Received: 11 February 2019 | Revised: 28 July 2019 | Accepted: 30 July 2019

DOI: 10.1002/ece3.5590

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

The impact of genetic diversity on the accuracy of DNA barcoding to identify species: A study on the genus *Phellodendron*

Zhi-peng Zhang[1] | Xiao-yue Wang[1] | Zhao Zhang[1] | Hui Yao[1] | Xiao-mei Zhang[2] | Yang Zhang[1] | Ben-gang Zhang[1]

[1]Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicine, Ministry of Education/Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China

[2]China-ASEAN Traditional Medicine Cooperation and Exchange Center, Guangxi Botanical Garden of Medicinal Plants, Nanning, China

Correspondence
Zhao Zhang, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China.
Email: zhangzhao1962@tom.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 81473305; National Science and Technology Foundational Special Project, Grant/Award Number: 2015FY111500

Abstract

DNA barcoding is widely used in species identification, but there is considerable controversy regarding the extent of sampling in research methods. Some scholars have proposed that this small sample size underestimates the intraspecific genetic diversity, which would impact on the accuracy of DNA barcoding to identify species. In study, we selected all *Phellodendron* species (including *P. amurense* Rupr., *P. chinense* Schneid., and *P. chinense* var. *glabriusculum* Schneid.) as the materials, collected 59 *P. amurense* samples from 35 populations greatly to represent the genetic diversity, and analyzed the haplotype, genetic distance, barcoding gap, and Neighbor-Joining (NJ) trees based on *psb*A-*trn*H and internal transcribed spacer gene sequences. Additionally, a sampling simulation was conducted to assess the correlation between genetic diversity and the number of populations. Finally, analysis of critical geographical populations was performed. Based on analysis of haplotype, genetic distance, barcoding gap, and NJ trees, we found that eight *P. amurense* samples impacted on the effectiveness of DNA barcoding, which genetic information were very important to identify *Phellodendron* species. Moreover, the result of the NJ tree analysis performed the small-scale *P. amurense* sample size did not completely match the objective phylogenetic relationship in *Phellodendron*. In simulation sampling analysis, the data showed the genetic diversity indexes at the same population level gradually decreased and stabilized as the number of simulation sampling populations increased. We found that 1–2 samples from over 24 populations based on uniform geographical distribution could represent 80% of the genetic diversity of *P. amurense* and ensure authenticity and reliability of DNA barcoding. Thus, we proposed it is particularly important adequately samples to cover infraspecific genetic diversity in order to ensure identification accuracy of DNA barcoding.

KEYWORDS
accuracy, DNA barcoding, genetic diversity, identify species, *Phellodendron*

1 | INTRODUCTION

DNA barcoding, which is based on one or more common and standard DNA sequences for species identification and characterization techniques, is widely used in the survey and inventory of biodiversity, species identification, and discovery of new species (Hebert, Cywinska, Ball, & Dewaard, 2003; HollingsworthGraham & Little, 2011; Kress, Wurdack, Zimmer, Weigt, & Janzen, 2005). Cytochrome oxidase I gene (*COI* gene) sequences from the mitochondrial genome can be used for the identification of multiple animal groups, making it an ideal animal barcode sequence (Tavares & Baker, 2008; Ward, Zemlak, Innes, Last, & Hebert, 2005). In terms of plants, no one barcode can be a universal barcode similar to the animal *COI* gene. The Consortium for the Barcode of Life used the ribulose bisphosphate carboxylase large chain (*rbc*L) and maturase K (*mat*K) genes as core DNA barcodes for seed plants, as well as the *psb*A-*trn*H intergenic spacer (*psb*A-*trn*H) and internal transcribed spacer (ITS) genes as supplementary DNA barcodes (CBOL Plant Working Group, 2009). Since Hebert proposed the concept of DNA barcodes in 2003 (Hebert et al., 2003), many scholars have performed studies to verify the feasibility of identifying plant species through DNA barcoding. The China Plant BOL Group (2011) assessed *rbc*L, *mat*K, *psb*A-*trn*H, and ITS candidate barcodes with 75 families, 141 genuses, 1,757 species, and 6,286 samples from seed plants. Many studies have further confirmed the effectiveness of DNA barcoding for species identification. For example, Yao, Song, and Ma (2009) reported that the *psb*A-*trn*H sequence can be used to identify *Dendrobium* species. Liu, Zhang, et al. (2012) showed a 100% identification using the *psb*A-*trn*H sequence in 38 species from the genus *Rhododendron*. Li, Chen, Wang, and Xiong (2012) discovered that the efficiency of the ITS sequence is 70% when examining 63 species in the genus *Ficus*. Yang, Wang, Möller, Gao, and Wu (2012) used ITS + *psb*A-*trn*H sequences to identify 30 species from the genus *Parnassia* with 90% efficiency.

However, the genetic diversity of each species was neglected by almost all DNA barcoding studies, which reflected in sampling strategies. The International Barcode of Life Project calls for at least 10 samples for each species (Gao et al., 2012). To obtain more species sequences, one must decrease the number samples for each species when there is limited funding (Matz & Nielsen, 2005). Currently, most species have only 5–10 sequences in the established DNA barcode database, though several species have only 1–2 sequences (http://www.barcodinglife.org/views/ligin.php), which is far from sufficient to speculate the number of samples needed to establish a database to represent the genetic diversity of each species. The efficiency and accuracy of DNA barcoding depends on the degree of sampling per species, because a large enough sample size is needed to provide a reliable estimate of genetic polymorphism and for delimiting species (Luo et al., 2015). Incomplete sample surveys, errors in sample identification and weak taxonomic are obstacles (Young, McKelvey, Pilgrim, & Schwartz, 2013). A small range and unevenness in sampling could cause differences in thresholding between intraspecific and interspecific variations (Meyer & Paulay, 2005). We believed that mostly representation of genetic variation in samples had defects, as many studies only analyze and collect 1 or 2 samples from one species. Therefore, we chose *Phellodendron* genus as a case to discuss genetic diversity impacting on the accuracy of DNA barcoding to identify species.

Highlights

- Used a case to prove the genetic diversity impacting on the identification accuracy of DNA barcoding.
- Analyzed the relationship between samples size and genetic diversity parameters of *P. amurense* by simulation sampling.
- Proposed adequately samples covering infraspecific variation of species being the key to DNA barcoding.

The *Phellodendron* genus (Rutaceae) has two species and one variant, including *P. amurense* Rupr., *P. chinense* Schneid., and *P. chinense* var. *glabriusculum* Schneid., are distributed in China (Huang, 1997). As tertiary paleotropical flora relict plants, *Phellodendron* has scientific value for studying ancient flora, paleogeography, and quaternary glacial climate (Huang, 1958). A large number of wild populations drastically reduced, especially since the late 19th century, because its cortex is a kind of precious Chinese traditional medicine, and its wood is widely used for its hard texture and beautiful grain and color (Qin, Wang, & Yan, 2006). The endangered plant *P. amurense* is distributed in northeastern China, and *P. chinense* is distributed in southwestern China (Huang, 1997; State Bureau of Environmental Protection of China & Institute of Botany Chinese Academy of Sciences, 1987); the gap in their distribution area is approximately 1,000 km. These genetic studies mainly focused on intraspecific genetic diversity, the analyses of intraspecies and intergenus have not been reported (Wang, Bao, Wang, & Ge, 2014; Yan, Zhang, Zhang, & Yu, 2006; Yu et al., 2013). *Phellodendron* genus does not exist controversy in plant taxonomy and have the geographical isolation of species. These studies mainly focused on intraspecific genetic diversity and concerned the species in China; however, analyses of intraspecies and intergenus have not been reported in the genus *Phellodendron*. Thus, it is an ideal material. To assess the impact of genetic diversity on DNA barcoding, we selected the *Phellodendron* genus as a model and collected numerous samples to represent intraspecific diversity in study.

Also, the data of previous study showed that there was no variation of *rbc*L, *mat*K among individuals of *Phellodendron*, *psb*A-*trn*H and ITS were polymorphism in *Phellodendron*. Thus, we assessed the accuracy of DNA barcoding to determine *Phellodendron* species by ITS and *psb*A-*trn*H.

2 | MATERIALS AND METHODS

2.1 | Plant materials

We collected 1 or 2 samples from each population, which based on a viewpoint that DNA barcoding variations within a population are usually less than that between populations (Liu, Provan, Gao, & Li, 2012). In total, 59 *P. amurense* cortex samples were densely collected from 35 populations

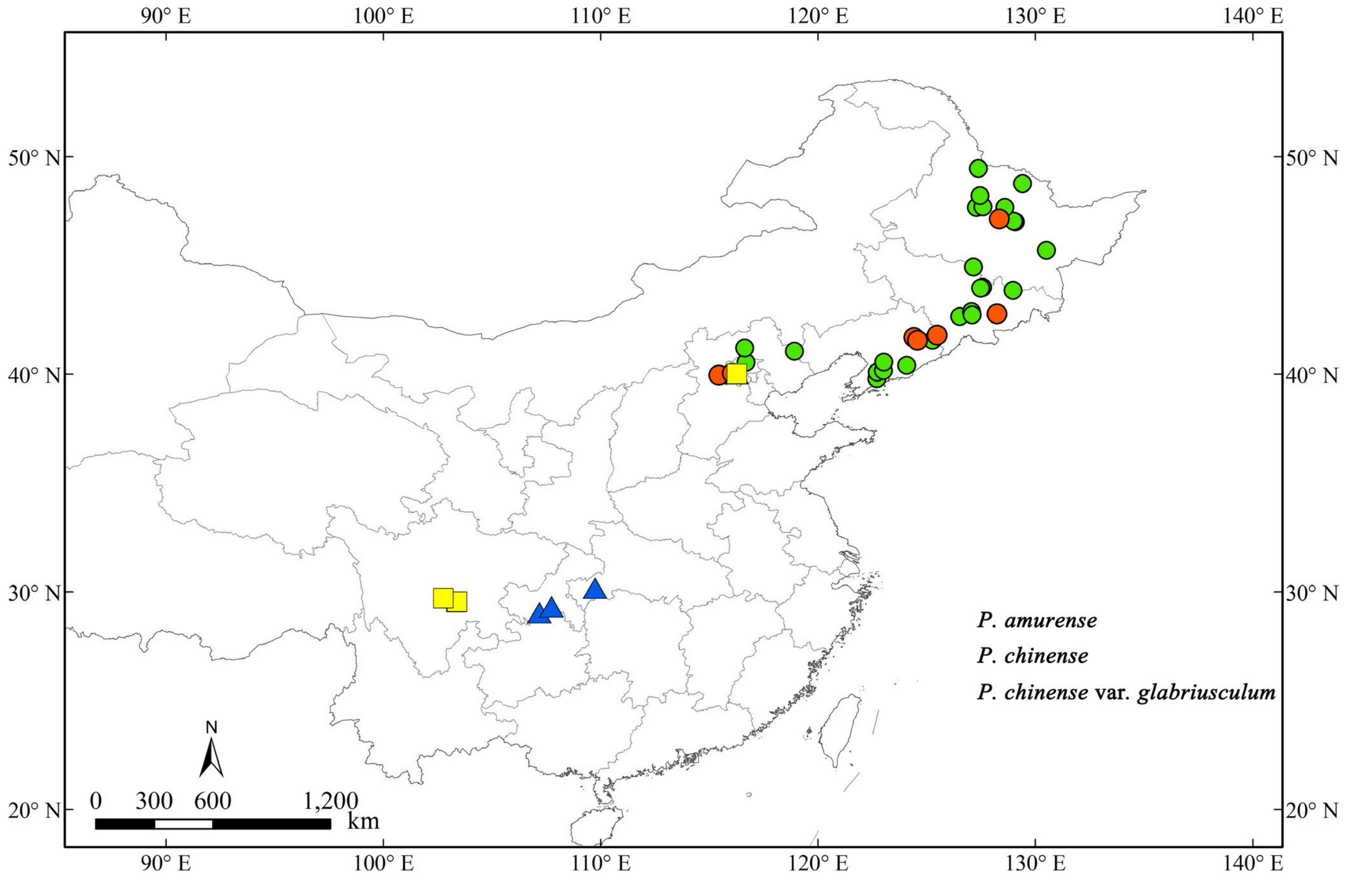

FIGURE 1 Geographical distribution of 41 sample populations (red circles mark critical populations)

throughout the distribution area, which enabled us to have a large size and representative samples to ensure the credibility of this study.

The survey found that *P. chinense* and *P. chinense* var. *glabriusculum* populations included types that were both were rare and wild, as well as cultivated; thus, we collected one wild population from *P. chinense* and *P. chinense* var. *glabriusculum*. Fourteen *P. chinense* and *P. chinense* var. *glabriusculum* samples were collected from eight populations to ensure the samples represent the entire distribution area. A total of 41 populations were surveyed and are shown in Figure 1.

Additionally, four samples of *Tetradium ruticarpum* were collected as an outgroup for the molecular analysis. The significant differences in the geographical distribution between *P. amurense* and other *Phellodendron* species would completely guarantee the accuracy of the classification. Although five samples from three plants were collected from a botanical garden, detailed records of their original location provided evidence to ensure their identity. The botanical identities of all the samples were confirmed by Professor Zhao Zhang. All corresponding voucher samples were deposited in the Herbarium of the Institute of Medicinal Plant Development.

2.2 | DNA extraction and amplification

All samples (40 mg) were rubbed for 2 min at a frequency of 30 r/s. Total genomic DNA was extracted using the Plant Genomic DNA Kit (Tiangen Biotech Co.) according to the manufacturer's instructions. The extracted genomic DNA was amplified by polymerase chain reaction (PCR) using the ITS (ITS5F and ITS4R) and *psbA-trnH* (fwdPA and revTH) primers (Chen, 2012). The PCR mixtures and conditions were described by Chen (2012). PCR products were separated and detected by 1.5% agarose gel electrophoresis. Purified products were sequenced in both directions using the PCR primers on a 3730XL sequencer (Applied Biosystems).

2.3 | Statistical analysis

Sequences were assembled and aligned with the CodonCode Aligner 3.7.1 (CodonCode Co.) as well as the base quality was evaluated to avoid technical error. The inter/intraspecific genetic distances and barcoding gap were analyzed with the *P* Language based on Kimura's 2-parameter theory (Tamura et al., 2011). The variable sites and the bootstrap Neighbor-Joining (NJ) tree were conducted with MEGA (4.0 version) according to Kimura's 2-parameter method with 1,000 replicate bootstrap testing (Tamura et al., 2011).

To assess the relationship between a number of populations and the genetic diversity of *P. amurense*, a sampling simulation was conducted in this study. In total, 3, 6, 9, 12, 15, 18, 23, 28, and 33 populations levels were randomly selected from 35 *P. amurense* population samples by a computer sampling simulation, with each population sampling being repeated 50 times. The number of haplotypes (*H*) was used as an important indicator of genetic diversity in studies

TABLE 1 GenBank accession No. for the samples in this study

			GenBank accession No.	
Codes	Scientific name	Family	ITS	*psb*A-*trn*H
A1-A59	*P. amurense*	Rutaceae	MK419239–MK419297	MK419162–MK41922
B1-B6	*P. chinense*	Rutaceae	MK419298–MK419303	MK419221–MK419226
C1-C8	*P. chinense* var. *glabriusculum*	Rutaceae	MK419304–MK419311	MK419227–MK419234
D1-D4	*Tetradium ruticarpum*	Rutaceae	MK419312–MK419315	MK419235–MK419238

TABLE 2 ITS sequence haplotypes and variation sites in *Phellodendron* species

				Sites/bp											
				1	1	1	1	2	3	4	4	4	4	5	6
				5	7	7	9	0	9	1	3	4	8	9	2
Scientific name	Haplotype	Number of samples	Percentage	7	3	7	0	8	1	9	5	3	6	9	3
P. amurense	A1	50	84.7	C	C	G	G	C	C	C	C	C	C	G	A
	A2	3	5.1	.	.	.	.	.	T	.	.	.	.	.	.
	A3	1	1.7	.	.	.	A	.	.	.	.	.	.	.	.
	A4	1	1.7	.	.	.	.	.	.	G	.	.	.	.	.
	A5	1	1.7	.	.	.	.	.	.	.	.	.	.	.	C
	A6	3	5.1	.	.	A	.	.	.	.	.	.	.	.	.
P. chinense	A7	2	33.3	.	T	.	.	.	.	.	.	.	.	A	.
	A8	2	33.3	.	T	.	.	.	.	.	.	.	.	.	.
	A9	1	16.7	.	T	.	.	.	.	.	.	.	T	.	.
	A10	1	16.7	T	T	.	.	.	.	.	.	.	.	.	.
P. chinense var. *glabriusculum*	A11	6	75	.	.	.	.	T	.	.	.	.	.	.	.
	A12	1	12.5	.	.	.	.	.	.	.	.	A	.	.	.
	A13	1	12.5	.	.	.	.	.	.	.	T	A	.	.	.

(.) indicated the same base as the first row.

of populations calculated by the MEGA (4.0 version) (Tamura et al., 2011) and DnaSP V5.0 (Librado & Rozas, 2009) software. The haplotype diversity (H_d) and nucleotide diversity (P_i) were also studied at the same time. The haplotype diversity (H_d), the frequencies of two different haplotypes randomly extracted in the sample, is an important indicator of the degree of variation in a population. The nucleotide diversity (P_i), the average of all pairwise distances in a sample, is commonly used to estimate genetic polymorphism. The relationship of H with a population's number in the sampling simulation was analyzed by nonlinear regression (curve fit) with GraphPad Prism 5 (https://www.graphpad.com/), while the relationships of numbers of populations with H_d and P_i were analyzed with a scatter plot.

3 | RESULTS

3.1 | Haplotype analysis

The GenBank accession No. for the ITS and *psb*A-*trn*H contig sequences from all the samples in this study are shown in Table 1. The average base quality value (QV) of forward sequence or reverse sequence was ≥30, and the coincidence ratio of forward and reverse sequences was 100%. The haplotype and variable sites in the ITS sequence are shown in Table 2. The quality value (QV) of variable sites was verified as ≥30 by traceability. The data revealed that *P. amurense* had 6 haplotypes, *P. chinense* had four haplotypes, and *P. chinense* var. *glabriusculum* had three haplotypes. *P. chinense* was different from *P. amurense* and *P. chinense* var. *glabriusculum* at bp 173 with a T; simultaneously *P. chinense* var. *glabriusculum* was also different from *P. amurense* and *P. chinense* at bp 208 with a T, except in haplotypes A12 and A13.

The haplotype and variable sites in the *psb*A-*trn*H sequence (Table 3), which the quality value (QV) of variable sites, were verified as ≥30 by traceability, demonstrated that *P. chinense* and *P. chinense* var. *glabriusculum* had one haplotype (B10) in common, and that *P. amurense* had nine haplotypes, with 54.1% of the samples being haplotype B1. Based on the variable sites, haplotypes B2 and B3 in *P. amurense* were the same as haplotype B10 with 379, 380, 381, and 382 bp.

TABLE 3 *psb*A-*trn*H sequence haplotypes and variation sites in *Phellodendron* species

				Sites/ bp										
Scientific name	Haplotype	Number of samples	Percentage	2	3	14	59	206	379	380	381	382	397	445
P. amurense	B1	32	54.1	G	C	C	A	T	T	T	G	C	A	T
	B2	5	8.5	.	.	.	.	.	G	C	A	A	.	.
	B3	3	5.1	.	.	T	.	.	G	C	A	A	.	.
	B4	13	22.1	.	.	T	.	.	.	.	.	.	.	.
	B5	1	1.7	C	.	T	.	.	.	.	.	.	.	.
	B6	2	3.4	C	.	.	C	.	.	.	.	.	.	.
	B7	1	1.7	.	A	.	.	.	.	.	.	.	.	.
	B8	1	1.7	-	G	T	.	.	.	.	.	.	.	.
	B9	1	1.7	.	.	.	.	.	.	.	.	.	.	C
P. chinense	B10	6	100	.	.	.	.	G	G	C	A	A	G	.
P. chinense var. *glabriusculum*	B10	8	100	.	.	.	.	G	G	C	A	A	G	.

(.) indicated the same base as the first row.

TABLE 4 Analysis of the intergenus-specific divergence and intraspecific variation with the two barcodes

Markers	ITS	*psb*A-*trn*H
Theta(avg_intra_avg)	0.0018 ± 0.0020	0.0018 ± 0.0026
coalescent depth(avg_intra_max)	0.0048 ± 0.0023	0.0081 ± 0.0115
All intraspecific distance(avg_between_intraspecies)	0.0005 ± 0.0010	0.0035 ± 0.0041
Theta prime(avg_interbyG_avg)	0.0024 ± 0.0010	0.0137 ± 0.0041
minimum interspecific distance(avg_interbyG_min)	0.0016 ± 0.0010	0.0046 ± 0.0041
all interspecific distance	0.0024 ± 0.0009	0.0137 ± 0.0036

3.2 | Genetic distance and barcoding gap analysis

Six parameters were used to analyze intraspecific variation and interspecific divergence with two barcodes (Table 4). In this instance, the maximum intraspecific distance was higher than the minimum interspecific distance for two barcodes, which indicated that the two barcodes did not perform well in the discrimination of *Phellodendron* species.

The barcoding gap presents the remarkable variation between inter- and intraspecies and demonstrates the separate or overlapping distributions between intra- and interspecific samples. In this study (Figure 2), the ITS sequence did not exhibit gaps in the intra- and interspecific variation distributions. In contrast, the *psb*A-*trn*H sequence displayed murky barcoding gaps with overlapping intra- and interspecific variation distributions. Through calculation and traceability, we found that eight *P. amurense* samples had overlapping regions with haplotypes B2 and B3.

3.3 | NJ tree analysis

An NJ tree illustrates the relationship among species and facilitates their clustering. In this study, DNA barcode NJ trees were built based on the K2P model. A NJ tree built on the ITS sequence demonstrated that *P. amurense*, *P. chinense*, and *P. chinense* var. *glabriusculum* were short of monophyly. The NJ tree built on the *psb*A-*trn*H sequences clustered into two major branches (Figure 3), with 86.4% of *P. amurense* clustering into one branch, and the other eight samples clustering with *P. chinense* and *P. chinense* var. *glabriusculum* into the other branch. This finding meant that the phylogenetic relationship of *P. amurense* with *P. chinense* and *P. chinense* var. *glabriusculum* was paraphyly. Furthermore, eight *P. amurense* samples were particularly important for their intraspecific genetic variation and for species identification; these samples were haplotypes B2 and B3 by traceability.

When we performed the NJ tree analysis using a small-scale randomly selected samples from *P. amurense*, two typical NJ tree patterns existed. One pattern was the reciprocal monophyly in Figure 4, in which members from *P. amurense* and other species shared a unique common ancestor. The other was paraphyly (Figure 5), in which the *P. amurense* species is monophyletic but nests within another recognized species. Therefore, the adequately samples to cover infraspecific variation is essential for DNA barcoding.

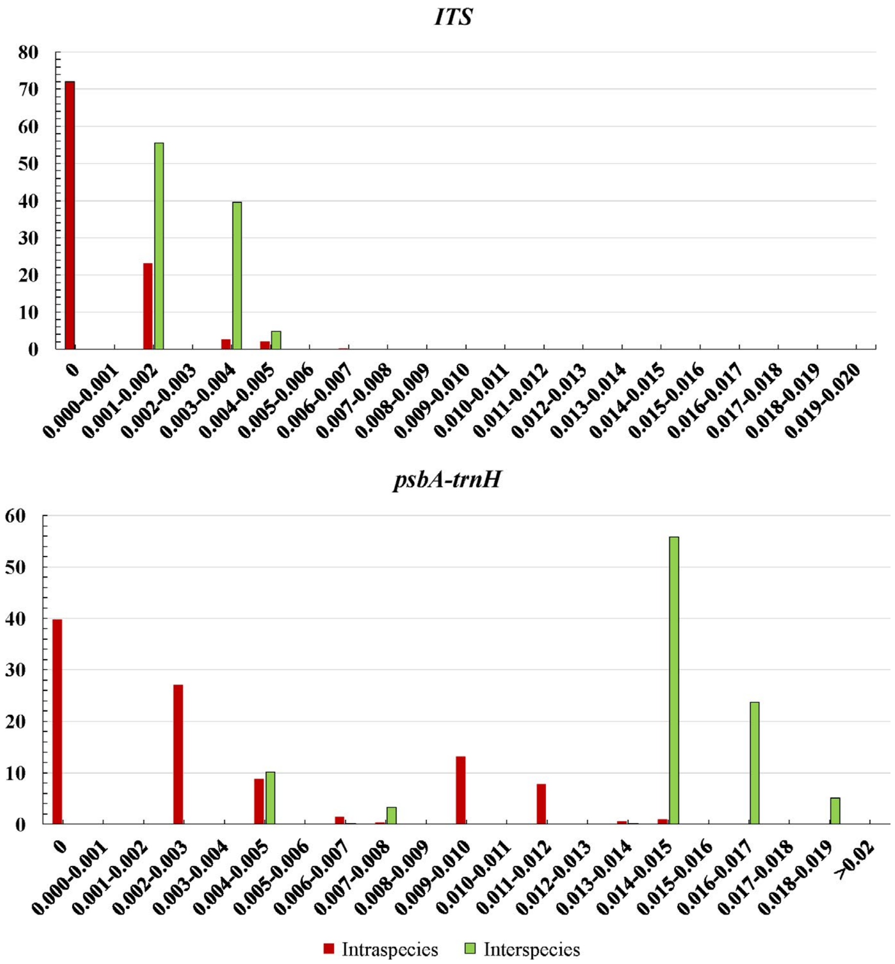

FIGURE 2 Relative distribution of the interspecific and intraspecific variation using the two barcodes based on the K2P genetic distance

3.4 | Genetic diversity parameters in the simulation sampling analysis

Since *psb*A-*trn*H performed better than ITS for identifying the *Phellodendron* species, we analyzed the relationship between the number of populations and the genetic diversity parameters based on *psb*A-*trn*H for *P. amurense*. The Michaelis–Menten equation [$f(x)$ = ax/(b + x)] is used to perform nonlinear regression (curve fit), which satisfied the requirements in this study. The haplotype discovery curve (HDC) is presented in Figure 6 with the theoretical equation $f(x)$ = 7.072x/(5.756 + x) and an r^2 = .8082. The results showed that the number of haplotypes (H) index gradually increased with the increase in the simulation sampling populations and had an overall sample level as shown in Figure 6. The scatter plots of the population number with haplotype diversity (H_d) and nucleotide diversity (P_i) are shown in Figures 7 and 8, respectively. These plots explained that the dispersion of the genetic diversity index at the same population level gradually decreased and stabilized as the number of simulation sampling populations increased.

3.5 | Theoretical key number of sample sizes in the simulation sampling analysis

The number of haplotypes (H) was a pivotal criterion for estimating the genetic diversity of *P. amurense* in simulation sampling. We arrived at the haplotype discovery curve (HDC) in Figure 6 with the theoretical equation $f(x)$ = 7.072x/(5.756 + x). We focused on the following two theoretical key parameters for sample sizes in this study: (a) the threshold of sample sizes where new haplotypes become considerably more difficult to identify with extra sampling efforts (the first-order derivative of the HDC curve is equal to zero) and (b) the sample size that includes the majority of haplotypes (indicating that 80% of haplotypes are found).

The theoretical key parameters are presented in Figure 6 after careful calculation based on the HDC curve theoretical equation. The figure showed that the theoretical threshold of the sample size is 5.756, which meant that discovering a new haplotype is much harder with only six populations in the simulation sampling. Furthermore, we could obtain 80% of *P. amurense* haplotypes when the theoretical threshold of the sample size was 23.024 populations using simulated sampling with the theoretical equation or sample sizes with no less than 24 populations in an actual sampling. To obtain a higher percentage of *P. amurense* haplotypes, we should increase the sample size.

3.6 | Critical analysis on populations

Since *psb*A-*trn*H performed better than ITS for identifying *Phellodendron* species, we analyzed the key populations based

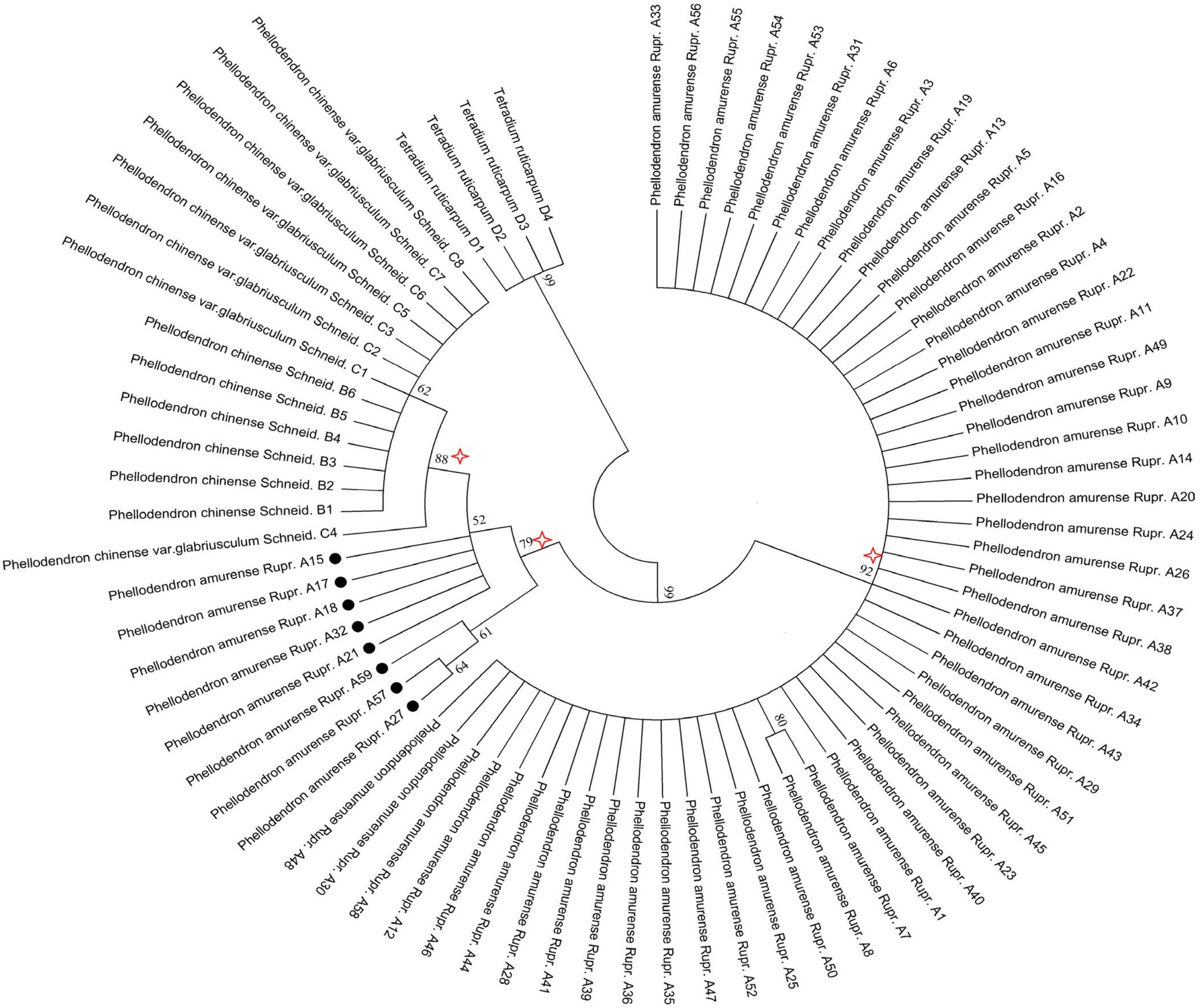
Tetradium ruticarpum D4
Tetradium ruticarpum D3
Tetradium ruticarpum D2
Tetradium ruticarpum D1
Phellodendron chinense var.glabriusculum Schneid. C8
Phellodendron chinense var.glabriusculum Schneid. C7
Phellodendron chinense var.glabriusculum Schneid. C6
Phellodendron chinense var.glabriusculum Schneid. C5
Phellodendron chinense var.glabriusculum Schneid. C3
Phellodendron chinense var.glabriusculum Schneid. C2
Phellodendron chinense var.glabriusculum Schneid. C1
Phellodendron chinense Schneid. B6
Phellodendron chinense Schneid. B5
Phellodendron chinense Schneid. B4
Phellodendron chinense Schneid. B3
Phellodendron chinense Schneid. B2
Phellodendron chinense Schneid. B1
Phellodendron chinense var.glabriusculum Schneid. C4
Phellodendron amurense Rupr. A15 ●
Phellodendron amurense Rupr. A17 ●
Phellodendron amurense Rupr. A18 ●
Phellodendron amurense Rupr. A32 ●
Phellodendron amurense Rupr. A21 ●
Phellodendron amurense Rupr. A59 ●
Phellodendron amurense Rupr. A57 ●
Phellodendron amurense Rupr. A27 ●
Phellodendron amurense Rupr. A48
Phellodendron amurense Rupr. A30
Phellodendron amurense Rupr. A58
Phellodendron amurense Rupr. A12
Phellodendron amurense Rupr. A46
Phellodendron amurense Rupr. A44
Phellodendron amurense Rupr. A28
Phellodendron amurense Rupr. A41
Phellodendron amurense Rupr. A39
Phellodendron amurense Rupr. A36
Phellodendron amurense Rupr. A35
Phellodendron amurense Rupr. A47
Phellodendron amurense Rupr. A52
Phellodendron amurense Rupr. A25
Phellodendron amurense Rupr. A50
Phellodendron amurense Rupr. A8
Phellodendron amurense Rupr. A7
Phellodendron amurense Rupr. A1
Phellodendron amurense Rupr. A40
Phellodendron amurense Rupr. A23
Phellodendron amurense Rupr. A45
Phellodendron amurense Rupr. A51
Phellodendron amurense Rupr. A29
Phellodendron amurense Rupr. A43
Phellodendron amurense Rupr. A34
Phellodendron amurense Rupr. A42
Phellodendron amurense Rupr. A38
Phellodendron amurense Rupr. A37
Phellodendron amurense Rupr. A26
Phellodendron amurense Rupr. A24
Phellodendron amurense Rupr. A20
Phellodendron amurense Rupr. A14
Phellodendron amurense Rupr. A10
Phellodendron amurense Rupr. A9
Phellodendron amurense Rupr. A49
Phellodendron amurense Rupr. A11
Phellodendron amurense Rupr. A22
Phellodendron amurense Rupr. A4
Phellodendron amurense Rupr. A2
Phellodendron amurense Rupr. A16
Phellodendron amurense Rupr. A5
Phellodendron amurense Rupr. A13
Phellodendron amurense Rupr. A19
Phellodendron amurense Rupr. A3
Phellodendron amurense Rupr. A6
Phellodendron amurense Rupr. A31
Phellodendron amurense Rupr. A53
Phellodendron amurense Rupr. A54
Phellodendron amurense Rupr. A55
Phellodendron amurense Rupr. A56
Phellodendron amurense Rupr. A33
99
62
88
52
79
99
92
61
64
80

FIGURE 3 *Phellodendron* species NJ tree with the *psb*A-*trn*H sequence (The bootstrap scores [1,000 replicates] are shown for each branch)

on *psb*A-*trn*H. Combining the haplotype analysis with the NJ tree analysis, we discovered eight special samples with seven populations existing in all the *P. amurense* populations. The samples' variable sites and genetic relationship were similar to *P. chinense* and *P. chinense* var. *glabriusculum* based on our data (Table 3 and Figure 3). Therefore, missing these *P. amurense* sample haplotypes would directly affect the reliability of sampling research results. We defined the eight *P. amurense* samples as critical samples and their populations as critical populations based on sufficient evidence. The geographical distribution of the seven critical *P. amurense* populations is shown in Figure 1 as red circles. We observed that they were uniformly distributed in the whole *P. amurense* area. This meant there was no key geographical group of *P. amurense*. Thus, it is highly meaningful to assess the sampling size from the uniform geographical distribution based on the genetic diversity.

4 | DISCUSSION

It is fact that different plant species will vary in the amount of infraspecific variation they contain, because of variation in length of time since divergence, and variation. Speciation is a gradual process, meaning that the discussion of species based on the evolutionary history of a cross section has significant limitations (Hennig, 1966). Intraspecific genetic variation of species is not only an integral part of evolutionary history but also changes continuously. Objectively, each species possesses abundant intraspecific genetic variation, and the intraspecific variation of ITS and *psb*A-*trn*H sequences is natural. It is necessary to assess the intraspecific genetic diversity of species. Furthermore, lack the genetic information of key populations will lead to DNA barcoding gaps, and the number of samples for collection is crucial to establish a reliable reference database for species identification (Meyer & Paulay, 2005; Wiemer & Fiedler, 2007).

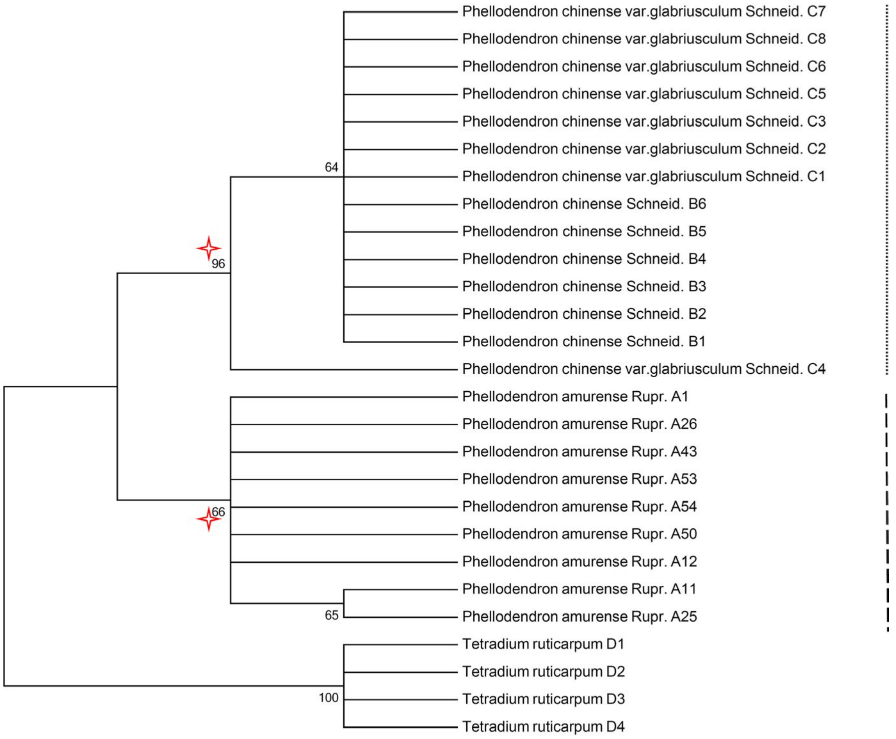
Phellodendron chinense var.glabriusculum Schneid. C7
Phellodendron chinense var.glabriusculum Schneid. C8
Phellodendron chinense var.glabriusculum Schneid. C6
Phellodendron chinense var.glabriusculum Schneid. C5
Phellodendron chinense var.glabriusculum Schneid. C3
Phellodendron chinense var.glabriusculum Schneid. C2
64
Phellodendron chinense var.glabriusculum Schneid. C1
Phellodendron chinense Schneid. B6
Phellodendron chinense Schneid. B5
96
Phellodendron chinense Schneid. B4
Phellodendron chinense Schneid. B3
Phellodendron chinense Schneid. B2
Phellodendron chinense Schneid. B1
Phellodendron chinense var.glabriusculum Schneid. C4
Phellodendron amurense Rupr. A1
Phellodendron amurense Rupr. A26
Phellodendron amurense Rupr. A43
Phellodendron amurense Rupr. A53
Phellodendron amurense Rupr. A54
66
Phellodendron amurense Rupr. A50
Phellodendron amurense Rupr. A12
Phellodendron amurense Rupr. A11
65
Phellodendron amurense Rupr. A25
Tetradium ruticarpum D1
Tetradium ruticarpum D2
100
Tetradium ruticarpum D3
Tetradium ruticarpum D4

FIGURE 4 Typical sample NJ tree pattern for reciprocal monophyly based on small-scale *P. amurense* with the *psb*A-*trn*H sequence (The bootstrap scores [1,000 replicates] is shown for each branch)

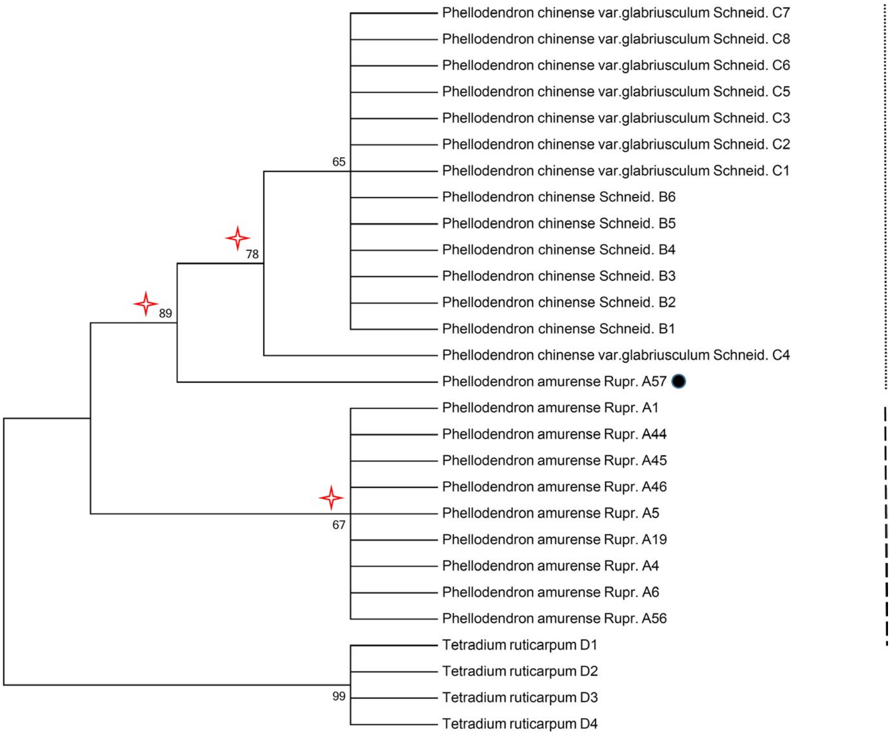
Phellodendron chinense var.glabriusculum Schneid. C7
Phellodendron chinense var.glabriusculum Schneid. C8
Phellodendron chinense var.glabriusculum Schneid. C6
Phellodendron chinense var.glabriusculum Schneid. C5
Phellodendron chinense var.glabriusculum Schneid. C3
Phellodendron chinense var.glabriusculum Schneid. C2
65
Phellodendron chinense var.glabriusculum Schneid. C1
Phellodendron chinense Schneid. B6
Phellodendron chinense Schneid. B5
Phellodendron chinense Schneid. B4
78
Phellodendron chinense Schneid. B3
89
Phellodendron chinense Schneid. B2
Phellodendron chinense Schneid. B1
Phellodendron chinense var.glabriusculum Schneid. C4
Phellodendron amurense Rupr. A57 ●
Phellodendron amurense Rupr. A1
Phellodendron amurense Rupr. A44
Phellodendron amurense Rupr. A45
Phellodendron amurense Rupr. A46
Phellodendron amurense Rupr. A5
67
Phellodendron amurense Rupr. A19
Phellodendron amurense Rupr. A4
Phellodendron amurense Rupr. A6
Phellodendron amurense Rupr. A56
Tetradium ruticarpum D1
Tetradium ruticarpum D2
99
Tetradium ruticarpum D3
Tetradium ruticarpum D4

FIGURE 5 Typical NJ tree pattern of paraphyly based on small-scale *P. amurense* samples with the *psb*A-*trn*H sequence (The bootstrap scores [1,000 replicates] are shown for each branch)

Dasmahapatra and Mallet (2006) believes that many studies only analyze and collect 1 or 2 samples from one species within a limited geographical scope, which seriously underestimates the intraspecific variation and may lead to false-positive results.

In study, we paid close attention to the genetic diversity of *P. amurense* and collected adequately samples throughout the distribution area to ensure covering infraspecific variation. We conducted a comparative analysis of the overall samples and small-scale randomly samples. The results showed that small-scale samples could not demonstrate an objective phylogenetic relationship in *Phellodendron* (Figures 3, 4 and 5). It meant that an insufficient sample would reduce identification accuracy of the DNA barcoding. In the simulation sample analysis, we found 1–2 samples from 23 populations based on uniformed geographical distribution are necessary to obtain critical samples and represent 80% of the genetic diversity of *P. amurense* (Figure 6). It did not support other studies, such as Liu, Provan, et al.

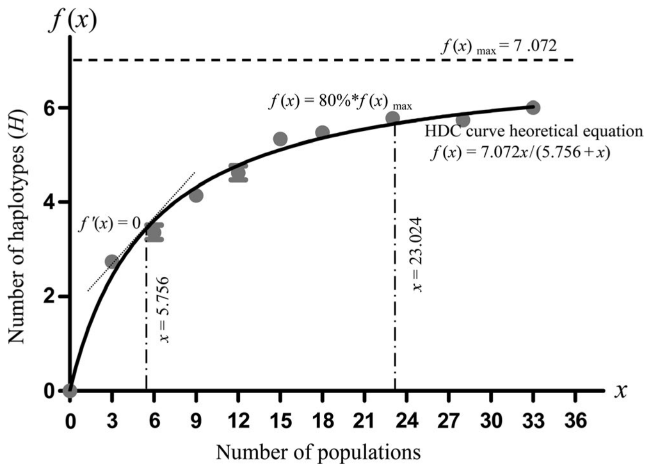

FIGURE 6 Haplotype number versus population number curve

(2012) believe that 8–10 samples of each species are representative of the genetic diversity in *Taxus*, and Matz and Nielsen (2005) recommended that 12 samples from a species can used to study DNA barcoding. We agreed with Zhang, He, Crozier, Muster, and Zhu (2010) hold that 5–10 samples for each species are not sufficient to represent species genetic diversity. When studying DNA barcoding with closely related species, obtaining a specific sample size for each species should become the research objective. Our research showed that errors may occur in the DNA barcoding identification method, which would establish unrepresentative samples sizes, including the lack of the B2 and B3 haplotypes. The key geographical populations have a great impact on the accuracy of DNA barcoding to identify species (Bergsten et al., 2012; Meyer & Paulay, 2005). The B2 and B3 haplotype populations were uniformly distributed in the entire *P. amurense* area, which meant there were no key geographical groups (Figure 1). *P. amurense* clearly underwent genetic variation among the existing populations. It was successful to take large-scale uniform sampling covering the entire distribution area in study.

The most important challenge for species identification is DNA barcodes used for closely related species and recently differentiated species (Newmaster, Fazekas, Steeves, & Janovec, 2008). DNA barcoding has been applied to crude drug identification (Kool et al., 2012; Techen, Parveen, Pan, & Khan, 2014), but we had to keep doubts about the identification accuracy of DNA barcoding based on the results of our study. Whether the sample size of each species may represent the actual levels of genetic diversity in the current database needs to be studied (Chen, 2015; Chen et al., 2014). In addition, the lack of key plant specimens will lead to narrow genetic levels in the database resulted in DNA barcoding failures. Furthermore, the input and output ratio is an important factor to restrict large sample strategies implemented in DNA barcoding database of each species.

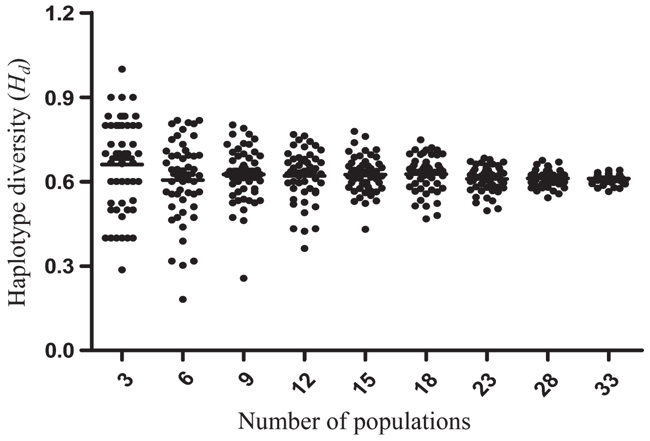

FIGURE 7 Scatter plot of haplotype diversity

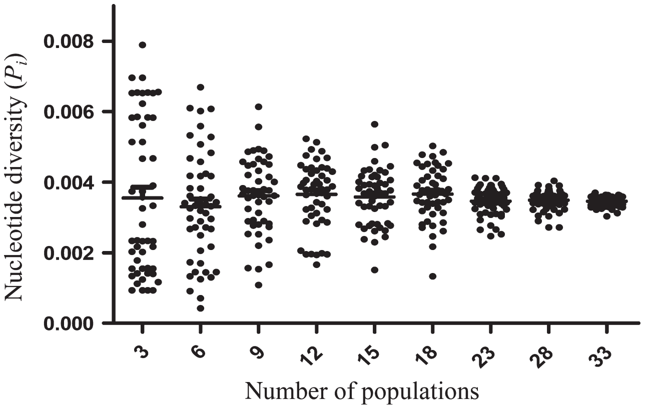

FIGURE 8 Scatter plot of nucleotide diversity

In order to ensure identification accuracy of DNA barcoding which is used as tools for species identification, it is particularly important to collect adequately samples covering infraspecific genetic diversity of species.

ACKNOWLEDGMENTS

The authors are grateful for the financial support provided by the National Natural Science Foundation of China (No. 81473305) and the National Science and Technology Foundational Special Project (2015FY111500).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHORS CONTRIBUTIONS

Zhi-peng Zhang: experimental operation, manuscript writing; Xiao-yue Wang: experimental operation; Zhao Zhang: corresponding author, experimental design; Hui Yao: manuscript modification; Xiao-mei Zhang: manuscript modification; Yang Zhang: sample Collection; Ben-gang Zhang: experimental design.

DATA AVAILABILITY STATEMENT

ITS and *psb*A-*trn*H sequence data can be accessed in GenBank. The GenBank accession No. is shown in Table 1.

ORCID

Zhao Zhang https://orcid.org/0000-0002-3209-9071

REFERENCES

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., ... Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, *61*, 851–869. https://doi.org/10.1093/sysbio/sys037

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.

Chen, S. L. (2012). *DNA Barcoding identification of Chinese materia medica*. Beijing, China: People's Medical Press.

Chen, S. L. (2015). *Standard DNA barcodes of Chinese materia medica in Chinese Pharmacopoeia*. Beijing, China: China Science Press.

Chen, S. L., Pang, X. L., Song, J. Y., Shi, L. C., Yao, H., Han, J. P., & Leon, C. (2014). A renaissance in herbal medicine identification: From morphology to DNA. *Biotechnology Advances*, *32*, 1237–1244. https://doi.org/10.1016/j.biotechadv.2014.07.004

China Plant BOL Group (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 19641–19646. https://doi.org/10.1073/pnas.1104551108

Dasmahapatra, K. K., & Mallet, J. (2006). DNA barcodes: Recent successes and future prospects. *Heredity*, *97*, 254–255.

Gao, L. M., Liu, J., Cai, J., Yang, J. B., Zhang, T., & Li, D. Z. (2012). A synopsis of technical notes on the standards for plant DNA barcoding. *Plant Diversity and Resources*, *34*, 592–606. https://doi.org/10.3724/SP.J.1143.2012.12138

Hebert, P. D. N., Cywinska, A., Ball, S. L., & Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hennig, W. (1966). *Phylogenetic systematics*. Urbana, IL: University of Illinois Press.

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE*, *6*, e19254. https://doi.org/10.1371/journal.pone.0019254

Huang, C. J. (1958). Preliminary study on Chinese Rutaceae (2). *Acta Phytotaxonomica Sinica*, *7*, 229–261.

Huang, C. J. (1997). Phellodendron Rupr. In C. J. Huang (Ed.), *Flora of Reipublicae Popularls Sinicae*, *Angiospermae*, *Dicotyledoneae*, *Rutaceae* (vol. *42*, Issue 2, pp. 99–103). Beijing, China: China Science Press.

Kool, A., de Boer, H. J., Krüger, Å., Rydberg, A., Abbad, A., Björk, L., & Martin, G. (2012). Molecular identification of commercialized medicinal plants in Southern Morocco. *PLoS ONE*, *7*, e39459. https://doi.org/10.1371/journal.pone.0039459

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (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. https://doi.org/10.1073/pnas.0503123102

Li, H. Q., Chen, J. Y., Wang, S., & Xiong, S. Z. (2012). Evaluation of six candidate DNA barcoding loci in *Ficus* (*Moraceae*) of China. *Molecular Ecology Resources*, *12*, 783–790. https://doi.org/10.1111/j.1755-0998.2012.03147.x

Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*, 1451–1452. https://doi.org/10.1093/bioinformatics/btp187

Liu, J., Provan, J., Gao, L. M., & Li, D. Z. (2012). Sampling strategy and potential utility of indels for DNA barcoding of closely related plant species: A case study in *Taxus*. *International Journal of Molecular Sciences*, *13*, 8740–8751. https://doi.org/10.3390/ijms13078740

Liu, Y. M., Zhang, L. H., Liu, Z., Luo, K., Chen, S. L., & Chen, K. L. (2012). Species identification of *Rhododendron* (*Ericaceae*) using the chloroplast deoxyribonucleic acid *PsbA-trnH* genetic marker. *Pharmacognosy Magazine*, *8*, 29. https://doi.org/10.4103/0973-1296.93311

Luo, A., Lan, H. Q., Ling, C., Zhang, A. B., Shi, L., Simon, Y. W. H., & Zhu, C. D. (2015). A simulation study of sample size for DNA barcoding. *Ecology and Evolution*, *5*, 5869–5879. https://doi.org/10.1002/ece3.1846

Matz, M. V., & Nielsen, R. (2005). A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1969–1974. https://doi.org/10.1098/rstb.2005.1728

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, *3*, 2229–2238. https://doi.org/10.1371/journal.pbio.0030422

Newmaster, S. G., Fazekas, A. J., Steeves, R. A. D., & Janovec, J. (2008). Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, *8*, 480–490. https://doi.org/10.1111/j.1471-8286.2007.02002.x

Qin, Y. J., Wang, Y., & Yan, X. F. (2006). Status and sustainable utilization strategy of *Phellodendron amurense* resources in China. *Chinese Traditional and Herbal Drugs*, *37*, 1104–1107.

State Bureau of Environmental Protection of China & Institute of Botany Chinese Academy of Sciences (1987). *The list of chinese rare and endangered plants*. Beijing, China: China Science Press.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*, 2731–2739. https://doi.org/10.1093/molbev/msr121

Tavares, E. S., & Baker, A. J. (2008). Single mitochondrial gene barcodes reliably identify sister-species in diverse clades of birds. *BMC Evolutionary Biology*, *8*, 81. https://doi.org/10.1186/1471-2148-8-8

Techen, N., Parveen, I., Pan, Z. Q., & Khan, I. A. (2014). DNA barcoding of medicinal plant material for identification. *Current Opinion in Biotechnology*, *25*, 103–110. https://doi.org/10.1016/j.copbio.2013.09.010

Wang, L. J., Bao, L., Wang, H. F., & Ge, J. P. (2014). Study on the genetic diversity of *Phellodendron amurense*. *China Sciencepaper*, *9*, 1397–1401.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1847–1857. https://doi.org/10.1098/rstb.2005.1716

Wiemer, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist-A case study in blue butterflies (*Lepidoptera*: *Lycaenidae*). *Frontiers in Zoology*, *4*, 8. https://doi.org/10.1186/1742-9994-4-8

Yan, Z. F., Zhang, B. G., Zhang, Z., & Yu, J. L. (2006). Genetic diversity in wild populations of *Phellodendron amurense*, a rare and endangered medicinal plant, detected by AFLP. *Biodiversity Science*, *14*, 488–497. https://doi.org/10.1360/biodiv.060041

Yang, J. B., Wang, Y. P., Möller, M., Gao, L. M., & Wu, D. (2012). Applying plant DNA barcodes to identify species of *Parnassia* (*Parnassiaceae*). *Molecular Ecology Resources*, *12*, 267–775. https://doi.org/10.1111/j.1755-0998.2011.03095.x

Yao, H., Song, J. Y., & Ma, X. Y. (2009). Identification of *Dendrobium* species by a candidate DNA barcode sequence: The chloroplast *psbA-trnH* intergenic region. *Planta Medica*, *75*, 667–669.

Young, M. K., McKelvey, K. S., Pilgrim, K. L., & Schwartz, M. K. (2013). DNA barcoding at riverscape scales: Assessing biodiversity among fishes of the genus *Cottus* (*Teleostei*) in northern Rocky Mountain streams. *Molecular Ecology Resources*, *13*, 583–595.

Yu, J. H., Chen, C. M., Han, S. J., Guo, X. R., Yuan, S. S., Wang, C. J., & Zu, Y. G. (2013). Development and characterization of polymorphic microsatellite loci in *Phellodendron amurense* (Rutaceae). *Applications in Plant Sciences*, *1*, 1200321.

Zhang, A. B., He, L. J., Crozier, R. H., Muster, C., & Zhu, C. D. (2010). Estimating sample sizes for DNA barcoding. *Molecular Phylogenetic and Evolution*, *54*, 1035–1039. https://doi.org/10.1016/j.ympev.2009.09.014

How to cite this article: Zhang Z-P, Wang X-Y, Zhang Z, et al. The impact of genetic diversity on the accuracy of DNA barcoding to identify species: A study on the genus *Phellodendron*. *Ecol Evol*. 2019;9:10723–10733. https://doi.org/10.1002/ece3.5590