RESEARCH ARTICLE Legume Research- An International Journal, Volume 45 Issue 6: 659-668 (June 2022)

Establishment of a DNA Barcoding Database for Legume and Grass Species Identification

Yongqing Li[1], Zunongjiang A. Bu La[1], Lijun Cao[1], Shengguo Zhao[2]

10.18805/AF-727

ABSTRACT

Background: DNA barcoding, an emerging approach, is being widely used to accurately and quickly identify species using conserved DNA sequences.

Methods: Herein we designed seven universal primers for *matK* (*matK1*, *matK2*, *matK3* and *matK4*), *rbcL*, *psbA-trnH* and *ITS* based on their nucleotide sequences in GenBank to amplify 40 species of leguminosae and grass forages. Sequence alignment was performed using MEGA 5.0 and haplotype and mutation sites were analyzed with DnaSP 5.10. PCR amplification efficiency on using the primers designed for *psbA-trnH* and *ITS* was relatively low, making these sequences unsuitable for DNA barcoding. Further, we optimized target fragment amplification conditions for all 40 species analyzed in this study. On purifying, sequencing and analysing amplification products, we selected 5′ - and 3′ -end conserved fragments in four marked fragments.

Result: Sequences of each maker loci showed that there were 12, 17 and 6 haplotypes of *matK1*, *matK2* and *matK3*, respectively and 13 of *rbcL*. Based on these haplotypes of *matK1*, *matK2*, *matK3* and *rbcL*, we established a DNA barcoding database for 20 forage species.

Key words: DNA barcoding, Forages, *MatK*, *RbcL*, Universal sequences.

INTRODUCTION

DNA barcoding is widely used for species identification considering that it is rapid, simple, economical and reliable. It involves amplifying one or more standard DNA regions and is particularly helpful when analyzing species that cannot be identified using biometric data. The success of this method is heavily dependent on distinct barcoding gap, *i.e.*, differences between intra- and interspecies nucleotide divergence (Kress *et al.*, 2005, Hollingsworth *et al.*, 2011). Hebert *et al.* (2003) studied variations of COI genes among 11 different animal phyla and were able to distinguish one species from another using them; however, in plants, COI and other mitochondrial genes show a low degree of variation and are thus not particularly effective for species identification. Chloroplast genes can be evidently used for plant species identification as they are abundant, demonstrate strong amplification ability and are largely immune to gene recombination, although available sequence data reveal the relatively conserved nature of chloroplast genomes in terms of both structure and gene content. As per the recommendation of the CBOL Plant Working Group, *rbcL*+*matK* can be used for DNA barcoding of terrestrial plants and species identification (CBOL Plant Working Group, 2009). Cai *et al.* (2021) found that the combination barcodes of *ITS*+*rbcL* achieve the accurate identification of *Uncaria* species when the DNA barcoding technology is applied to identify *Uncaria* species. Rashmi *et al.* (2020) analyzed characterize the intraspecific diversity among a total of 12 *Mucuna pruriens* accessions by using three barcode markers (*ITS2*, *matK* and *trnHpsbA*). Based on comparative analyses of a large sample, research team of plant barcode in China suggested that *ITS* or *ITS2* should be incorporated into the core barcode for seed plants (Yao *et al.*, 2010, China Plant BOL Group, 2011). Costion *et al.* (2015) used pedigree diversity index to identify tropical rainforest shelters and species differentiation centers to determine biodiversity conservation priorities. *ITS2* sequence has been used for quickly and effectively identifying traditional Chinese medicine from Amomum and Alpinia (Wang *et al.*, 2014). Dang *et al.* (2021) studied variations of three chloroplast DNA regions (*rbcL*, *matK*, *trnH-psbA*) among thirty three lotus samples and confirmed that the use of *matK*, *rbcL* and *trnH-psbA* or combine all three regions together is better discrimination within the genus Nelumbo. Erma *et al.* (2020) found that *ITS* could distinguish Sumatran Mulberry from other mulberry.

DNA barcoding for plants differs from that for animals, because in plants, it is still in the fragment research phase. Sequence alignment and manual correction are performed to remove unreliable bases at either end of the sequence

[1]Animal Husbandry Quality Standards Institute, Xinjiang Academy of Animal Sciences, Urumqi, China.
[2]College of Animal Science and Technology, Gansu Agricultural University, Lanzhou, China.

Corresponding Authors: Zunongjiang A. Bu La and Shengguo Zhao, Animal Husbandry Quality Standards Institute, Xinjiang Academy of Animal Sciences, Urumqi, China; College of Animal Science and Technology, Gansu Agricultural University, Lanzhou, China. Email: m18894036031@163.com; zhaosg@gsau.edu.cn

How to cite this article: Li, Y., La, Z.A.B., Cao, L. and Zhao, S. (2022). Establishment of a DNA Barcoding Database for Legume and Grass Species Identification. Legume Research. 45(6): 659-668. DOI: 10.18805/AF-727.

Submitted: 12-02-2022 **Accepted:** 03-03-2022 **Online:** 29-03-2022

and sequence data are analyzed with PAUP and MEGA; intra- and interspecies genetic distances are then calculated by pairwise uncorrected p-distance (Newmaster *et al*., 2008) or Kimura 2-parameter distance models (Meyer *et al*., 2005, Lahaye *et al*., 2008).

Herein we used molecular genetics to attempt and establish DNA barcode for 20 forages species. Our findings provide a theoretical basis and methodological reference for further identification of forage species.

MATERIALS AND METHODS

Sample collection

The research was conducted at Animal Husbandry Quality Standards Institute, Xinjiang Academy of Animal Sciences and research period was two years. We collected 40 samples belonging to 11 genera and 14 species of grass and 5 genera and 6 species of leguminosae forage plants. Sample information is shown in Table 1.

Primer design

Primers were designed using Primer 5.0 (http://www. bbioo. com/download/58-166-1.html) based on the nucleotide sequences of *matK* and *rbcL* in GenBank (http://www. ncbi. nlm.nih.gov). Nucleotide sequences are listed in Table 2 and primers are shown in Table 3. The primers were synthesized by Shanghai Sangon Biological Engineering Co., Ltd.

Genomic DNA extraction, PCR amplification and sequencing

Approximately 10 mg of plant leaves were ground into powder with liquid nitrogen and genomic DNA was isolated using the CTAB method (Luo, 2010). The reaction mixture for PCR included 2 µL dNTPs (2.5 mmol/L), 5 µL 10×buffer, 0.4 µL Taq DNA polymerase (5 U/µL), 2 µL DNA and 1 µL primers (10 µmol/L); ddH_2O was used to achieve a final volume of 50 µL. The cycling conditions were as follows: initial denaturation at 94°C for 2 min; 30 cycles of denaturation at 94°C for 40 s, annealing for 30s (Temperature was listed in Table 3) and elongation at 72°C for 30 s and final elongation at 72°C for 10 min. The amplicons thus obtained were stored at 4°C. They were assessed for quality using conventional 1-D gel electrophoresis (30 mm, 150 V) and sequenced by Shanghai Seiko Bio-Engineering Co., Ltd.

Data processing

Sequencing data were proofread and edited using Chromas 2.33 (http://www.seekbio.com/DownloadShow.asp?id=284). Sequence alignment was then performed using MEGA 5.0 (http://mega.software.informer.com/5.0/) and haplotype and mutation sites were analyzed with DnaSP 5.10 (http:// www.itopdog.cn/soft/4785.html).

RESULTS AND DISSCUSSION

Primer screening

According to the four pairs of universal primers used in this study, PCR conditions for target fragments were established and optimized for the 40 aforementioned grass samples. Clear target bands were visible on electrophoresis (Fig 1- Fig 4) and fragment sizes were as anticipated. Sequencing generated high-quality nucleotide sequences, indicating that the primers used in this study were suitable for DNA barcoding.

Recognition of conserved DNA regions

Using Chromas 2.33 and MEGA 5.0, 5′ - and 3′ -end conserved fragment sequences at marker loci were screened (Table 4).

MatK1 haplotype analysis

Fifty mutation sites and 12 haplotypes were found in the 40 samples (Table 5). $H1^B$, $H1^H$, $H1^J$, $H1^K$ and $H1^L$ were the characteristic haplotypes of *Zea mays*, *Avena sativa* Linn, *Medicago sativa* Linn, *Onobrychis viciaefolia* Scop and *Coronilla varia* Linn, respectively. Shared haplotypes were listed in Table 5. We herein found $H1^A$ to be a shared haplotype of leguminosae and grass forages. These results indicated that at *matK1*, some forages belonging to different families showed high homology, short genetic distance and close genetic relationship.

MatK2 haplotype analysis

Forty-five mutation sites and 17 haplotypes were found in the 40 samples (Table 6). The haplotype $H2^H$ showed the highest frequency. These findings indicated that all the six leguminosae forages showed unique haplotypes, indicating that *matK2* can be used for species identification; however, grass forages could not be completely distinguished as shared haplotypes were detected.

MatK3 haplotype analysis

Eleven mutation sites and six haplotypes were found in the 40 samples. Shared haplotypes and characteristic haplotypes were listed in Table 7. The results of the analysis indicated that these haplotypes cannot be alone used for species differentiation but can be used in combination with other haplotypes.

Rbcl haplotype analysis

Seventy-seven mutation sites and 13 haplotypes were found in the 40 samples (Table 8). $H4^A$ was a shared haplotype of *Sorghum bicolor*×*Sorghum sudanense* and *Zea mays*. $H4^B$ was the characteristic haplotype of *Stipa capillata*. $H4^C$ was a shared haplotype of *Lolium perenne* Plxie, Caddieshack, GT Fire Phoenix, Beryl, Fairway, Lark, *Triticale rimpau*, *Poa pratensis* Diamond, Barvictor, Bluebird, VN3, Rugby2, Nassau, Prize, Leopard, Snow wolf, Kentucky and Midnightall. $H4^D$ was a shared haplotype of *Festuca kansuensis*, *Festuca rubra* Bargena, Maxima, *Festuca elata* Keng ex E. Alexeev Houndog5, Roby, Barlexas, Pride, Red Elephant, *Dactylis glomerata* Linn and *Achnatherum splendens*. $H4^E$ was the characteristic haplotype of Poa forage type, $H4^F$ of *Avena sativa* Linn, $H4^G$ of *Bromus inermis* Leyss, $H4^H$ of *Medicago sativa* Linn, $H4^I$ of *Trifolium repens* Linn, $H4^J$ of *Trifolium pratense* Linn, $H4^K$ of

Table 1: Sample information.

Species	Cultivar	Genera	Families	Location
Sorghum bicolor × *Sorghum sudanense*	Unknown	*Sorghum*	Grass forage	College of
Zea mays	Unknown	*Zea*	Grass forage	Prataculture,
Stipa capillata	Wild	*Stipa*	Grass forage	Gansu Agricultural
Lolium perenne	Plxie	*Lolium*	Grass forage	University
Lolium perenne	Caddieshack	*Lolium*	Grass forage	Test Plots
Lolium perenne	GT Fire Phoenix	*Lolium*	Grass forage	
Lolium perenne	Beryl	*Lolium*	Grass forage	
Lolium perenne	Fairway	*Lolium*	Grass forage	
Lolium perenne	Lark	*Lolium*	Grass forage	
Triticale rimpau	Unknown	*Triticale*	Grass forage	
Festuca kansuensis	Wild	*Festuca*	Grass forage	
Festuca rubra	Bargena	*Festuca*	Grass forage	
Festuca rubra	Maxima	*Festuca*	Grass forage	
Festuca elata Keng ex E. Alexeev	Houndog5	*Festuca*	Grass forage	
Festuca elata Keng ex E. Alexeev	Roby	*Festuca*	Grass forage	
Festuca elata Keng ex E. Alexeev	Barlexas	*Festuca*	Grass forage	
Festuca elata Keng ex E. Alexeev	Pride	*Festuca*	Grass forage	
Festuca elata Keng ex E. Alexeev	Red Elephant	*Festuca*	Grass forage	
Poa pratensis	Diamond	*Poa*	Grass forage	
Poa pratensis	Barvictor	*Poa*	Grass forage	
Poa pratensis	Bluebird	*Poa*	Grass forage	
Poa pratensis	VN3	*Poa*	Grass forage	
Poa pratensis	Rugby2	*Poa*	Grass forage	
Poa pratensis	Nassau	*Poa*	Grass forage	
Poa pratensis	Prize	*Poa*	Grass forage	
Poa pratensis	Leopard	*Poa*	Grass forage	
Poa pratensis	Snow wolf	Poa	Grass forage	
Poa pratensis	Kentucky	*Poa*	Grass forage	
Poa pratensis	Midnight a!	*Poa*	Grass forage	
Poa forage type	Unknown	*Poa*	Grass forage	
Avena sativa Linn	Unknown	*Avena*	Grass forage	
Dactylis glomerata Linn	Unknown	*Dactylis*	Grass forage	
Bromus inermis Leyss	Unknown	*Bromus*	Grass forage	
Achnatherum splendens	Wild	*Achnatherum*	Grass forage	Gannan "sangke" grassland
Medicago sativa Linn	Unknown	*Medicago*	Leguminosae forages	College of
Trifolium repens Linn	Unknown	*Trifolium*	Leguminosae forages	Prataculture
Trifolium pratense Linn	Unknown	*Trifolium*	Leguminosae forages	Science Gansu
Onobrychis viciaefolia Scop	Unknown	*Onobrychis*	Leguminosae forages	Agricultural
Coronilla varia Linn	Unknown	*Coronilla*	Leguminosae forages	University
Vicia gigantea Bge	Unknown	*Vicia*	Leguminosae forages	Test Plots

Onobrychis viciaefolia Scop, $H4^L$ of *Coronilla varia* Linn and $H4^M$ of *Vicia gigantea* Bge.

Establishment of a DNA barcoding database

We established a DNA barcoding database for the 40 species based on the specificity of the expression of *matK1*, *matK2*, *matK3* and *rbcL*. The database consisted of three parts: Part one was specific primers for *matK1*, *matK2*, *matK3* and *rbcL* (Table 3); part two was 5′ - and 3′ -end conservative fragments at marker loci (Table 4) and part three was DNA identification code (Table 9). Sequencing data were analyzed and the haplotypes obtained using the four pairs of primers were combined to obtain a unique DNA identification code for all samples. We found that DNA barcoding could effectively distinguish between leguminosae and grass forages; different genera showed unique DNA barcoding and DNA barcoding was distinct for different species in the same genera. Further, different species of the same genera, such as the 11 species of *Poa pratense*, showed common DNA barcoding. It was verified that there

Table 2: List of reference sequences for primers.

Gene	Accession number	Species	Genera	Families
matK	AF522102	*Medicago platycarpos Linn. Trautv*	*Medicago*	Leguminosae
	AF522107	*Medicago ruthenica Linn. Trautv*		
	HM026402	*Vicia unijuga A. Br*	*Vicia*	
	AY386946	*Trifolium fragiferum Linn*	*Trifolium*	
	AF522138	*Trifolium pratense Linn*		
	AF164418	*Sorghum bicolor Linn. Moench*	*Sorghum*	Grass forage
	HF558520	*Sorghum halepense Linn. Pers*		
	DQ786940	*Festuca elata* Keng ex *E. Alexeev*	*Festuca*	
	GU367310	*Avena vaviloviana Linn*	*Avena*	
	GU367311	*Avena ventricosa Linn*		
	AB078131	*Hordeum brevisubulatum Linn*	*Hordeum*	
	AB078133	*Hordeum turkestanicum Linn*		
	AF489915	*Oryza malampuzhaensis Linn*	*Oryza*	
	AY768779	*Oryza longistaminata Linn*		
	KC129652	*Stipa lessingiana Trin. et Rupr*	*Stipa*	
	KC129653	*Stipa bungeana Trin*		
rbcL	AM234851	*Aster barbellatus Griers*	*Aster*	Asteraceae
	GQ436475	*Aster ageratoides Turcz*		
	KC589890	*Saussurea manshurica Kom*	*Saussurea*	
	U74222	*Coronilla varia Linn*	*Coronilla*	Leguminosae
	HQ644077	*Trifolium arvense Linn*	*Trifolium*	
	HM850419	*Trifolium pratense Linn*		
	JX848470	*Vicia venosa (Willd.) Maxim*	*Vicia*	
	KC700643	*Medicago sativa Linn*	*Medicago*	
	AB851490	*Medicago arabica Linn. Huds*		
	KC129652	*Stipa lessingiana Trin. et Rupr*	*Stipa*	Grass forage
	KC129653	*Stipa grandis P. Smirn*		
	HE573441	*Stipa kirghisorum P. Smirn*		
	HQ600430	*Achnatherum pekinense Hance. Ohwi*	*Achnatherum*	
	KC482943	*Festuca kryloviana Reverd*	*Festuca*	
	KC482922	*Festuca litvinovii Tzvel. E. Alexeev*		
	EU750602	*Trifolium repens Linn*	*Trifolium*	
	HQ596870	*Trifolium strepens Crantz*		
	HQ596889	*Vicia cracca Linn*	*Vicia*	
	JQ734426	*Medicago orbicularis Linn*		
	EU750098	*Trifolium pratense Linn*	*Trifolium*	
	HQ596890	*Vicia sativa Linn*	*Vicia*	
	HQ596891	*Vicia tetrasperma Linn. Schreber*		

Table 3: Optimized annealing temperature for PCR primers.

Fragment	Primer	Primer sequence	Size/bp	T_m
matK1	F-M1	5-TATACCCACTTATTTTTCGGGAGTATA-3	400	56.8
	R-M433	5-ATGGATAGGATATGGTATTCGTATATCTG-3		
matK2	F-M118	5-TTGTAAAACGTTTAATTACTCGAATGTAT-3	300	58.0
	R-M434	5-ATGGATAGGATATGGTATTCGTATATCTG-3		
matK3	F-M1262	5-TATATACTTCGGCTTTCTTGTATTAAAACTT-3	200	55.6
	R-M1472	5-CGTTTCTGAAAAGAATATCCAAATACCAAA-3		
rbcL	F-R1	5-CCAAAGATACTGATATCTTGGCAGCAT-3	450	54.4
	R-R452	5-AGACATTCATAAACAGCTCTACCGT-3		

were no differences within species, but large differences existed between species, which reached the standard of identification.

Marker site selection and primer design for leguminosae and grass forages

According to CBOL (2009), *matK*, *rpoB*, *rpoC1*, *rbcL* and *psbA-trnH* are the candidate fragments for plant DNA

Table 4: Identification of conserved regions in fragments.

Fragment	5′-Sequence	3′-Sequence	Size/bp
matK1	5′-GTAGGTT	3′-AATTCATT	256
matK2	5′-TTAATTAC	3′-CATTCCAT	219
matK3	5′-TATATACTT	3′-GATTCTTTC	131
rbcL	5′-GCAGGAGCTG	3′-ATTATCCGC	389

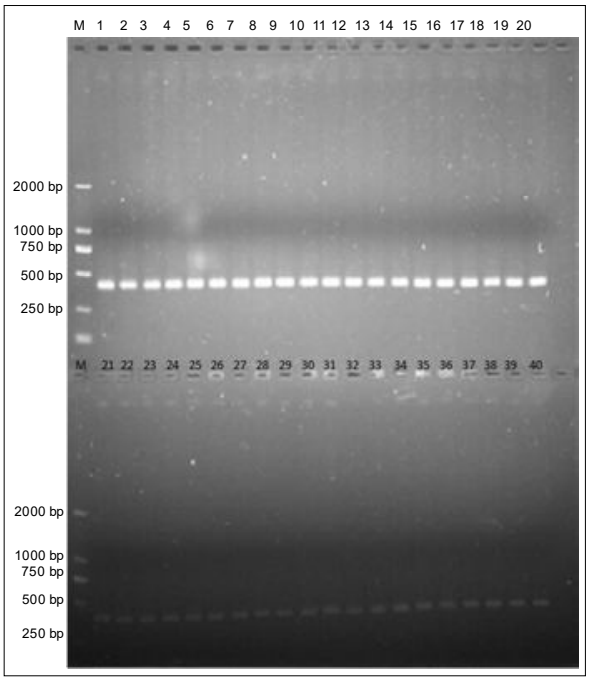

Fig 1: Amplicons generated on using primers F-M1/R-M433. Note: Lanes 1-40 correspond to sample numbers in Table 1.

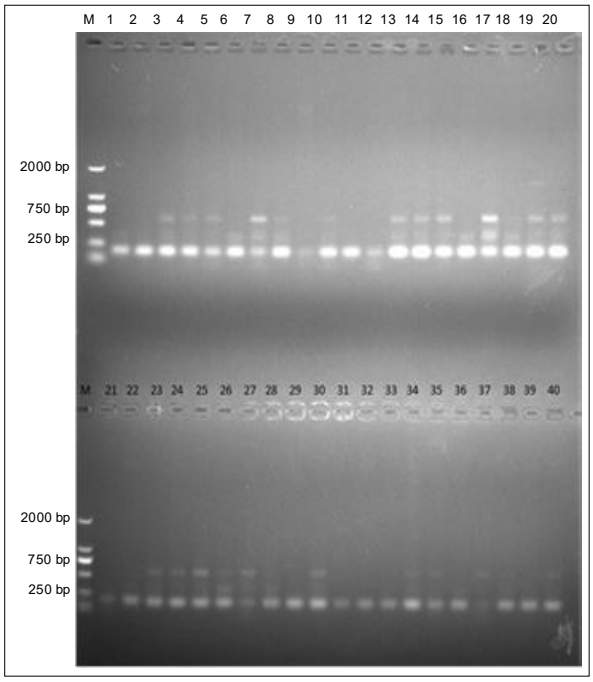

Fig 3: Amplicons generated on using primers F-M1262/R-M1472.

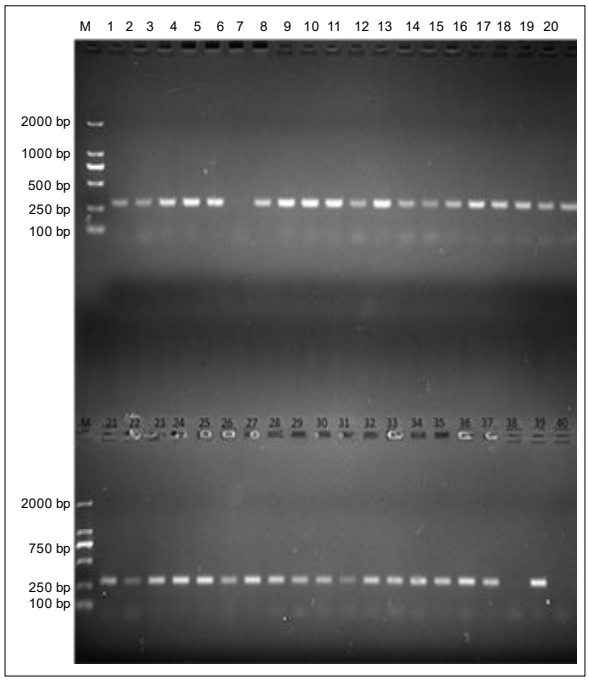

Fig 2: Amplicons generated on using primers F-M118/R-M434.

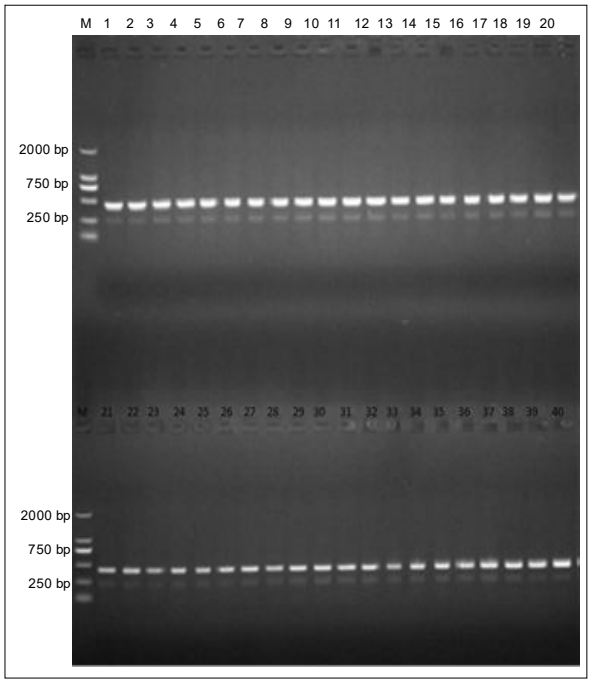

Fig 4: Amplicons generated on using primers F-R1/R-R452.

barcoding; *rbcL*, *matK* and *ITS* represent core barcodes and *psbA-trnH* represent supplementary barcodes (Pang *et al.*, 2012). *matK* reportedly shows a rapid evolution rate and is accordingly efficient at interspecies identification, but primer universality is difficult or impossible to achieve (Chase *et al.*, 2007). The amplification success rate of *matK* has been reported to be 93.5%, which is excellent, but when used alone, the species identification rate is as low as 22.2% (Fu *et al.*, 2011). *rbcL* sequences show the characteristics of high universality, easy amplification and easy alignment,

Table 5: *matK1* haplotype polymorphic sites.

Haplotypes	Mutation sites 81112345667778911111111111111111111111122222222222224582242458931001111 22335566677888999001222344444673457680323356295670340172357134 59	Frequency
H1[A]	CCGTCTTAACACACGCCGGGCAATTAGACGACTCCAAACTACAAATGAGC	3
H1[B]	A.....G.....G....CTT..G.....C.A	1
H1[C]	A.AG..G..A..CG.T....G.G.....G.G....CTT..G..C.....A	2
H1[D]	.....G.....G....GTTG.G.....A	7
H1[E]	.....G.....G....G....G.....	14
H1[F]	A.A.....GG.....G..G....CTTG.G...TG.C.A	2
H1[G]	A.A.....C.....GG.....G..G....G.TG.G...TG.CGA	5
H1[H]	..A..A.....TA.AG.....AG....CTT..G....GA...	1
H1[I]	..A.....ACGCA.A....GG.A...G..G....CTT..G.C.T..CA.	2
H1[J]	.....G.....G.....	1
H1[K]	A.A.....GG.....G....CTT..G.....G.	1
H1[L]	ATA.T..GG.GT....GT.AGG..GGT...GTCATCTT.GGA.C.C.C.A	1

Table 6: *matK2* haplotype polymorphic sites.

Haplotypes	Mutation sites 12233345556677778888899911111111111111111222202367145645 01350134601711233444455577888900001357345812914036 9279	Frequency
H2[A]	AACACACTAACCTGGGATAGATTCAGCTCTCCCATTGACAATACT	1
H2[B]	.....G.....A.....G.....A.....	1
H2[C]	.....G...T.....G.....	1
H2[D]	.....A....C.....	6
H2[E]	...G...G.....G..G.....A....C.....	1
H2[F]	.....T.....G.....A.....	1
H2[G]	.....CG...T.....G.....A....C.....	7
H2[H]	.....G.A.....	13
H2[I]	.....C.....	1
H2[J]	.....T.....G.A.....	1
H2[K]	.....T.....A.....G.....C.CG...A	1
H2[L]	.....C.....G.A.....	1
H2[M]	.....CC..G..ATAGA.GG....G.....T..A....G.AT	1
H2[N]	.....CC..G..A.....GG....G.....T..A....G.A.	1
H2[O]	.G.GT.....G.TA.....G.GT..TCAT.T.G..AC.CCAT	1
H2[P]	GG.GT.....G.TA.....G.GT..TCAT.T.G..AC.CCA.	1
H2[Q]	..A..CG...T.....AG.....G.....T.....C...A.	1

Table 7: *matK3* haplotype polymorphic sites.

Haplotypes	Mutation sites 13356778991836487843711	Frequency
H3[A]	TATTTCAAAGC	4
H3[B]	.G...CT.CCA	9
H3[C]	.G...C....A	22
H3[D]	.GCCCT.G..A	3
H3[E]	.GCC.T...AA	1
H3[F]	GGCC.T....A	1

Table 8: *rbcL* haplotype polymorphic sites.

Haplotypes	Mutation sites 233345667899111111111111111111111111112222222222222222222222222233333333333333 103657392439000011222222223333456678900011223333444555566667899001112233 4 444702581503456789458906254081373628124706726894789154736258176925892	Frequency
H4[A]	GTTTATACTCATCCCGCCGCCCAGATATCTGACAGTTCTGCATACTACTTTGCACTCCCGTTGCTATAGGGCTTTTA	2
H4[B]	C.....T..AGACA.....GC....C..A.....C....C.....T.C.T.....T..C..	1
H4[C]	C.....TG.AGACA.CGG.....C.....C.T..C.....GT..C.T....A..T.....	18
H4[D]	.....T...G.....C.....T.....	10
H4[E]	C.....G..GACA.CGG.....C....C....C.T.....GT..C.T....A..T.....	1
H4[F]	C.....TG.AGACA.CGG.....C.....C.T..C.....T....A..T.....	1
H4[G]	C.....T..AGACAGCGG.....C.....G.....T.....A.C.T....A..T.....	1
H4[H]	CC.G.C...TCC...ATGAAGAGAC.T.TGCC.TA.C..AT.GG..GT..CATG.C.GT...T..A.GAAATA.CCT	1
H4[I]	CC.G.C...TC.....TGAAGATAC.T.TGCC.TA.CT.AT.GG..GT..CATG.C.GT...T.A..GAA.TA.CCT	1
H4[J]	CC...C...TC.....TGAAGA..C.T.TGCC.TA.CT.AT.GG..GT..CATG.C.GT...T.A..GAA.....	1
H4[K]	CC.GCCG..T.C..T..GAAGA.AG.TCTGCCTTA.C..TT.GG.....C.TGAC.....T.A..GAA.T...CG	1
H4[L]	C....CGT...C.....AAGA.AG.TCTGCT.TA....AT.AGTCT.C.C.TG..TAA.G...AGG.AA.T.C.CT	1
H4[M]	ACA..C..CTCCGGA.T.AAGATAG.T.TGCC.TA.C..AT.GG..G..GCATG.CTAA...T.A..GAA.TA.CC.	1

but variations mainly exist at the level of species or above and variations at the level of species are usually not large enough (Fazekas *et al.*, 2008; Kress *et al.*, 2007; Sass *et al.*, 2007). Consequently, the use of *rbcL* is recommended in combination with one or more fragments (Fu *et al.*, 2011, Newmaster *et al.*, 2006) as *rbcL* alone cannot identify all species, but it can differentiate between many plants of the same genus (Newmaster *et al.*, 2006).

Herein we selected *matK* and *rbcL* to assess whether they are suitable for DNA barcoding. Identification of highly conserved general primers is the prerequisite for obtaining an ideal DNA barcoding sequence (Taberlet *et al.*, 2007). Highly conserved regions were screened and primers were designed for *matK1-3* and *rbcL* (Table 3). Target fragments were successfully amplified and sequenced and the results were satisfactory, facilitating the establishment of a DNA barcoding database. PCR amplification is pivotal for DNA barcoding and annealing temperature markedly affects PCR amplification. Within the range of T_m value, choosing a higher denaturation temperature can relatively reduce non-specific binding between primers and templates, improving the specificity of PCR amplification (Li *et al.*, 2014). The primers used herein showed high amplification efficiency, high success rate and suitable reaction conditions, which laid the foundation for species identification.

Identification of common forages using haplotype combinations

The introduction of DNA barcoding resolved the many problems associated with species identification (Vilaça *et al.*, 2013, Moritz *et al.*, 2004). However, according to few studies, many species still cannot be identified by DNA barcoding, which could be due to various factors, such as plant species polyploidization, hybridization and radiation evolution (Percy *et al.*, 2014, Collins *et al.*, 2013). These limitations can be overcome by using a combination of haplotypes.

matK is widely used for DNA barcoding in plants, particularly medicinal plants (Selvaraj *et al.*, 2008, Will *et al.*, 2004). *matK* has been proven to be the dominant candidate sequence in Zingiberaceae plants, but it has not been widely used for DNA barcoding in leguminosae and grass species. In this study, species identification efficiency using *matK* could not reach 100% when four marker sites were used alone and specific bands corresponding to *matK4* were not visible on agarose gel electrophoresis. We believe that this could be attributed to the designed primer amplifying a sequence showing a low conserved pattern or the primer not meeting the principles of universal primer design, leading to the amplification efficiency being 0%. Furthermore, identification efficiency associated with *matK* was relatively high and its sequence showed the unique poly-A structure, leading to hurdles in amplification and sequencing. *ITS* sequences behaved in a similar manner. Therefore, structures such as poly-A should be avoided when designing general primers (Hollingsworth *et al.*, 2009).

There were many variation sites in *rbcL* fragments, but interspecies variation rate was low, which is consistent with the results of previous studies (Hollingsworth *et al.*, 2009, Drumwright *et al.*, 2014). Haplotype analysis of *rbcL* amplified fragments showed that shared haplotypes existed in the 15 genera of Leguminaceae and Gramineae; for example, *Sorghum* and *Zea* showed shared haplotypes, as with *Lolium*, *Triticale* and *Poa* and also *Festuca*, *Dactylis* and *Achnatherum*. However, other genera showed characteristic haplotypes. The identification success rate on using *rbcL* for leguminosae and grass species was 53.3%. We also assessed haplotype results among different

Table 9: DNA barcoding database for the 40 samples analyzed in this study.

Plant species		Variance of base pair				DNA barcoding
		matK1	*matK2*	*matK3*	*rbcL*	
Sorghum bicolor × *Sorghum sudanense*		$H1^{A}$	$H2^{A}$	$H3^{A}$	$H4^{A}$	$MRH1^{A}H2^{A}$ $H3^{A}$ $H4^{A}$
Zea mays		$H1^{B}$	$H2^{B}$	$H3^{A}$	$H4^{A}$	$MRH1^{B}H2^{B}$ $H3^{A}$ $H4^{A}$
Stipa capillata		$H1^{C}$	$H2^{C}$	$H3^{B}$	$H4^{B}$	$MRH1^{C}H2^{C}$ $H3^{B}$ $H4^{B}$
Lolium perenne	Plxie	$H1^{D}$	$H2^{D}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{D}$ $H3^{B}H4^{C}$
	Caddieshack	$H1^{D}$	$H2^{D}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{D}$ $H3^{B}H4^{C}$
	GT Fire Phoenix	$H1^{D}$	$H2^{D}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{D}$ $H3^{B}H4^{C}$
	Beryl	$H1^{D}$	$H2^{D}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{D}$ $H3^{B}H4^{C}$
	Fairway	$H1^{D}$	$H2^{D}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{D}$ $H3^{B}H4^{C}$
	Lark	$H1^{D}$	$H2^{D}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{D}$ $H3^{B}H4^{C}$
Triticale rimpau		$H1^{D}$	$H2^{E}$	$H3^{B}$	$H4^{C}$	$MRH1^{D}H2^{E}$ $H3^{B}$ $H4^{C}$
Festuca kansuensis		$H1^{E}$	$H2^{F}$	$H3^{C}$	$H4^{D}$	$MRH1^{E}H2^{F}H3^{C}H4^{D}$
Festuca rubra	Bargena	$H1^{F}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{F}H2^{G}H3^{C}H4^{D}$
	Maxima	$H1^{F}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{F}H2^{G}H3^{C}H4^{D}$
Festuca elata Keng	Houndog5	$H1^{G}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{G}H2^{G}H3^{C}H4^{D}$
ex E. Alexeev	Roby	$H1^{G}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{G}H2^{G}H3^{C}H4^{D}$
	Barlexas	$H1^{G}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{G}H2^{G}H3^{C}H4^{D}$
	Pride	$H1^{G}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{G}H2^{G}H3^{C}H4^{D}$
	Red Elephant	$H1^{G}$	$H2^{G}$	$H3^{C}$	$H4^{D}$	$MRH1^{G}H2^{G}H3^{C}H4^{D}$
Poa pratensis	Diamond	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Barvictor	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Bluebird	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	VN3	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Rugby2	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Nassau	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Prize	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Leopard	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Snow wolf	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Kentucky	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
	Midnight II	$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{C}$	$MRH1^{E}H2^{H}H3^{C}H4^{C}$
Poa forage type		$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{E}$	$MRH1^{E}H2^{H}H3^{C}H4^{E}$
Avena sativa Linn		$H1^{H}$	$H2^{I}$	$H3^{D}$	$H4^{F}$	$MRH1^{H}H2^{I}H3^{D}H4^{F}$
Dactylis glomerata Linn		$H1^{I}$	$H2^{J}$	$H3^{D}$	$H4^{D}$	$MRH1^{I}H2^{J}H3^{D}H4^{D}$
Bromus inermis Leyss		$H1^{I}$	$H2^{K}$	$H3^{D}$	$H4^{G}$	$MRH1^{I}H2^{K}H3^{D}H4^{G}$
Achnatherum splendens		$H1^{E}$	$H2^{H}$	$H3^{C}$	$H4^{D}$	$MRH1^{E}H2^{H}H3^{C}H4^{D}$
Medicago sativa Linn		$H1^{J}$	$H2^{L}$	$H3^{C}$	$H4^{H}$	$MRH1^{J}H2^{L}H3^{C}H4^{H}$
Trifolium repens Linn		$H1^{A}$	$H2^{M}$	$H3^{A}$	$H4^{I}$	$MRH1^{A}H2^{M}H3^{A}H4^{I}$
Trifolium pretense Linn		$H1^{A}$	$H2^{N}$	$H3^{A}$	$H4^{J}$	$MRH1^{A}H2^{N}H3^{A}H4^{J}$
Onobrychis viciaefolia Scop		$H1^{K}$	$H2^{O}$	$H3^{E}$	$H4^{K}$	$MRH1^{K}H2^{O}H3^{E}H4^{K}$
Coronilla varia Linn		$H1^{L}$	$H2^{P}$	$H3^{F}$	$H4^{L}$	$MRH1^{L}H2^{P}H3^{F}H4^{L}$
Vicia gigantea Bge		$H1^{C}$	$H2^{Q}$	$H3^{B}$	$H4^{M}$	$MRH1^{C}H2^{Q}H3^{B}H4^{M}$

Note: "M" = *matK*, "R" = *rbcL*, "H" = haplotype, "1, 2 and 3" represent *matK1*, *matK2* and *matK3*, respectively and "4" represents *rbcL*. Superscripted letters indicate nucleotide haplotypes for marked fragments.

species in the same genus: *Lolium perenne* and *Triticale rimpau* showed shared haplotypes and *Festuca kansuensis*, *Festuca rubra* and *Festuca elata* Keng ex E. also showed shared haplotypes. The identification success rate on using *rbcL* for different species in the same genus was 75%. Different varieties of the same forage share the same haplotype, interspecies cannot be distinguished.

CONCLUSION

The DNA barcode database established in this study showed that each species had its own unique DNA barcode that was distinguishable from each other. Moreover, DNA barcodes for different species of the same herbage were identical. These factors met the conditions for the establishment of the DNA barcode database, ensuring no differences within

species and allowing large differences between species. Our findings indicate that *matK-rbcL* can be used as a common sequence combination for DNA barcoding in legumes and grasses; the combination can be used to identify legumes and grasses at the species level and above, while intraspecific identification cannot be achieved.

ACKNOWLEDGEMENT

Research was supported by Natural Science Foundation of Xinjiang Uygur Autonomous Region (2021D01B85).

Conflict of interest: None.

REFERENCES

Cai, Y.M., Dai, J.P., Zheng Y.X., Ren, Y.Y., Chen, H.M., Feng, T.T., Gao, X.X., Zhu, S. (2021). Screening of DNA barcoding sequences for molecular identification of Genus Uncaria. Chinese Traditional and Herbal Drugs. 1-10.

CBOL Plant Working Group. (2009). A DNA barcode for land plants. PNAS. 106(31): 12794-12797.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., Cassio, V.D.B., Madrinan, S., Petersen, G., Seberg, O., Jorgsensen, T., Cameron, K.M., Carine, M. (2007). A proposal for a standardised protocol to barcode all land plants. Taxon. 56: 295-299.

China Plant BOL Group. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be in-corporated into the core barcode for seed plants. PNAS. 108: 19641-19646.

Collins, R.A. and Cruickshank, R.H. (2013). The seven deadly sins of DNA barcoding. Molecular Ecology Resources. 13(6): 969-975.

Costion, C.M., Edwards, W., Ford, A.J., Metcalfe, D.J., Cross, H.B., Harrington, M.G., Richardson, J.E., Hilbert, D.W., Lowe, A.J., Crayn, D.M. (2015). Using phylogenetic diversity to identify ancient rain forest refugia and diversification zones in a biodiversity hotspot. Diversity and Distributions. 21: 279-289.

Dang, T.L., Hoang, T.K.H., Le, L.T.T., Nguyen, T.Q.T. (2021). Evaluation of Genetic Diversity by DNA Barcoding of Local Lotus Populations from Thua Thien Hue Province. Indian Journal of Agricultural Research. 55(2): 121-128.

Drumwright, A.M., Allen, B.W., Huff, K.A., *et al.* (2014). Survey and DNA Barcoding of Poaceae in Flat Rock Cedar Glades and Barrens State Natural Area, Murfreesboro, Tennessee. Castanea. 76(3): 300-310.

Erma, N., Nindi, A., Syamsuardi, N., Fitmawati, F. (2020). Clarification of Sumatran Mulberry (*Morus macroura* var. *macroura*, Moraceae) from West Sumatra, Indonesia using Nucleus Ribosomal *ITS* (*Internal Transcribed Spacer*) Gene. Indian Journal of Agricultural Research. 54(5): 635-640.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Perey, D.M., Hajibabaei, Barrett, S.C.H. (2008). Multiple multiloeus DNA barcodes from the plastid genome discriminate plants species equally well. PloS One. 3: e2802.

Fu, Y.M., Jiang, W.M., Fu, C.X. (2011). Identification of species within Tetrastigma (Miq.) Planch (Vitaceae) based on DNA barcoding techniques. Journal of Systematics and Evolution. 49(3): 237-245.

Hebert, P.D.N, Cywinska, A., Ball, S.L., de Waard, J.R. (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society of London Series B: Biological Sciences. 270: 313-321.

Hollingsworth, M.L. andra, C.A., Forrest, L.L. (2009). Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divegent groups of land plants. Molecular Ecology Resources. 9: 439-457.

Hollingsworth, P.M., Graham, S.W., Little, D.P. (2011). Choosing and using a plant DNA barcode. PloS One. 6: el9254.

Kress, W.J., Eriekson, D.L. (2007). A two locus global DNA barcode form land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region. PloS One. 2: e508.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., Janzen, D.H. (2005). Use of DNA barcodes to identify flowering plants. PNAS. 102: 8369-8374.

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G., Vincent, S. (2008). DNA barcoding the floras of biodiversity hot-spots. PNAS. 105: 2923-2928.

Li, Y., Wu, S.M., Chen, Q. (2014). Establishment of DNA barcode of common species of Lolium. Plant Quarantine. 28(6): 376-385.

Liu, G.S. (2003). Advances of molecular biology and biotechnology used in gramineal forage species. Acta Botanica Boreali-Occidentalia Sinica. 23(4): 682-687.

Luo, K. (2010). Assessment for universal plant DNA barcodes based on species of rutaceae and araceae family. PhD Thesis. Wuhan: Hubei University of Traditional Chinese Medicine. (in Chinese).

Meyer, C.P., Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. PloS Biology. 3(12): 2229-2238.

Moritz, C., Cicero, C. (2004). DNA barcoding: Promise and pitfalls. Plos Biology. 2(10): e354.

Newmaster, S.G., Fazekas, A.J., Ragupathy, S. (2006). DNA barcoding in land plants: Evaluation of rbcL in a multi genetiered approach. Canadian Journal of Botany. 84: 335-341.

Newmaster, S.G., Fazekas, A.J., Sleeves, R.A.D., Janovec, J. (2008). Testing candidate plant barcode regions in the Myristica ceae. Molecular Ecology Resources. 8: 480-490.

Pang, X.H., Liu, C., Shi, L.C., Liu, R., Liang, D., Li, H., Cherny, S.S., Chen, S.L. (2012). Utility of the trnH-psbA intergenic spacer region and ITS combinations as plant DNA barcodes: A meta analysis. PloS One. 7: e48833.

Percy, D.M., Argus, G.W., Cronk, Q.C., *et al.* (2014). Understanding the spectacular failure of DNA barcoding in willows (Salix): does this result from a trans-specific selective sweep? Molecular Ecology. 23(19): 4737-4756.

Rashmi, K.V., Sathyanarayana, N., Vidya, S.M. (2020). Variations in the *trnHpsbA* region of *Mucuna pruriens* L. (DC.) varieties of India: An insight on intraspecific diversity. Indian Journal of Agricultural Research. 53(3): 284-290.

Sass, C., Little, D.P., Stevenson, D.W., Specht, C.D. (2007). DNA barcoding in thecycadales: Testing the potential of proposed barcoding markers for species identification of cycads. PloS One. 2: el 154.

Selvaraj, D., Sarma, R.K., Sathishkumar, R. (2008). Phylogenetic analysis of chloroplast matK gene from Zingiberaceae for plant DNA barcoding. Bioinformation. 3(1): 24-27.

Taberlet, P., Coissac, E., Pompanon, F. (2007). Power and limitations of the chlorop last trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Research. 35(3): el4.

Vilaça, S.T., Lacerda, D.R., Sari, E.H.R., *et al.* (2013). DNA-based identification applied to Thamnophilidae (Passeriformes) species: The first barcodes of Neotropical birds. Revista Brasileira De Ornitologia. 14(1): 7-14.

Wan, J.J., Yu, L., Lu, W.H., Yang, G.L., Zhang, Q.B., Yang, J.B. (2014). Comprehensive evaluation of nutritive value of dominant gramineous grass in Shaertao Mountain, Zhaosu County in Xinjiang. Pratacultural Science. 31(11): 2141-2147.

Wang, X.Y., Chen, X.C., Liao, B.S., Wang, L.L., Han, J.P. (2014). Identification of Amomi Fructus Rotundus based on DNA barcoding. Abstract of papers on the 14th National Symposium on Traditional Chinese Medicine and Natural Medicine. BeiJing: Chinese Pharmaceutical Association, 17.

Will, K.W., Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics-the International Journal of the Willi Hennig Society. 20(1): 47-55.

Yao, H., Song, J.Y., Liu, C., Luo, K., Han, J. P., Li, Y., Pang, X.H., Xu, H.X., Zhu, Y.J., Xiao, P.G., Chen, S.L. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS One. 5: el3102.