to be a DNA barcode for animal species (Hebert et al., 2003). However, consensus is still to be reached regarding gene fragments for a plant DNA barcode, although the Consortium for the Barcode of Life (CBOL) Plant Working Group (2009) has suggested *matK* + *rbcL*. The difficulty in selecting specific genes to be a plant barcode is due to the imperfection of any gene from either the chloroplast, mitochondrial, or nuclear genomes. The plant mitochondrial genes evolve slowly and, therefore, are ineffective for distinguishing between different plant species. Plant nuclear genes often occur in multiple copies and are highly variable, making the design of universal primers difficult. The search for a plant DNA barcode has focused on genes of the chloroplast genome and several candidates, such as *acc*D, *atpF–atpH*, *matK, nhdJ*, *psbK–psbI*, *rbcL*, *rpoB*, *rpo*C1, and *trnH–psbA* have been proposed (Chase et al., 2005; Kress et al., 2005; Newmaster et al., 2006; Yoo et al., 2006). Unfortunately, very few of these loci are variable enough to identify or distinguish between plant species when used alone, leading to the use of several combinations of loci, such as *rpoC1* + *rpoB* + *matK*, *rpoC1* + *matK* + *trnH–psbA*, *rbcL* + *trnH–psbA*, *matK* + *atpF-H* + *psbK-I*, and *matK* + *atpF-H* + *trnH–psbA* (Yoo et al., 2006; Chase et al., 2007; Kress & Erickson, 2007; CBOL Plant Working Group, 2009).

Among the candidate barcode genes, *matK* is one of the most promising candidates for a plant barcode. The *matK* gene is approximately 1570 bp in length and codes for a maturase protein. The coding region of *matK* is generally located within an intron of the chloroplast *trnK* gene, except in some ferns, in which it encodes $tRNA^{Lys}$(UUU) (Neuhaus & Link, 1987). Being a coding region, the very high evolutionary rate of *matK* has made it usable in phylogenetic reconstructions at high taxonomic levels, such as Order or Family, and sometimes also at low taxonomic levels, such as Genus or Species (Wolfe, 1991; Hilu et al., 2003; Müller et al., 2006; Chase et al., 2007; Lahaye et al., 2008). Although extensive divergence in the *matK* sequence for higher taxonomic categories results in questionable positions and relationships of some phylogenetic clades, *matK* is very useful in DNA barcoding for the identification of

Received: 16 December 2010 Accepted: 8 March 2011
* Author for correspondence. E-mail: slzhou@ibcas.ac.cn; Tel.: 86-10-6283-6503; Fax: 86-10-6259-0843.

© 2011 Institute of Botany, Chinese Academy of Sciences

plant families (Qiu et al., 1999; Li & Zhou, 2007; Gao et al., 2008).

In systematics, the full-length *matK* sequence is amplified and sequenced with primers designed within the *trnK* region (Wang et al., 2006; Li & Zhou, 2007); however, for DNA barcoding, a fragment of 600–800 bp is usually sufficient. Primers designed within the *matK* region can be problematic for PCR amplification and sequencing, leading to the design of new taxon-specific primers or modifications of existing primers. For example, primers 390F (5′-CGA TCT ATT CAT TCA ATA TTT C-3′) and 1326R (5′-TCT AGC ACA CGA AAG TCG AAG T-3′) have been designed for the order Caryophyllales (Cuenoud et al., 2002), primers F (5′-GGG TTG CTA ACT CAA CGG GTA G-3′) yes and R (5′-GGT TGC CCG GGA CTC GAA CCC GGA ACT CGT CGG-3′) have been designed for *Equisetum* (Hausner et al., 2006), and primers SpRend (5′-TTA TGT AAT CCG TGT AAA AT-3′) and 5UTRSp (5′-CAT TTC ATA ACA CAA GAA-3′) have been designed for orchids (Barthet, 2006). Primers 2.1-Myristicaceae and 5-Myristicaceae, designed for the family Myristicaceae (Newmaster et al., 2008), were based on primers 2.1F (5′-CCT ATC CAT CTG GAA ATC TTA G-3′) and 5R (5′-GTT CTA GCA CAA GAA AGT CG-3′), respectively. Some of the taxon-specific primers are applicable to other taxa and were thus recommended for use in plant DNA barcoding. However, difficulties were often encountered when the available primers were used (Sass et al., 2007; Fazekas et al., 2008). Therefore, order-specific primers were proposed by Dunning & Savolainen (2010). However, it is important to note that plant DNA barcoding would best be suited by only a few universal primers, as opposed to many taxon-specific primers, which would be beneficial in the event that the identification of a plant sample must be based solely on knowledge that the sample is a seed plant.

Despite these issues, *matK* remains one of the best choices for plant DNA barcoding. Determination of the individual region of *matK* that best represents the entire gene and design of additional universal primers is an important next step. In the present study, we first determined the individual region of *matK* that best represents the entire gene and then designed primers in adjacent and more conserved regions of the area. The universality of the new primers is tested by amplifying and sequencing some samples selected randomly from 33 orders. The ultimate goal of the present study was to provide new alternative primers for *matK* for use in plant DNA barcoding.

1 Material and methods

1.1 *matK* sequence analysis and primer design

The *matK* gene sequences were obtained from 106 chloroplast genomes available from GenBank, aligned using Clustal X 2.0 (Larkin et al., 2007), and adjusted manually with Se-Al Version 2.0 a11 (Rambaut, 1996; Larkin et al., 2007). Ten 600-bp data subsets were generated by sampling the data set of the entire length and analyzed with PAUP 4.0b10 (Swofford, 2000) to reconstruct the phylogeny. COMPONENT Version 2.00a software (Page, 1993) was used to calculate the distances between the phylogenetic trees based on the entire gene length and the subsets with the smallest distance considered to be the best representative of the entire length. Variability of *matK* was estimated by DNAsp Version 5.0 (Librado & Rozas, 2009) using the sliding window method. Conserved regions were identified and primers were designed with the help of Primer Premier 5.0 (Premier Biosoft International, Palo Alto, CA, USA).

1.2 Amplification and sequencing with new primers

Gradient PCR was performed using five samples representing basal angiosperms, monocots, basal eudicots, rosids, and asterids in order to determine the optimum annealing temperatures of the newly designed primers *matK*472F (5′-CCC RTY CAT CTG GAA ATC TTG GTT C-3′) and *matK*1248R (5′-GCT RTR ATA ATG AGA AAG ATT TCT GC-3′). The universality of the primers was tested using 58 samples from 47 families (7 species from 6 families of basal angiosperms, 12 species from 10 families of monocots and 39 species from 31 families of dicots, Table S1), with 390F/1326R as a control. Total DNA was extracted using the cetyltrimethyl ammonium bromide (CTAB) method (Doyle & Doyle, 1987). The PCR reaction mixture consisted of 1× PCR buffer, 0.5 mmol/L dNTPs, 0.25 μmol/L each primer, 1 U Taq polymerase, and 5–50 ng template DNA. Thermal cycling conditions were as follows: 94 °C for 3 min, followed by 40 cycles of 94 °C for 30 s, 48 °C for 40 s, 72 °C for 1 min, and a final extension at 72 °C for 10 min. The PCR products were verified by electrophoresis in 1% agarose gels stained with ethidium bromide. Repetitions of PCR were performed again under the same conditions for those samples of no band or weak bands. The PCR products were sent to the Beijing Genomic Institute (BGI; Shenzhen, China) for sequencing after simple purification. Sequences were analyzed by Sequencher 4.7 (Gene Codes, Ann Arbor, MI, USA).

© 2011 Institute of Botany, Chinese Academy of Sciences

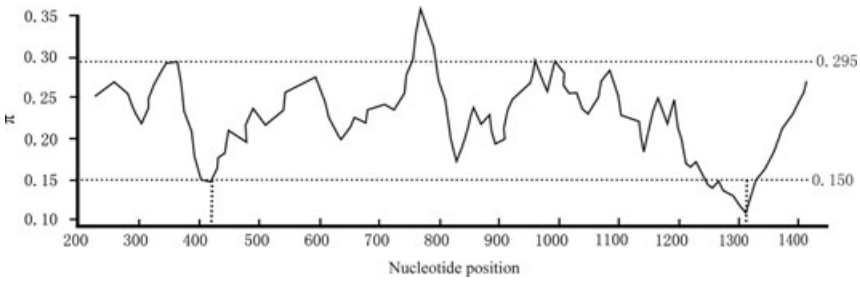

Fig. 1. Variability of *matK* sequences of seed plants parameterized by nucleotide diversity (π). Four peaks (above the upper horizontal dotted line) indicate the most variable regions and two valleys (indicated by vertical dotted lines) mark the most conservative regions.

2 Results

2.1 Variability of the *matK* region and priming sites

There are 129 entire-chloroplast genomes of seed plants deposited in GenBank thus far. Of these, 18 gymnosperms, 25 monocots, and 86 dicots contain the *matK* gene. The size of the *matK* gene is 1535 bp in monocots and 1515–1580 bp in dicots. The aligned length was 1524 bp after removal of some taxa-unique inserts in a few genomes. The sequence was remarkably variable throughout the entire length. There are four peaks with $\pi \geq 0.295$, and two valleys with $\pi < 0.150$ (Fig. 1). The conserved regions of the valleys are ideal for primer design.

2.2 Most representative segments of the entire length of *matK*

Using the length of the *CO1* gene as a reference, the total length of *matK* was sequentially grouped into 10 data subsets of 600 bp, with 100-bp intervals. Results from the neighbor joining (NJ) method showed that data subset 8 best represented the entire length, followed by data subsets 9 and 7. However, the maximum parsimony (MP) method suggested data subset 7 and the maximum likelihood (ML) method suggested data subset 9 (Table 1). Taken together, the sequences from 600 to 1400 bp are more representative of the entire length than the other regions, according to the relatively smaller distances.

2.3 Priming sites, primer universality, PCR amplification, and sequencing successes

The exact sites of the primers were adjusted within the two most conserved regions for better priming and excluding a poly-A region. The priming site of *matK*472F is roughly at the same area as 2.1F/2.1aF and AST_F/*matK*_KewR/1R_KIM, and the *matK*1248R overlaps partly with MALV_R1 (Fig. 2). Gradient PCR revealed an optimum temperature below 52 °C for *matK*472F and *matK*1248R, with the strongest bands at 48 °C. However, some samples amplified well at 60 °C. Therefore, an annealing temperature range of 48–52 °C can be considered feasible. In total, 58 samples (Table S1) were amplified and sequenced using the primer pair *matK*472F/*matK*1248R, with 390F/1326R as a control in amplification under the same conditions. The primers worked well for angiosperms, with an average amplification success of 93.1% compared with 25.9% in the control (Table 2). Most amplicons were strong enough for isolation of bands or direct sequencing. Although weak bands were observed, no apparently adverse effects were detected. After trimming off the two ends of low quality, the average sequence length of reads was 684 bp, with a mean quality value 97.8%, using *matK*472F and 655 bp, with a mean quality value 94.7%, using *matK*1248R.

3 Discussion

The major problem of using *matK* as a barcode is the difficulty in amplifying and sequencing the gene

Table 1 Distances between the phylogenetic tree based on the entire length of *matK* sequences and the trees based on the length of 600 bp

		Distance		
Data subset	Sequence position (bp)	NJ	MP	ML
1	1–600	18	14	28
2	100–700	21	13	22
3	200–800	17	14	36
4	300–900	15	17	28
5	400–1000	13	13	32
6	500–1100	12	16	36
7	600–1200	11	**12**	26
8	700–1300	**7**	14	22
9	800–1400	10	15	**10**
10	900–1524	13	18	20

The neighbor-joining (NJ), maximum parsimony (MP), and maximum likelihood (ML) tree-building methods, were compared. The smallest distances are indicated by boxes.

© 2011 Institute of Botany, Chinese Academy of Sciences

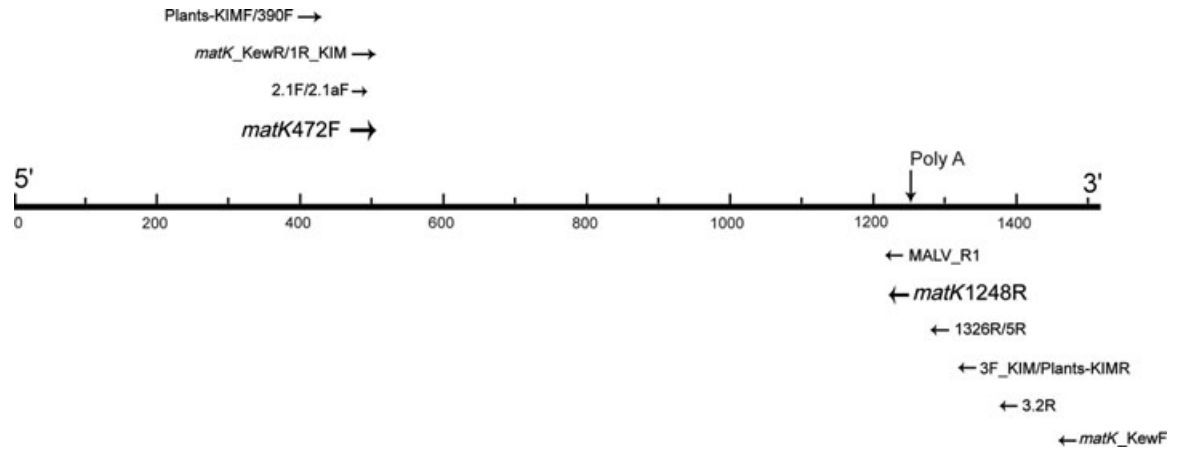

Fig. 2. The position and orientation of the new *matK* primer pair *matK*472F + *matK*1248R and some other primers on the entire length of the *matK* gene of *Arabidopsis thaliana* (1515 bp). The primer sequences are given in Table S2.

fragments. Early studies adopted a strategy of using the more highly conserved *trnK* regions as priming positions for fragment amplification and several internal primers for sequencing. The internal primers were typically designed for specific taxa and the universality was limited (Fazekas et al., 2008). Of the primers designed, primer pair 390F/1326R is one pair of the best, as exemplified by Lahaye et al. (2008), although in the present study the amplification success was only 25.9%. Within the *matK* region there are a few hot spots for designing primers (Fig. 2). There are two major hot spots for forward primers: one is represented by 390F and the other by 1R_KIM. The *matK*472F falls in the second hot spot. The reverse primers were tiered mainly at three positions represented by 3F_KIM, 1326R, and *matK*_KewR. The *matK*1248R happens to overlap partially with MALV_R1. The new primers produce significantly improved amplification success for angiosperms. Many primers designed within the *matK* gene amplify fragments that may include a long mononucleotide repeat, which may cause problems in sequencing. The new primer pair excludes that region and thereby may increase sequencing success and quality.

There are many possible regions of the entire *matK* gene that may be used as a plant barcode. We determined the region from 600 to 1400 bp to be the best representative of the whole-gene sequence based on the smallest distances (Table 1). Considering the conservative regions for primer design and the mononucleotide repeats that cause sequencing problems, the final region of *matK* was adjusted to contain the sequence from 472 to 1248 bp. The new primers were named according to the starting and ending sites of the whole *matK* gene sequence of *Arabidopsis thaliana*.

Although *matK*472F + *matK*1248R worked reasonably well for the samples tested, its usefulness for all angiosperms needs to be evaluated. To avoid using too many degenerate bases, we have to ignore rare base substitutions. More specific primers can be designed by modifying *matK*472F and *matK*1248R according to the desired sequence. Failure of amplification can also be attributed to a low quality of template DNA. Amplification of the same samples was significantly enhanced after purifying the DNA. The new primer pair identified in the present study is able to function over a broad

Table 2 Universality test for the newly designed primer pair *matK*472F and *matK*1248R with 390F + 1326R as a control

		*matK*472F + *matK*1248R				Control	
Order	*n*	*n1*	*n1*/*n* (%)	*n2*	*n2*/*n1* (%)	*n3*	*n3*/*n* (%)
Angisperms	58	54	93.1	50	92.6	15	25.9
Basal angiosperms	7	7	100	6	85.7	3	42.9
Monocots	12	11	91.7	11	100	5	41.7
Eudicots	39	36	92.3	33	91.7	7	17.9
Basal eudicots	6	5	83.3	5	100	0	0
Asterids	12	11	91.7	10	90.9	3	25
Rosids	21	20	95.2	18	90.0	4	19.0
Total	58	54	93.1	50	92.6	15	25.9

n, number of samples (species) used; *n1*, number of samples amplified successfully; *n2*, number of samples sequenced successfully; *n3*, number of samples amplified successfully using 390F and 1326R.

© 2011 Institute of Botany, Chinese Academy of Sciences

range of annealing temperatures (from 38 to 52 °C, or even to 60 °C). We suggest using an annealing temperature of 48–52 °C for this pair of primers. However, performing an initial gradient PCR will identify the best annealing temperature before performing many amplifications. Additional optimization methods for PCR can be found in standard DNA barcoding protocols (Chase et al., 2007).

The sequences obtained using the primers given are the best representatives of the whole *matK* region and are thus of phylogenetic value. Taking the representative together with the variability of barcoding regions into consideration will potentiate species identification using phylogenetic methods when species delimitations are problematic.

Acknowledgements This study was supported by the Research Fund for the Large-scale Scientific Facilities of the Chinese Academy of Sciences (grant no. 2009-LSF-GBOWS-01 and KSCX2-YW-N-0807).

References

Barthet MM. 2006. Expression and function of the chloroplast-encoded gene *matK*. Blacksburg: Virginia Polytechnic Institute and State University.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences USA 106: 12 794–12 797.

Chase MW, Cowan RS, Hollingsworth PM, Petersen G, Seberg O, Jorgsensen T, Cameron KM, Carine M. 2007. A proposal for a standardised protocol to barcode all land plants. Taxon 56: 295–299.

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V. 2005. Land plants and DNA barcodes: short-term and long-term goals. Philosophical Transactions of the Royal Society Series B Biological Sciences 360: 1889–1895.

Cuenoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW. 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. American Journal of Botany 89: 132–144.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19: 11–15.

Dunning LT, Savolainen V. 2010. Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. Botanical Journal of the Linnean Society 164: 1–9.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH. 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLOS one 3: e2802.

Gao X, Zhu YP, Wu BC, Zhao YM, Chen JQ, Hang YY. 2008. Phylogeny of Dioscorea sect. Stenophora based on chloroplast *matK*, *rbcL* and *trnL-F* sequences. Journal of Systematics and Evolution 46: 315–321.

Hausner G, Olson R, Simon D, Johnson I, Sanders ER, Karol KG, McCourt RM, Zimmerly S. 2006. Origin and evolution of the chloroplast *trnK* (*matK*) intron: a model for evolution of group II intron RNA structures. Molecular Biology and Evolution 23: 380–391.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences 270: 313–321.

Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R. 2003. Angiosperm phylogeny based on *matK* sequence information. American Journal of Botany 90: 1758–1776.

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, Van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V. 2008. DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences USA 105: 2923–2928.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

Li X-X, Zhou Z-K. 2007. The higher-level phylogeny of monocots based on *matK*, *rbcL* and 18S rDNA sequences. Acta Phytotaxonomica Sinica 45: 113–133.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.

Müller KF, Borsch T, Hilu KW. 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK*, *trnT-F*, and *rbcL* in basal angiosperms. Molecular Phylogenetics and Evolution 41: 99–117.

Neuhaus H, Link G. 1987. The chloroplast tRNA Lys (UUU) gene from mustard (*Sinapis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. Current Genetics 11: 251–257.

Newmaster SG, Fazekas AJ, Ragupathy S. 2006. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. Botany 84: 335–341.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J. 2008. Testing candidate plant barcode regions in the Myristicaceae. Molecular Ecology Resources 8: 480–490.

Page RDM. 1993. COMPONENT: tree comparison software for Microsoft Windows Ver 2.00a. User's Guide. London: Natural History Museum.

Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404–407.

Rambaut A. 1996. Se-Al: Sequence Alignment Editor (Se-Al v2.0a11) [online]. Available from: http://tree.bio.ed.ac.uk/software/seal [accessed 24 March 2009].

© 2011 Institute of Botany, Chinese Academy of Sciences

Sass C, Little DP, Stevenson DW, Specht CD. 2007. DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. PLOS one 2: e1154.

Swofford DL. 2000. PAUP* 4.0: Phylogenetic analysis using parsimony (* and other methods). Version 4.0 b10. Sunderland, MA: Sinauer Associates.

Wang Y-L, Li Y, Zhang S-Z, Yu X-S. 2006. The utility of *matK* gene in the phylogenetic analysis of the genus Magnolia. Acta Phytotaxonomica Sinica 44: 135–147.

Wolfe KH. 1991. Protein-coding genes in chloroplast DNA: compilation of nucleotide sequences, data base entries, and rates of molecular evolution. In: Bogorad L, Vasil IK eds. Cell culture and somatic cell genetics of plants. Vol. 7B. SanDiego: Academic Press. 467–482.

Yoo HS, Eah J, Kim JS, Kim Y, Min M, Paek WK, Lee H, Kim C. 2006. DNA barcoding Korean birds. Molecules and Cells 22: 323–327.

Supporting Information

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

Table S1. Taxa used for testing the universality of new *matK* primers and the length and quality of the sequence reads using *matK*472F and *matK*1248R.

Table S2. Sequences of some representative primers for *matK* together with the new *matK*272F and *matK*1248R.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

© 2011 Institute of Botany, Chinese Academy of Sciences