Received: 28 February 2018 | Revised: 1 May 2018 | Accepted: 7 May 2018

DOI: 10.1111/1755-0998.12903

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Integrating a comprehensive DNA barcode reference library with a global map of yews (*Taxus* L.) for forensic identification

Jie Liu[1,2] | Richard I. Milne[3] | Michael Möller[4] | Guang-Fu Zhu[1,2,5] | Lin-Jiang Ye[1,2,5] | Ya-Huang Luo[1] | Jun-Bo Yang[2] | Moses C. Wambulwa[6] | Chun-Neng Wang[7] | De-Zhu Li[2,5] | Lian-Ming Gao[1]

[1]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

[2]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China

[3]Institute of Molecular Plant Sciences, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

[4]Royal Botanic Garden Edinburgh, Edinburgh, Scotland, UK

[5]College of Life Sciences, University of Chinese Academy of Sciences, Kunming, Yunnan, China

[6]Biochemistry Department, South Eastern Kenya University, Kitui, Kenya

[7]Institute of Ecology and Evolutionary Biology, Department of Life Science, National Taiwan University, Taipei, China

Correspondence
De-Zhu Li and Lian-Ming Gao, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China.
Emails: gaolm@mail.kib.ac.cn and dzl@mail.kib.ac.cn

Funding information
National Key Basic Research Program of China, Grant/Award Number: 2014CB954100; National Natural Science Foundation of China, Grant/Award Number: 41571059, 31200182, 31370252; Interdisciplinary Research Project of Kunming Institute of Botany, Grant/Award Number: KIB2017003; Ministry of Science and Technology, China, Grant/Award Number: 2013FY112600; China Scholarship Council, Grant/Award Number: 201504910423

Abstract

Rapid and accurate identification of endangered species is a critical component of biosurveillance and conservation management, and potentially policing illegal trades. However, this is often not possible using traditional taxonomy, especially where only small or preprocessed parts of plants are available. Reliable identification can be achieved via a comprehensive DNA barcode reference library, accompanied by precise distribution data. However, these require extensive sampling at spatial and taxonomic scales, which has rarely been achieved for cosmopolitan taxa. Here, we construct a comprehensive DNA barcode reference library and generate distribution maps using species distribution modelling (SDM), for all 15 *Taxus* species worldwide. We find that *trnL-trnF* is the ideal barcode for *Taxus*: It can distinguish all *Taxus* species and in combination with ITS identify hybrids. Among five analysis methods tested, NJ was the most effective. Among 4,151 individuals screened for *trnL-trnF*, 73 haplotypes were detected, all species-specific and some population private. Taxonomical, geographical and genetic dimensions of sampling strategy were all found to affect the comprehensiveness of the resulting DNA barcode library. Maps from SDM showed that most species had allopatric distributions, except *T. mairei* in the Sino-Himalayan region. Using the barcode library and distribution map data, two unknown forensic samples were identified to species (and in one case, population) level and another was determined as a putative interspecific hybrid. This integrated species identification system for *Taxus* can be used for biosurveillance, conservation management and to monitor and prosecute illegal trade. Similar identification systems are recommended for other IUCN- and CITES-listed taxa.

KEYWORDS
comprehensive sampling, DNA barcoding, forensic identification, geographical origin, sampling strategy, species distribution modelling

1 | INTRODUCTION

The extinction risk for plants and animals is driven by multiple natural and anthropogenic factors, but varies between regions and taxa (Ceballos, Ehrlich, & Dirzo, 2017; Tilman et al., 2017). Anthropogenic-induced factors, such as climate and land-use change, overexploitation and deforestation, are pushing the Earth's biota towards a sixth "mass extinction" (Ceballos et al., 2015). A particular threat to some taxa comes from the overexploitation for commercial trade in plants and their products, which has dramatically increased in recent decades. International conventions like CITES (Convention on International Trade in Endangered Species) and efforts at the national level are designed to combat illegal trades for endangered and threatened species, but the effectiveness of their governing rules and measures is highly dependent upon the rapid and accurate identification of the threatened species. The same applies to successful management of habitats and populations: It is essential to know exactly which taxa are present.

Until recently, plant identification has been largely dependent upon morphology-based approaches, which in turn depended upon taxonomical specialists, who are generally the only experts on some specific groups of plant (Godfray, 2002; Li et al., 2011). Moreover, where available material is sterile, juvenile and/or poor in quality, accurate identification even by an expert may be impossible. Furthermore, traditional taxonomic approaches can rarely be scaled up for high throughput (Li et al., 2011), making it inconvenient for routine forensic applications in species identification.

DNA-based approaches, such as DNA barcoding, are more universally applicable than morphological approaches, often less subjective, and do not rely on expertise in the specific group under investigation (Hebert, Cywinska, Ball, & deWaard, 2003). DNA barcoding sensu stricto compares short sequences from a standardized portion of the genome with a known DNA barcode reference library, to identify the species to which a particular specimen belongs (Hebert et al., 2003; Valentini, Pompanon, & Taberlet, 2009). It shows powerful universality and versatility at the species level and can sometimes provide insights beyond those obtained through morphological analysis alone (Blaxter, 2004). In the presence of a well-established reference library, an unknown sample can theoretically be identified to species using its DNA barcode sequences.

Huge amounts of DNA barcode data are now available, providing invaluable insights for understanding species' boundaries, community ecology and trophic interactions in ecology and evolution (Joly et al., 2014; Kress, 2017; Valentini et al., 2009). Furthermore, the technology is gradually gaining popularity in such fields as forensic identification (Ferri et al., 2015), authentication of medicinal herbs (Chen et al., 2010) and timber identification (Dormontt et al., 2015). However, genetic variation occurs, often abundantly, within species and populations, and especially across the distribution range of particularly of widespread taxa (Avise, 2000). Therefore, a comprehensive, solid and reliable DNA reference library is an indispensable prerequisite for any of these applications (deWaard, Hebert, & Humble, 2011; Ogden & Linacre, 2015), and this requires ample sampling within and across populations, covering the full range of a taxon (Bergsten et al., 2012; Ekrem, Willassen, & Stur, 2007). A broad taxonomic barcode coverage has been achieved for certain groups in recent years, but this success has so far always been limited to animal groups and restricted to specific geographical regions, for example, Canadian spiders (Blagoev et al., 2016), German mayflies, stoneflies and caddisflies (Morinière et al., 2017) and perciform fishes in the South China Sea (Hou, Chen, Lu, Cheng, & Xie, 2018). A comprehensive barcode library could be defined as one that captures 95% of genetic variation, and this has been estimated to require a minimum of 70 (Bergsten et al., 2012) or 156 (Zhang, He, Crozier, Muster, & Zhu, 2010) individuals per species; this is also affected by the geographical scale of sampling (Bergsten et al., 2012) and the population structure of the species sampled (Zhang et al., 2010). However, it is often difficult or impossible to obtain material from the full distribution range of a species; therefore, many existing libraries are incomplete, introducing bias and possible misidentifications. These issues could cause serious problems for conservation and especially law enforcement regarding IUCN- and CITES-listed taxa.

Taxus is the most diverse genus within Taxaceae, with 13 recognized species (Farjon, 2010; Möller et al., 2013; Spjut, 2007) plus two additional cryptic species (currently known as the Emei type and Qinling type; Liu, Möller, Gao, Zhang, & Li, 2011), hereafter referred to as "species" for simplicity. The genus is broadly distributed across temperate of the northern Hemisphere, covering North America, Europe, North Africa and Asia (Supporting information: Figure S1; Farjon & Filer, 2013). It has acquired great medical significance as the source of taxol, a natural antitumour agent with high potential for cancer treatments (Itokawa & Lee, 2003). However, its species are slow-growing and scattered distribution, thus rarely occur in large numbers (Fu, Li, & Mill, 1999). Consequently, commercial exploitation and the illegal trade of its bark and leaves for taxol have caused a sharp decline in its natural populations (Schippmann, 2001). According to IUCN (2017), *T. floridana* is critically endangered due to deforestation and land-use change, whereas *T. brevifolia*, *T. globosa*, *T. contorta* (synonym *T. fuana*), *T. chinensis* and *T. wallichiana* are either endangered or near threatened (Table 1); furthermore, the latter three species plus *T. cuspidata* are listed by CITES (2007) in appendix II. Three of the remaining nine recognized species have yet to be evaluated (IUCN, 2017; Table 1). At the national level, all native Chinese *Taxus* species are listed as first-class national protected plants (State Forestry Administration and Ministry of Agriculture P.R. China 1999), and the export of native *Taxus* from India is prohibited (Sajwan & Prakash, 2007). Nevertheless, illegal exploitation is still rampant; for example, there were 34 convictions involved from a single case in China (Tang, 2010).

Policing this illegal trade requires accurate identification of species. However, morphological characters tend to vary greatly within species and often with overlap among species, leading to ongoing taxonomic controversy (Möller et al., 2013), especially in Asia (Fu et al., 1999; Spjut, 2007), which in turn causes uncertainty about the

TABLE 1 List of 15 species (=types) of *Taxus* included in this study, and their recent common synonyms, status in IUCN and CITES; number of occurrence points (N_O), AUC and the thresholds selected in the species distribution modelling (SDM)

Taxon	Reference	Common synonyms	IUCN (2013)[a]	CITES (2007)[b]	N_O	AUC[c]	Threshold[d]
T. baccata	Farjon (2010)		Least Concern		223	0.964	0.3196
T. brevifolia	Farjon (2010)		Near Threatened		99	0.971	0.2924
T. calcicola	Möller et al. (2013)				25	0.997	0.1650
T. canadensis	Farjon (2010)		Least Concern		147	0.932	0.3821
T. chinensis	Farjon (2010)	*T. wallichiana* var. *chinensis*	Endangered A2d	Appendix II	69	0.995	0.1483
T. contorta	Farjon (2010)	*T. fuana*	Endangered A2acd	Appendix II	52	0.996	0.2013
T. cuspidata	Farjon (2010)		Least Concern	Appendix II	53	0.967	0.2717
Emei type	Liu et al. (2011, 2012), Möller et al. (2013)				40	0.998	0.2917
T. floridana	Farjon (2010)		Critically Endangered B1ab (iii,v)		7	0.998	0.6717
T. florinii	Spjut (2007), Möller et al. (2013)				65	0.997	0.3134
T. globosa	Farjon (2010)		Endangered A2c		119	0.994	0.0909
T. mairei	Farjon (2010)	*T. sumatrana*; *T. wallichiana* var. *mairei*	Vulnerable A2d		103	0.975	0.2548
T. phytonii	Spjut (2007)				30	0.999	0.1431
Qinling type	Liu et al. (2011, 2012; Möller et al. (2013)				60	0.995	0.2640
T. wallichiana	Farjon (2010)	*T. yunnanensis*; *T. wallichiana* var. *wallichiana*	Endangered A2acd	Appendix II	94	0.995	0.2209

Notes. [a]IUCN, population status assessed according to IUCN categories & criteria 2001 (version 2017-2) in 2013. [b]CITES Appendix II, population status evaluated in 2007. [c]AUC, area under the receiver operating characteristic curve. [d]Thresholds selected according to Liu et al. (2005).

distribution range of some species. Hence, species identification is difficult even from complete specimens of known origin and often impossible from the limited parts of the plant typically used in illegal trade (e.g., bark, leaves and timber). Therefore, an accurate, quick, cost-effective and universally applicable identification system for *Taxus* species is badly needed to support and enforce international and national plant protection laws. DNA barcoding, if supported by adequate sampling plus species distribution modelling (SDM), provides an ideal solution.

The goal of this study, therefore, was to use comprehensive sampling to create a practical DNA barcode identification system, supported by SDM, for the genus *Taxus*, which could be applied to identify material up to species or population level and hence set up a repeatable workflow for other IUCN- and CITES-listed taxa. The three specific objectives were to (a) determine the ideal DNA barcode, identification method and sampling strategy; (b) construct a comprehensive DNA barcode reference library and a global map; and (c) demonstrate the potential forensic applications of the data for species identification.

2 | MATERIALS AND METHODS

2.1 | Sampling

As noted above, *Taxus* is a notoriously taxonomically difficult genus, with ongoing uncertainty and disagreement about its classification (Fu et al., 1999; Möller et al., 2013) (Table 1). For this study, we recognized 10 species according to Farjon (2010), plus two from Möller et al. (2013) and two cryptic species from China revealed by our previous studies (Liu et al., 2011; Möller et al., 2013) (Table 1). The species *T. sumatrana* was not recognized by Farjon (2010), and morphological and molecular data have also indicated that Indonesian material previously included in *this species* was identical to the Sino-Himalayan species *T. mairei*, in which it is now included. However, material previously included in *T. sumatrana* from the Philippines and Taiwan was distinct (Liu et al., 2011). Based on our preliminary analysis (data not shown), the name *T. phytonii* (Spjut, 2007) was suitable to represent this material and is therefore used for it here, making 15 recognized species in total.

A total of 2,636 accessions were available from previous studies, though mostly only as *trnL-trnF* sequences (Gao et al., 2007; Kozyrenko, Artyukova, & Chubar, 2017; Liu, Provan, Gao, & Li, 2012; Liu et al., 2013; Mayol et al., 2015; Poudel et al., 2012; Poudel, Möller, Li, Shah, & Gao, 2014; Poudel, Möller, Liu et al., 2014; Rachmat, Subiakto, & Kamiya, 2016). To these were added 1,515 newly sampled individuals, collected from 73 populations between 2012 and 2016, making a total of 4,151 accessions and 251 populations representing all 15 species and the global distribution range of *Taxus* (Figure 1; Supporting information: Figure S2, Supporting information: Table S1). Sampling was conducted at higher density for the more taxonomically difficult Asian species. Healthy and clean needles were collected, dried and stored in silica gel for DNA extraction. Voucher specimens for most sampled accessions were deposited at the herbarium of Kunming Institute of Botany, Chinese Academy of Sciences (KUN).

2.2 | Laboratory procedures

Total genomic DNA was isolated from dried leaves using a modified CTAB method (Liu & Gao, 2011). The quality and quantity of DNA were measured on 1% TAE agarose gels and using a NanoDrop® ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). The DNA was diluted to a final concentration of 30–50 ng/μl for PCR amplifications.

For land plants, two "core barcodes" (*rbcL* plus *matK*) plus two complementary barcodes (ITS and *psbA-trnH*) have been proposed (CBOL Plant Working Group 2009; Kress, Wurdack, Zimmer, Weigt, & Janzen, 2005; Li et al., 2011). For this study, therefore, we examined all four of these markers, plus *trnL-trnF* which has already been used extensively for DNA barcoding (Liu et al., 2011, 2012) and phylogeographical analyses within *Taxus* (Gao et al., 2007; Kozyrenko et al., 2017; Liu et al., 2013; Mayol et al., 2015; Poudel, Möller, Li, Shah, & Gao, 2014; Poudel, Möller, Liu et al., 2014; Rachmat et al., 2016).

Following Liu et al. (2011), universal primers were used for four regions (*rbcL*, *psbA-trnH*, *trnL-trnF* and ITS), whereas for *matK*, new primers specifically developed for gymnosperms were employed (Supporting information: Table S2). PCRs were carried out on a Veriti® 96-Well Thermal Cycler (Applied Biosystems, Foster City, USA) as Liu et al. (2011). PCR products were purified using ExoSAP-IT (GE Healthcare, Cleveland, OH, USA). Purified PCR products were sequenced bidirectionally on an ABI 3730xl DNA Sequencer (Applied Biosystems).

2.3 | Data analysis

2.3.1 | Sequence and data set assembly

The forward and reverse chromatograms of each sequence were assembled and aligned in GENEIOUS v9.1.4 (Biomatters Ltd, Anzac Avenue, New Zealand) and subsequently adjusted manually where necessary. All variable sites in the matrices were rechecked in the original trace files. We generated 60, 62, 53, 81 and 1500 new sequences for *rbcL*, *matK*, *psbA-trnH*, ITS and *trnL-trnF*, respectively. Newly generated sequences as well as selected sequences downloaded from GenBank (Supporting information: Tables S1, S3) were used to construct DNA barcode data sets. In total, we used 110, 173, 167, 195 and 4151 individuals for *rbcL*, *matK*, *psbA-trnH*, ITS and *trnL-trnF*, respectively.

To determine the effectiveness of different barcodes and sampling strategies for each species, three data sets were constructed, each of which included representatives of all 15 species. Data set I comprised 72 individuals, each having all the five barcode sequences, and with individuals of each species selected so as to cover the entire distribution range of the species (Figure 2a). This data set was used to screen the candidate barcodes and compare species identification methods for *Taxus*. Data set II comprised 201 accessions, for which at least two barcode sequences were available, including all those in Data set I, although not all were sequenced for every

barcode: 110, 173, 167, 190 and 195 were sequenced for *rbcL*, *matK*, *psbA-trnH*, *trnL-trnF* and ITS within this data set, respectively (Supporting information: Figure S2a, Supporting information: Table S1). This data set was used to verify the reliability and confidence of the proposed barcode from Data set I, and from it was generated an ITS ribotype reference library based on 195 individuals, intended for applications in routine identification of samples. Finally, Data set III comprised a comprehensive haplotype DNA barcode library for *trnL-trnF* from all of the 4,151 individuals sampled across 251 populations worldwide (Figure 1; Supporting information: Table S1), and this data set was further used to confirm the sampling strategy.

The discriminatory properties of every individual barcode, plus every possible combination of all the five barcodes, were examined. All combinations were concatenated in SequenceMatrix v1.7.8 (Vaidya, Lohman, & Meier, 2011). To estimate the levels of variation and barcoding gap within the five examined DNA regions, the mean intra- and interspecific pairwise Kimura two-parameter (k2p) distance for each DNA region and their combinations were calculated using MEGA v5.0 (Tamura et al., 2011).

2.3.2 | Barcode evaluation and identification methods comparison with Data set I

To determine the optimal species identification method, three widely used methods were applied, both to each marker individually and to every concatenated combination: tree-based, coalescent-based and distance-based methods. Maximum likelihood (ML) and neighbour-joining (NJ) tree construction were performed for each marker and their combinations, respectively, using the RAXML web server (Stamatakis, Hoover, & Rougemont, 2008) and MEGA. Indels were treated as missing data in ML and pairwise deletion in NJ analyses. For ML analyses, the model GTR + G was selected with JMODELTEST 2 (Darriba, Taboada, Doallo, & Posada, 2012) for all data sets, and a rapid bootstrap analysis with 999 trees was conducted. The NJ tree was constructed under the *P*-distance substitution model. The bootstrap support of the NJ tree was assessed using 999 replicates. In the NJ and ML analyses, species identification was considered to be successful as long as all the conspecific individuals formed a species-specific monophyletic clade. The ratio of successfully identified species to all sampled species was calculated as the discrimination efficiency. Additionally, we also adopted a coalescent-based tree building method, the Poisson tree process model (PTP) (Zhang, Kapli, Pavlidis, & Stamatakis, 2013), to test the species discrimination rate. The analysis was implemented in a bPTP web server (http://species.h-its.org/ptp/) with default parameters and without outgroup. The phylogenetic tree from RAXML analysis was used as the input file.

The Automated Barcode Gap Discovery (ABGD) method (Puillandre, Lambert, Brouillet, & Achaz, 2012) (available at http://wwwabi.snv.jussieu.fr/public/abgd/) was employed to detect the barcode gap in the distribution of pairwise distances to test species delimitation. The k2p distance matrixes generated in MEGA were submitted and processed in ABGD with the range of prior intraspecific divergence set between 0.0001 and 0.003.

SpeciesIdentifier v1.7.8 from the TaxonDNA (Meier, Shiyang, Vaidya, & Ng, 2006) was used to test the individual-level discrimination rates for each single marker and their combinations with a 95%

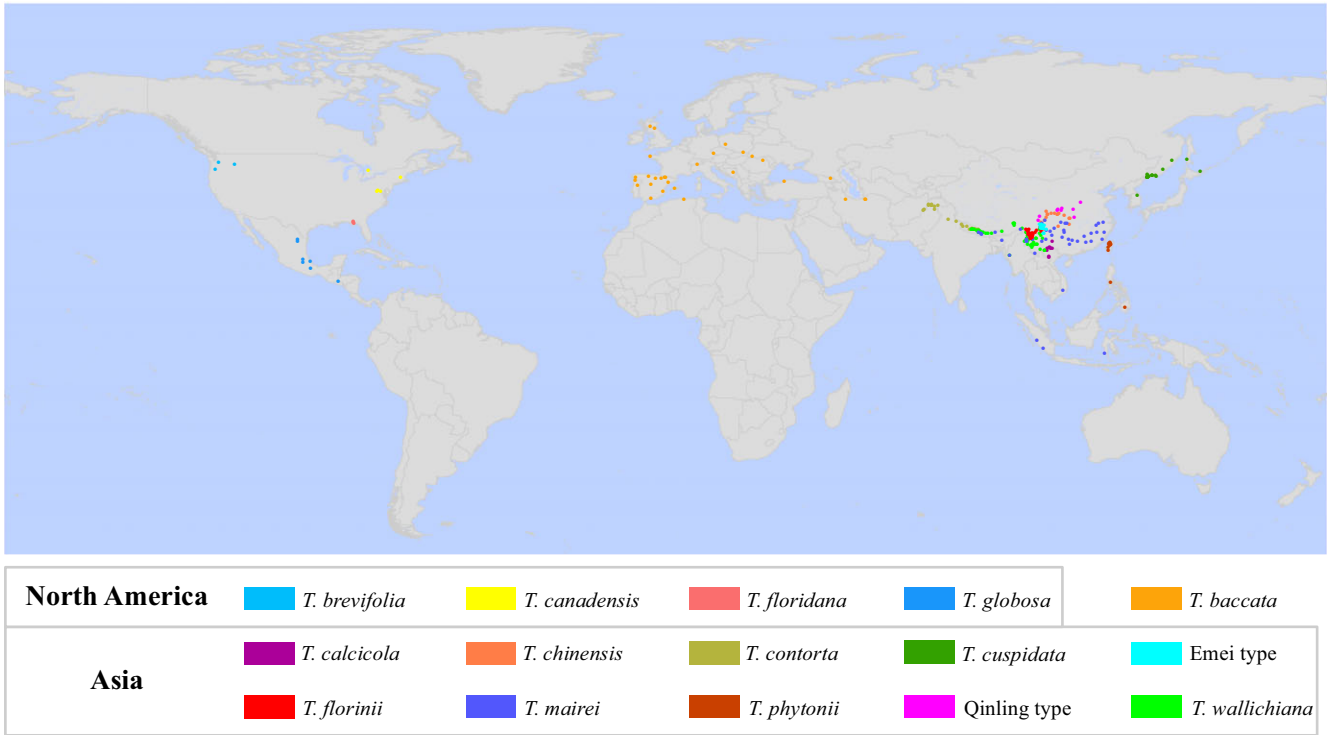

FIGURE 1 Location of the *Taxus* populations sampled around the world according to Supporting information: Table S1. The different colours represent various species shown in the legend [Colour figure can be viewed at wileyonlinelibrary.com]

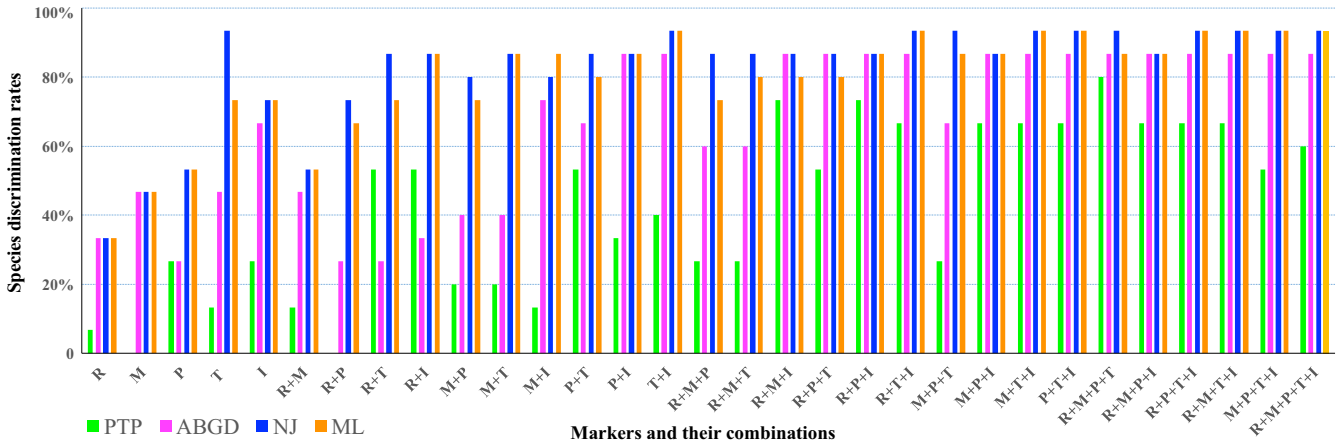

FIGURE 2 The species discrimination rates of five single barcodes and their concatenations for 72 individuals of 15 *Taxus* lineages based on PTP, ABGD, NJ and ML methods (R: *rbcL*; M: *matK*; P: *psbA-trnH*; T: *trnL-trnF*; and I: ITS) [Colour figure can be viewed at wileyonlinelibrary.com]

threshold value based on sequence similarity. Each sequence was treated as a query against the entire data set of identified sequences, and a species name was assigned according to three criteria as proposed by Meier et al. (2006): Best Match (BM), Best Close Match (BCM) and All Species Barcode (ASB).

2.3.3 | Barcodes verifying with extended Data set II

NJ tree constructions and k2p genetic distance calculations were carried out using MEGA, as detailed above. The kernel density estimates of intra- and interspecific k2p distances between Data set I and Data set II were plotted using ggplot2 (Wickham, 2009). Two independent samples *t* test analyses were implemented with stats package in R v3.3.1 (R Development Core Team 2016) to detect the differences between the mean k2p distance of Data sets I and II. For all statistical analyses, differences were considered to be significant when *p* values were lower than 0.01.

2.3.4 | Comprehensive haplotype-based barcode library based on Data sets II and III

Haplotypes of ITS and *trnL-trnF* sequences were defined using DnaSP v5.10 (Librado & Rozas, 2009). Species discrimination rate, including subregions for ITS (ITS1 and ITS2), was then visualized using a NJ tree. Genetic diversity indices H_d and π of *trnL-trnF* for each species were calculated in DnaSP.

2.3.5 | Species distribution modelling (SDM)

To predict the potential current distribution ranges of yews, SDM was performed for each of the 15 *Taxus* species. Georeferenced points of species occurrence data were obtained from the Global Biodiversity Information Facility (GBIF; GBIF.org), National Specimen Information Infrastructure (NSII; http://www.nsii.org.cn), literature and our field observation. To reduce potential errors in species locations, all points of occurrence for each species were carefully scrutinized using GOOGLE EARTH, and duplicate locations, or those that appeared to be wrong, were removed. To mitigate the sampling bias, occurrence data were further adjusted by a spatial filtering method (Kramer-Schadt et al., 2013). Because questionable taxonomic delimitation can often decrease the accuracy of models (Bittencourt-Silva et al., 2017), the occurrence points for each species were based on the most up-to-date taxonomic studies (Farjon, 2010; Farjon & Filer, 2013; Möller et al., 2013) and further adjusted according to their phylogeographical patterns (Gao et al., 2007; Liu et al., 2013; Mayol et al., 2015; Poudel, Möller, Liu et al., 2014). Finally, a total of 1186 occurrence points remained, with the fewest (seven) for *T. floridana* and the most (223) for *T. baccata* (Table 1).

Nineteen BIOCLIM variables from the WorldClim (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) were first examined for multicollinearity, excluding those with a Pearson's correlation $r > 0.8$ with another variable. Five to eleven bioclimatic variables were used for each species in the downstream SDM analysis (Supporting information: Tables S4, S5). To eliminate the impact of background geographical extent of the models on modelling results (Merow, Smith, & Silander, 2013), we limited our model extent to the distributional range of each *Taxus* species with a buffered zone. SDMs were generated at 30-s resolution using the Maxent v3.2 software package (Phillips, Anderson, & Schapire, 2006). The specific thresholds of modelling results were selected using the sensitivity–specificity equality approach (Liu, Berry, Dawson, & Pearson, 2005). To reflect the realized distribution range, any region where there was no possibility of distribution, such as large water bodies, was removed for each species.

2.3.6 | Forensic applications

Our laboratory received several different unknown biological samples submitted by various organizations which were suspected to be of *Taxus*. From these, three were chosen for further investigation.

Unknown X1 was bark powder confiscated by wildlife police from a suspect's lorry, following suspicion that a wildlife crime had been committed. Unknown X2 was a seedling which was assumed to be associated with fraudulent trading between two companies, provided by one of the Judicial Expertise Centers of Yunnan province, China. Unknown X3 was a specimen supplied by a company wishing to develop industrial *Taxus* cultivation in order to produce taxol. In all three cases, the possible origin of the samples was unknown to the investigating officers, and morphological identification was not possible. These scenarios provided, therefore, real tests of the identifying power of DNA barcoding in identifying unknown samples where no other identification method was possible.

3 | RESULTS

3.1 | Barcode universality and sequence characteristics

All five barcodes were successfully amplified and sequenced for all 72 individuals in Data set I (Supporting information: Table S6, q.v. for sequence characteristics), providing 360 sequences in total for DNA barcode evaluation. Alignment length varied from 702 to 1,360 bp across the five regions, and the full concatenated length was 4,868 bp. Intraspecific and interspecific K2p distances were highest for ITS (0–0.0054 and 0–0.0220, respectively) and lowest for *rbcL* (0 and 0–0.0058, respectively).

3.2 | Species discrimination efficiency: comparing methods and barcodes

Species discrimination rates varied according to both the analysis methods and the barcode combinations used (Figure 2). Among single barcodes or their combinations, the NJ tree-based method provided the highest species discrimination, followed by ML, ABGD and finally PTP as the least effective (Supporting information: Table S7).

For the sequence similarity method of BM, BCM and ASB analysis, the proportion of individuals that could be identified correctly depended on which barcodes, or combinations thereof, were used (Supporting information: Table S8). The proportions of correct identifications were equal between BM and BCM, ranging from 34.72% to 100% for both depending on barcodes; but ASB had the most favourable range, from 70.83% to 100%.

Species discriminatory ability was compared across barcodes using the NJ method (Figure 2, Table 2; Supporting information: Table S7). Of the five single barcodes, *trnL-trnF* showed the highest discriminatory power with 93% (i.e., all species except *T. floridana*), followed by ITS (73%), ITS1 (67%), *psbA-trnH* (53%), *matK* (47%) and finally ITS2 and *rbcL* (33%). Among the two-marker combinations, *trnL-trnF* + ITS provided the highest species resolution (93%), while other combinations ranged from 53% to 87%. Curiously, *T. mairei* failed to be discriminated when *trnL-trnF* was combined with any other cpDNA marker. Leaving aside ITS1 or ITS2 as separate markers, combinations of three or four markers that excluded *trnL-trnF* always had a discrimination rate of 87%, whereas any combination of three or more markers that included *trnL-trnF* had a rate of 93%. Hence, all five markers together had a rate of 93% and gave a bootstrap value of >96% for the monophyly of every *Taxus* species except *T. floridana*, which no combination could discriminate or resolve it as monophyletic (Figure 3; Supporting information: Table S7).

3.3 | Comparisons between data sets

Data set II contained 129 more individuals and 475 more sequences than Data set I, making a total of 835 sequences, with the number of individuals sampled per barcode ranging from 110 to 195 (Supporting information: Table S1). In general, Data set II captured the same distribution range of intra- and interspecific k2p distance as Data set I (Supporting information: Figure S3). However, for each individual marker, the interspecific k2p distance was slightly but significantly larger in Data set I than Data set II (Supporting information: Figure S4b; Supporting information: Table S9). Conversely, intraspecific distance for ITS in Data set II was around twice that in Data set I, whereas there was no significant difference between data sets for any of the cpDNA markers (Supporting information: Figure S4a; Supporting information: Table S9). Species discrimination rates were the same between Data sets I and II for *rbcL*, *matK* and *psbA-trnH* (Table 2). However, using ITS, Emei type was discriminated in Data set I but not in Data set II. In both data sets, *T. globosa* could be discriminated by *trnL-trnF* alone, but strangely, *trnL-trnF* + ITS could discriminate *T. globosa* in Data set I but not in Data set II.

3.4 | Comprehensive haplotype-based DNA barcode reference libraries

For ITS, a total of 63 ITS ribotypes were obtained from 195 individuals representing 15 species from Data set II (Supporting information: Table S10). Clustering of these ribotypes resolved ten well-supported haplotype clades, each representing one species; the other five species were not discriminated (Figure 4b). ITS1 can recognize nine species, whereas only five of 15 species could be identified by the ITS2 (Table 2).

A total of 4151 sequences were obtained for *trnL-trnF* in Data set III, and 73 haplotypes were defined. Molecular genetic diversity indices N_H, N_P, H_d and π ranged from 1 to 24, 0 to 13, 0.000 to 0.781 and 0.000 to 0.156, respectively (Table 3). As with Data set II, 14 of 15 species were discriminated, though with lower bootstrap support for some clades. In Data set III, the species not discriminated was *T. mairei*, which was polyphyletic (Figure 4a; Supporting information: Figure S5). Conversely, *T. floridana* was discriminated in Data set III only, albeit with only 78% support (Table 2).

3.5 | *Global map of* Taxus

In species distribution modelling, each one of the *Taxus* species had an area under the receiver operating characteristic curve (AUC) value of ≥0.932 (Table 1), indicating a far better than random prediction.

TABLE 2 Bootstrap values of monophyletic clades of *Taxus* lineages based on single DNA loci, and one combination

	DNA regions							
	rbcL	*matK*	*psbA-trnH*	*trnL-trnF*	ITS	ITS1	ITS2	*trnL-trnF* + ITS
No. of samples	72/110	72/173	72/167	72/190/4151	72/195	72/195	72/195	72/185
Taxon								
T. baccata	65/66	64/64	n.d./n.d.	99/98/98	98/91	95/87	87/61	99/99
T. brevifolia	62/63	99/99	99/98	59/61/60	96/97	96/97	n.d./n.d.	94/95
T. calcicola	n.d./n.d.	n.d./n.d.	n.d./n.d.	68/70/58	88/48	69/n.d.	n.d./n.d.	98/92
T. chinensis	42/42	66/65	63/60	87/84/85	n.d./n.d.	n.d./n.d.	n.d/n.d.	88/80
T. canadensis	n.d./n.d.	80/79	52/46	99/99/99	99/99	85/87	94/94	99/99
T. contorta	66/66	51/51	n.d./n.d.	94/82/42	99/100	99/99	88/86	99/99
T. cuspidata	n.d./n.d.	n.d./n.d.	98/98	92/92/94	n.d./n.d.	n.d./n.d.	n.d./n.d.	99/99
Emei type	n.d./n.d.	n.d./n.d.	61/67	99/99/99	49/n.d.	53/n.d.	n.d./n.d.	99/98
T. floridana	n.d./n.d.	n.d./n.d.	n.d./n.d.	n.d./n.d./78	n.d./n.d.	n.d./n.d.	n.d./n.d.	n.d./n.d.
T. florinii	87/86	40/41	86/80	84/85/43	98/96	98/94	n.d./n.d.	99/99
T. globosa	n.d./n.d.	n.d./n.d.	n.d./n.d.	99/81/84	n.d./n.d.	n.d./n.d.	n.d./n.d.	98/n.d.
T. mairei	n.d./n.d.	n.d./n.d.	n.d./n.d.	12/n.d./n.d.	94/85	90/81	n.d./n.d.	94/85
T. phytonii	n.d./n.d.	n.d./n.d.	55/55	68/53/66	83/80	76/78	53/54	89/84
Qinling type	n.d./n.d.	48/47	n.d./n.d.	60/77/39	99/100	98/98	64/63	99/99
T. wallichiana	n.d./n.d.	n.d./n.d.	95/48	81/80/31	96/75	91/76	n.d./n.d.	99/99
Monophyly	5/5	7/7	8/8	14/13/14	11/10	10/9	5/5	14/13
Discrimination rate	33/33	47/47	53/53	93/87/93	73/66	67/60	33/33	93/87

Notes. Figures cited are for Data set I/II/III for *trnL-trnF*; Data set I/II for others. Species identification mismatches among data sets are shown in grey shade. We used 73 haplotypes of *trnL-trnF* for the NJ analysis in Data set III. n.d: indicates not discriminated.

Current potential distribution predictions were generally good representations of the actual distributions of all *Taxus* species (Figure 5; Supporting information: Figure S1). *Taxus* is most variable, diverse and complex in Asia where ten species occur, and less so in North America with four species, whereas Europe is straightforward with only *T. baccata* present (Figure 5). Distribution ranges tended to be broader for northern temperate species than for tropical/subtropical ones. Among the latter, *T. mairei* showed the largest distribution range and *T. floridana* the smallest. Most species did not have overlapping distribution ranges; however, an exception was *T. mairei*, which had large areas of sympatry with each of *T. chinensis* and *T. wallichiana* across China and the Himalaya region, respectively. However, *T. mairei* displayed elevational separation from both *T. chinensis* and *T. wallichiana* (Poudel, Möller, Liu et al., 2014), occupying lower altitudes; this was also observed in the field (Supporting information: Table S1), suggesting that each species occupies a separate ecological niche. Narrow contact areas were also revealed among the other six species in the Himalaya–Hengduan Mountains (Figure 5).

3.6 | Forensic application test

A species identification workflow of *Taxus* was established based on the above analyses (Figure 6). Using this identification protocol, DNA barcodes (*trnL-trnF* plus ITS) of all three unknowns were successfully sequenced. The sequences from unknown X1 and X3 had 100% *trnL-trnF* and ITS identity with *T. florinii and T. mairei*, respectively (Figure 4), and the samples could hence be identified as belonging to these species. Unknown X2 had 100% identical *trnL-trnF* to *T. cuspidata* (Figure 4a), but 100% identical ITS to *T. baccata* (Figure 4b). This discrepancy can be explained if the sample is a hybrid between *T. cuspidata* and *T. baccata*, known as *T.* × *media*.

4 | DISCUSSION

4.1 | DNA barcoding of *Taxus*

4.1.1 | Comparison of species identification methods

Of the data analysis methods used, sequence similarity method showed the highest discrimination between *Taxus* species (Supporting information: Table S8), while the coalescent-based PTP method consistently exhibited the lowest (Figure 2; Supporting information: Table S7), although the PTP method always recognized more species (data not shown). Tree-based methods (NJ and ML) had higher species discrimination power than distance-based ABGD, with NJ performing better than ML (Figure 2). In routine DNA-based forensic application, there will always be a trade-off between accuracy and convenience of the species identification method. The sequence similarity method has been shown to be reliable, feasible and

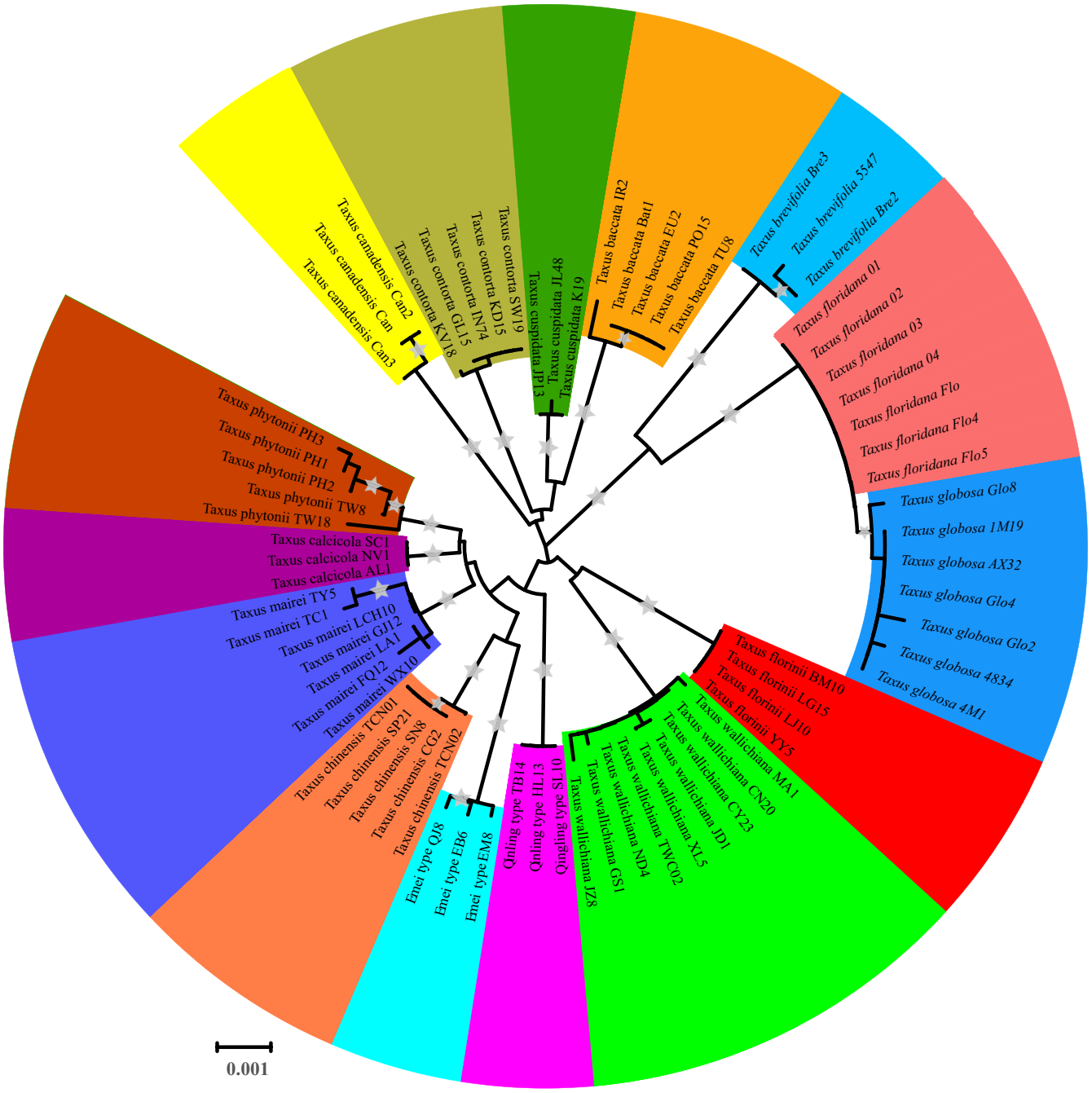

FIGURE 3 Neighbour-joining (NJ) tree of the 15 *Taxus* species in this study, based on a concatenated alignment of five barcodes (*rbcL*, *matK*, *psbA-trnH*, *trnL-trnF* and ITS) and inferred using MEGA. The tree displays 360 DNA barcodes assigned to 72 individuals from across the world. An asterisk indicates bootstrap value ≥0.85 shown for each coloured lineage. The scale bar represents base substitutions per site [Colour figure can be viewed at wileyonlinelibrary.com]

computationally tractable (Virgilio, Backeljau, Nevado, & De Meyer, 2010), but the results are often counter-intuitive; meaning that it is not easy to assign an unknown sample to a specific species. The tree-based NJ method is the most widely used method for DNA barcoding in the literature (Sandionigi et al., 2012) and has been tested and validated many times (Little & Stevenson, 2007; Sandionigi et al., 2012; van Velzen, Weitschek, Felici, & Bakker, 2012). Considering the robustness of the NJ method in the present study, as well as its popularity, rapidity and intuitiveness (van Velzen et al., 2012; Yan et al., 2015), we recommend it as a routine analysis method for DNA barcode-based identification of *Taxus* species. The discussion below therefore focuses on results from NJ analyses unless stated otherwise.

4.1.2 | Barcodes for *Taxus*

The first step for plant DNA barcoding is to find one or a suitable combination of barcodes. In *Taxus*, the discrimination power of each

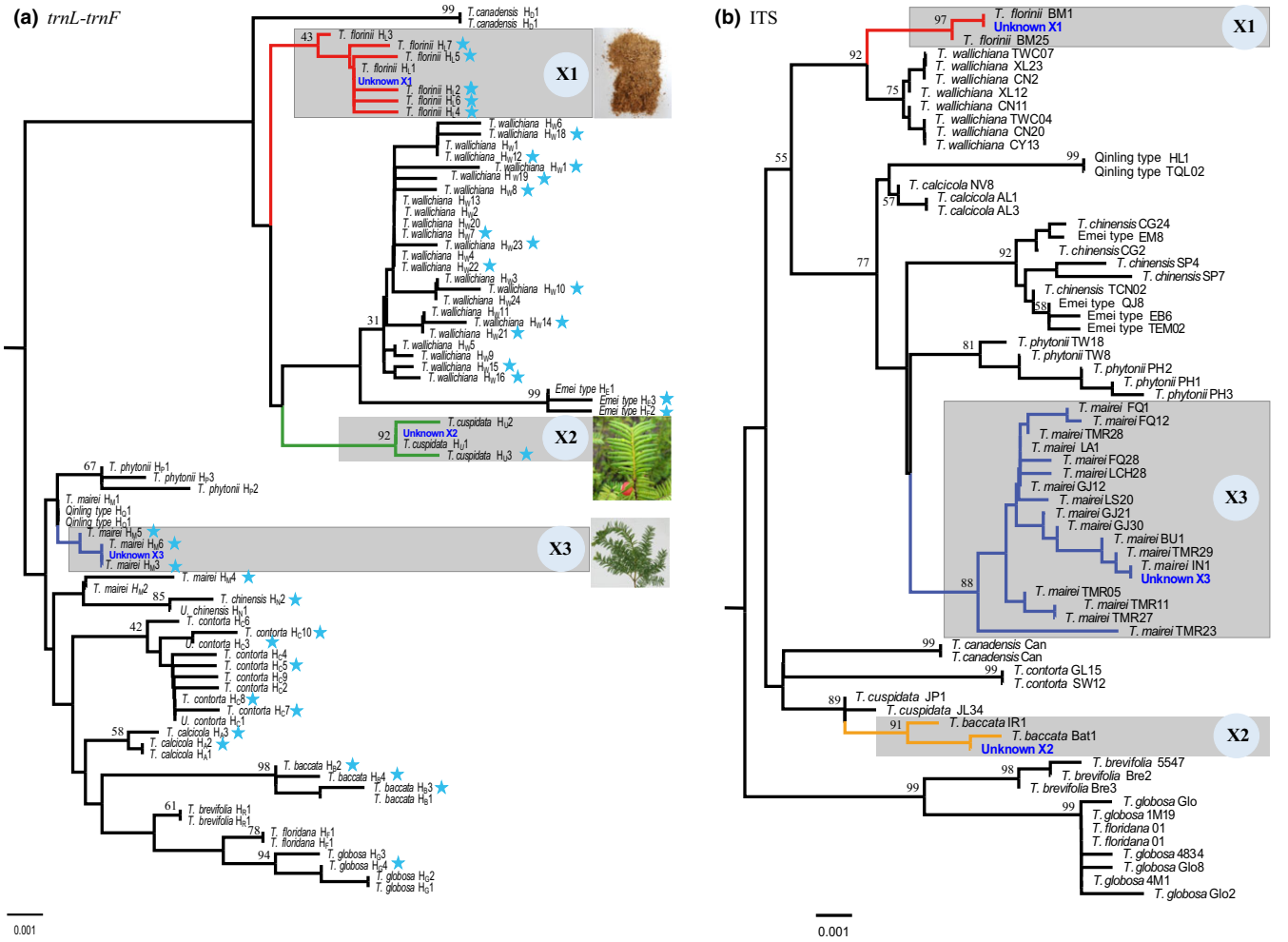

FIGURE 4 The clustering relationship of three unknown samples in the *trnL-trnF* (a) and ITS (b) NJ tree of reference haplotypes. The cyan stars in (a) indicate private *trnL-trnF* haplotypes. Clades that include "Unknowns" are highlighted with grey squares, X1, X2 and X3 represent the three unknown samples. Bootstrap values are shown on the branches [Colour figure can be viewed at wileyonlinelibrary.com]

of the five individual barcodes, and also the two subregions ITS1 and ITS2 treated separately, was assessed via the occurrence of monophyletic clades in the NJ tree. This showed that *trnL-trnF* had the highest discriminatory power (93%), followed by ITS (73%) and ITS1 (67%), while ITS2 and *rbcL* had the lowest (33%) (Figure 2; Table 2; Supporting information: Table S7). Only *T. floridana* could not be identified in the NJ analysis with the criteria used here for *trnL-trnF* (Table 2). However, the *trnL-trnF* sequence of *T. floridana* could be distinguished from its closest relative *T. globosa* by two point mutations (519 bp, T to C) and (520 bp G to A), and the position of one insertion from 531 to 540 bp, which is useful for species identification in closely related *Taxus* species (Supporting information: Figure S6). Such differences were shown in the NJ tree (Figures 3, 4a; Supporting information: Figure S5), where although the individuals from *T. floridana* did not form a monophyletic clade, they did show a consistent difference from *T. globosa*. Another case is *T. mairei*, whose six haplotypes clustered into two clades: haplotypes H_M2 and H_M4 are close to *T. chinensis*, and the rest are close to the Qinling type (Figure 4a; Supporting information: Figure S5). However, comparing with *T. mairei*, one species-specific insertion was observed from 670 to 676 bp in *T. chinensis*, and one deletion was detected from 849 to 850 bp in Qinling type (Supporting information: Figure S6) differentiating them from *T. mairei*. Thus, if we take the indel into account, *trnL-trnF* alone can distinguish all 15 species of *Taxus*, making it the ideal single barcode for *Taxus*. It meets all criteria for an ideal barcode, that is, primer universality, consistent ability to generate high-quality sequences from the target taxa, high species resolving power (Hollingsworth, Graham, & Little, 2011; Kress et al., 2005) and clear differentiation ("barcode gap") between species (Meyer & Paulay, 2005).

Combination of DNA barcodes has been proposed to increase species discrimination power (CBOL Plant Working Group 2009; Kress et al., 2005; Li et al., 2011) and has been adopted for many specific taxa (Liu et al., 2011; Yan et al., 2015). However, in *Taxus*, no combination gave a higher discrimination rate than *trnL-trnF* alone (Figure 2, Table 2; Supporting information: Table S7). Moreover, the other three cpDNA regions examined (*rbcL*, *matK* and *psbA-trnH*) did little to improve the discrimination power; even combined, their discrimination rate (86.7%) was lower than *trnL-trnF* alone (Supporting information: Table S7). Therefore, while these may have roles in

TABLE 3 Number of populations (P_N) and individuals (N) for each *Taxus* species and their genetic diversity based on the *trnL-trnF* region

Species	P_N	N	N_H	N_P	H_d	π ($\times 10^{-2}$)	Haplotypes (number of individuals), private haplotypes in bold
T. baccata	22	42	4	3	0.264	0.044	H_B1 (36); **H_B2** (3, IR2, Iran); **H_B3** (1, EU1, Scotland); **H_B4** (2, IR1, Iran)
T. brevifolia	3	3	1	0	–	–	H_R1 (3)
T. calcicola	7	109	3	2	0.139	0.004	H_A1 (101); **H_A2** (6, SC19, Sichou); **H_A3** (2, MLP15, Malipo)
T. canadensis	5	6	1	0	–	–	H_D1 (6)
T. chinensis	16	292	2	1	0.007	0.001	H_N1 (291); **H_N2** (1, CG10, Chengu)
T. contorta	19	373	10	5	0.124	0.017	H_C1 (348); H_C2 (3); **H_C3** (2, GL11, Jilong); H_C4 (3); **H_C5** (1, HZ3); H_C6 (8); **H_C7** (1, ME10); **H_C8** (1, MK11); H_C9 (4); **H_C10** (2, SW10)
T. cuspidata	22	306	3	1	0.051	0.062	H_U1 (166); H_U2 (139); **H_U3** (1, DPEL01)
Emei type	10	197	3	2	0.041	0.005	H_E1 (193); **H_E2** (1); **H_E3** (3)
T. floridana	3	7	1	0	–	–	H_F1 (7)
T. florinii	29	574	7	6	0.031	0.004	H_L1 (565); **H_L2** (2, HB02, Haba); **H_L3** (1, KPG18, Kangpu); **H_L4** (1, KPG29, Kanpu); **H_L5** (1, LJ20, Lijiang); **H_L6** (3, LJS23, Lijiang); **H_L7** (1, ML03, Meili)
T. globosa	8	10	4	1	–	–	H_G1 (4); H_G2 (2); H_G3 (3); **H_G4** (1)
T. mairei	32	739	6	4	0.494	0.064	H_M1 (443); H_M2 (284); **H_M3** (2, TC01, Tengchong); **H_M4** (6, HSH25, Huangshan); **H_M5** (2, LA19, Lianan); **H_M6** (2, TMR33, Tengchong)
T. phytonii	12	197	3	0	0.642	0.156	H_P1 (83); H_P2 (75); H_P3 (39)
Qinling type	11	278	1	0	0.007	0.001	H_Q1 (278)
T. wallichiana	52	1018	24	13	0.781	0.079	H_W1 (323); H_W2 (212); H_W3 (27); H_W4 (230); H_W5 (157); H_W6 (2); **H_W7** (1, CX01); **H_W8** (4, CY08); H_W9 (4); **H_W10** (2, CY29); H_W11 (3); **H_W12** (1, DS13); H_W13 (3); **H_W14** (4, GS13); **H_W15** (7, GS18, YG02); **H_W16** (1, JD22); **H_W17** (20, JZ01); **H_W18** (3, KC1); **H_W19** (5, MA01); H_W20 (3, XP03); **H_W21** (2, XB20); **H_W22** (1, XL03); **H_W23** (2, YG04); **H_W24** (1, YG15)
Total	**251**	**4151**	**73**	**38**			

Notes. N_H: number of total haplotypes; N_P: number of private haplotypes at population level; H_d: haplotype diversity; π: nucleotide diversity; –: not estimate due to small sample size.

determining if samples are *Taxus* or not, they are not useful at the infrageneric level.

Although ITS1 or ITS2 has been recommended as separate potential barcodes (Chen et al., 2010; Liu et al., 2011), we found that each has a lower species discrimination rate than the complete ITS (Table 2) and therefore cannot serve as alternatives to ITS. Unlike *trnL-trnF*, ITS (or ITS1 alone) consistently discriminated *T. mairei* (Table 2); however, within Data set II, *T. globosa* could be discriminated by *trnL-trnF* alone, but not by *trnL-trnF* + ITS combined. This was the effect of the incongruent *trnL-trnF* signal, which could be due to genetic introgression between species or incomplete lineage sorting. Considering that information from the nuclear DNA is essential for tracing species boundaries in DNA barcoding (Hollingsworth, Li, van der Bank, & Twyford, 2016; Li et al., 2011) and also for hybrid identification (see below), we recommend the incorporation of ITS in the panel of *Taxus* barcodes. Moreover, within-species variation might help trace the origin of an unknown sample to one part of a species' range, as demonstrated below.

4.2 | Sampling strategy in DNA barcoding

The size and extent of necessary sampling are one of the central issues in DNA barcoding (Bergsten et al., 2012; Zhang et al., 2010). In the current study, based on the *trnL-trnF* sequences among *Taxus* species, four species possessed only one haplotype, six had two to four, three had between six and 10, and *T. wallichiana* contained by far the most with 24 (Table 3). However, this species also had, by some margin, the most sampled populations (52) and individuals (1,021). As a general rule, the number of haplotypes tended to increase with the number of sampled populations and individuals, and the spatial scale of sampling, with two notable exceptions. Qinling type contained just one haplotype among 11 populations and 274 individuals analysed here, perhaps indicating a past genetic bottleneck. The European *T. baccata* had just four haplotypes detected among 22 samples from across Europe and into SW Asia and N Africa; moreover, two of these were private to Iran, leaving only two detected from Europe. While this could reflect few individuals sampled per population (39 sampled plants in total), it is consistent with a broad pattern of lower haplotype diversity in Europe likely resulting from a range contraction during Pleistocene ice ages (Hewitt, 2000). Therefore, the diversity of detected haplotypes for each species is likely to depend both on their Pleistocene population histories (Liu et al., 2013; Mayol et al., 2015; Poudel, Möller, Liu et al., 2014) and the breadth of sampling. Nonetheless, the possible existence of undetected haplotypes for less sampled species should not be an issue given that all but one species were resolved as monophyletic for haplotypes. The exception, *T. mairei*, was monophyletic but with extremely weak (12%) support only in Data set I, but not in II or III.

Due to access limitations, the sampling size was small for the four North American species (Table 3) and hence is unlikely to cover

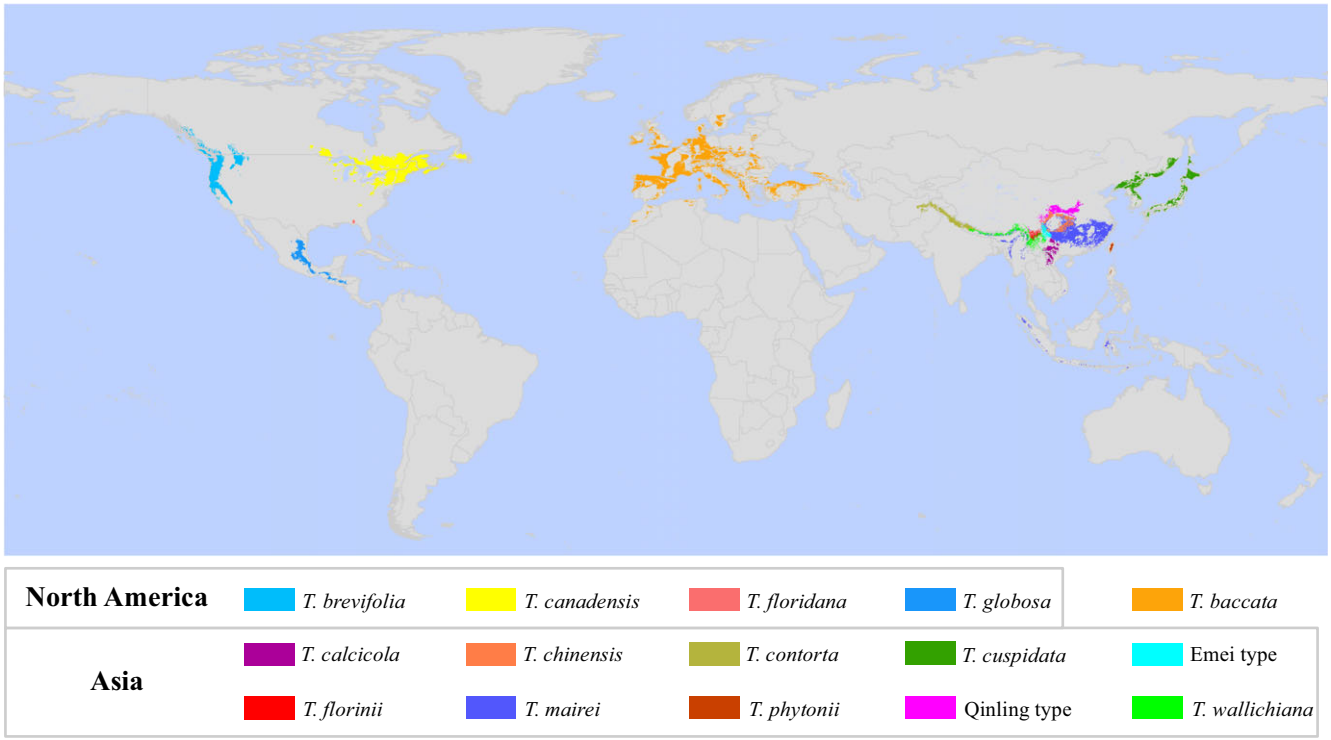

FIGURE 5 The potential global geographical distribution of 15 *Taxus* species predicted using species distribution modelling [Colour figure can be viewed at wileyonlinelibrary.com]

all the genetic variation of these species based on *trnL-trnF*. However, the four species showed a clear allopatric distribution (Figure 5; Supporting information: Figure S1), which is strongly determined by climatic factors, and the species may have different ecological niches (Farjon & Filer, 2013). Moreover, the phylogeographical history of North American plants was largely affected by the Quaternary glaciations (Avise, 2000). These factors would together accelerate lineage sorting and genetic divergence processes among the four species, increasing interspecific genetic distance ("barcoding gap") at the expense of intraspecific distance (Meyer & Paulay, 2005). Nevertheless, given the low number of samples from North America, more sampling is needed for *T. globosa* to further validate the robustness of the library.

For ITS, our results suggested that increasing the sample size from 72 individuals (Data set I) to 195 (Data set II) greatly increased the intraspecific k2p distance (Supporting information: Figure S4a), implying that more individuals are needed to capture available genetic variation. However, Emei type could be discriminated in Data set I but not in II for ITS, and a similar decrease in species identification rate was also observed in ITS1 for both Emei type and *T. calcicola*. The same also applied for *T. globosa* with ITS + *trnL-trnF* combination (Table 2). Likewise, based on a total of 47 samples, ITS or ITS1 had previously distinguished all 11 Eurasian species (Liu et al., 2011), whereas in the current study based on 195 individuals, four of these (Emei type, *T. calcicola*, *T. chinensis* and *T. cuspidata*) could not be discriminated by either ITS1 or complete ITS. This strongly indicates that the relatively intense within-taxon sampling of the current study reduced species-level resolution from ITS, presumably because it captured a higher level of within-species variation, reducing interspecific distances (Bergsten et al., 2012; Zhang et al., 2010). Introgression between species might have contributed to this effect, because shared or transferred ribotypes, for example, between Emei type and *T. chinensis*, are more likely to be detected as sampling size increases. Moreover, with an increase in species and population sampling, the strong reduction in species resolution rate decrease in ITS may also reflect the effective population size difference between chloroplast and nuclear genome (Birky, Maruyama, & Fuerst, 1983).

When we consider all the above evidence, together with previous work (Bergsten et al., 2012; Ekrem et al., 2007; Liu et al. 2012; Zhang et al., 2010), the main implications for DNA barcode library construction, for other groups that include IUCN- and/or CITES-listed taxa, include the following: First, the potential utility of a DNA barcode is closely associated with sampling completeness at the taxonomic level. As the number of included species increases, the discrimination rate may decrease (see ITS for *Taxus* here). Hence, comprehensive species-level sampling is one of the prerequisites for developing a reliable DNA barcode library. Second, the exact number of samples needed within any given species will largely depend on its population history and geographical extent, which determine its intraspecific genetic diversity. Generally, sampling of multiple individuals from several localities covering the entire distribution range is recommended. Third, because the effective population size varied between chloroplast and nuclear regions (Birky et al., 1983), the appropriate sampling size for developing a barcode library to represent these two genomes is different, and at least in this case, the latter needed more sampling to capture complete genetic variation.

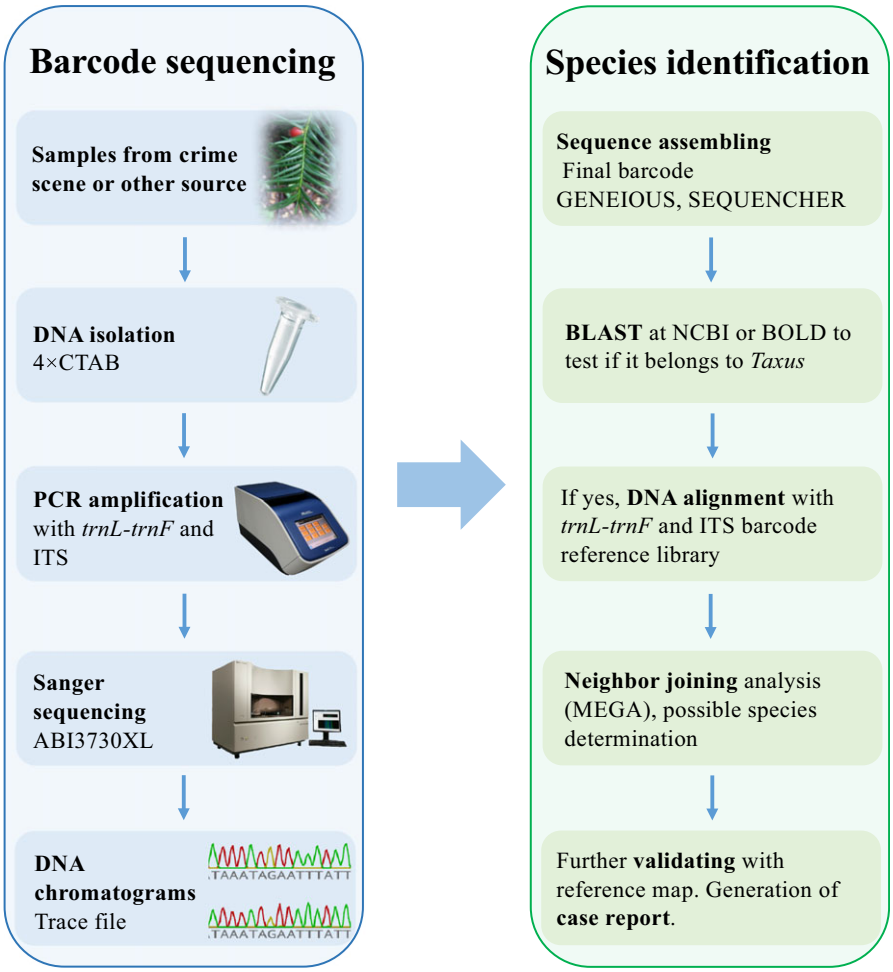

FIGURE 6 The applicable workflow for species identification of unknown *Taxus* samples using the DNA barcode libraries and reference map generated in this study [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | Forensic test cases: geographical origin inference and hybrid identification

Geographical structuring of genetic variation results from mutation, genetic drift, limitations to gene flow and selection (Avise, 2000); this offers the possibility of determining the region or even population of origin for a sample, for which DNA barcoding is an appropriate tool (Moritz & Cicero, 2004; Ogden & Linacre, 2015). Such knowledge can be useful in forensic identification and conservation, providing vital information for investigating wildlife crime and securing convictions.

Locating the origin of any *Taxus* sample can first be narrowed down by species identity, using distribution maps (Figure 5), as only three species exhibit significant overlap in distribution range (see above). Within-species variation for *trnL-trnF* was detected for 10 of the 15 *Taxus* species, including nine of 11 Eurasian species, where some of the haplotypes are private at population level (Figure 4, Table 3). For ITS, intraspecific variation was observed in 13 of the 15 species. Hence, in many, but not all, cases, the point of origin could be further narrowed down using within-species variation.

These principles were tested on three unknown samples. Unknown X1 was seized in Deqin County, northwest Yunnan Province of China, and the sample was identified by DNA barcoding as *T. florinii*, which is consistent with the *T. florinii* distribution map (Figure 5). In this instance, the sample's haplotype (H_L1) is widespread within the species and did not allow us to further narrow down its location.

Unknown X3 had *trnL-trnF* haplotype H_M6 (Supporting information: Figure S6), which is private to a single population of *T. mairei* sampled from Tenchong in the west Yunnan (Table 3; Supporting information: Table S1). Given that 31 other populations were sampled across the species' range (Table 3; Supporting information: Table S1), it is thus highly likely that this sample originated either from this population or from another nearby population that was not sampled. By contrast, the species *T. mairei* occurs in at least 12 Chinese provinces (Supporting information: Table S1). Hence in this instance, within-species variation for a single marker has narrowed down from a broad species range to a small and fairly precise point of origin.

Clearly, it is a matter of chance whether a private haplotype is present in any unknown sample; many species contain both private and widespread haplotypes (e.g., *T. wallichiana*), while *T. chinensis* and Qinling type have only one haplotype. Therefore, there will be some cases where determining location of origin requires additional markers, such as microsatellites or a fast-evolving DNA region (Ogden & Linacre, 2015), but that would of course incur significant extra costs. Future projects should focus on population genetic analysis with reproducible molecular markers that can better represent variation between parts of a species' distribution range.

Hybridization can often reduce the success of species discrimination from barcodes in plants (Hollingsworth et al., 2011; Tosh et al., 2016); thus, hybrids are usually excluded in DNA barcoding analyses (Meyer & Paulay, 2005). However, identifying hybrids is important, as for example they are often exempted from legislation (Dormontt et al., 2015). Because plastid inheritance is uniparental (paternal in *Taxus*; Collins, Mill, & Möller, 2003), a hybrid sample cannot be identified from cpDNA markers alone; thus, nuclear markers must also be used, as these are biparentally inherited (Hollingsworth et al., 2016; Li et al., 2011). The third unknown sample of *Taxus* from the current study, Unknown X2, had cpDNA from *T. cuspidata*, but clustered with *T. baccata* for the nuclear ITS data, suggesting that the sample is a hybrid between the two species. Two artificial *Taxus* hybrids, *T.* × *media* and *T.* × *hunnewelliana*, are widely cultivated (Hoffman, 2004). Based on RAPD markers and *trnL-trnF* sequence data, the former represents a cross between *T. baccata* × *T. cuspidata*, whereas the latter involves *T. cuspidata* and *T. canadensis* (Collins et al., 2003). Unknown X2, therefore, matches *T.* × *media*.

4.4 | Routine applications

To operationalize the application component of the DNA barcode reference library, we propose a workflow for the identification of unknown specimens (Figure 6). In routine forensic application of DNA-based methods, it is important to generate accurate and reproducible DNA sequencing results from suspect samples (Dormontt et al., 2015). In *Taxus*, the main illegal commercial products are leaves and bark, used for isolation of taxol and its derivatives, but these may also be processed into other forms, for example, powder. DNA isolation from these samples is straightforward, as with sample Unknown X1. However, illegal wood products such as chopping boards, chopsticks and bowls are also commonly found on the market in China, and DNA extracted from wood is generally of poor quality (Dormontt et al., 2015). This challenge, notwithstanding, advances in DNA isolation procedures from wood (Rachmayanti, Leinemann, Gailing, & Finkeldey, 2009), and the decreasing cost of next-generation sequencing raises the chances of usable DNA sequences being generated from *Taxus* timber.

Where the genus of a sample is not known, the popular barcodes *rbcL* and *matK*, together with ITS and *trnL-trnF*, can be used to confirm the identity to genus level, as part of a comprehensive tiered or hierarchical approach from unknown to genus to species to population. This approach follows the principle that DNA barcoding should establish a database centred on standardized barcodes with a solid taxonomic foundation, including adequate sampling of genetic variation linked to accurately verified voucher specimens (Moritz & Cicero, 2004) that can identify any plant (CBOL Plant Working Group 2009; Hollingsworth et al., 2011; Kress et al., 2005). A search tool in public databases (e.g., GenBank and BOLD) could be used to confirm whether a sample belongs to *Taxus*. However, inaccurate identifications and uneven quality of sequences deposited in GenBank are not uncommon (Nilsson et al., 2006). BOLD data are better curated with higher quality standards, but might still harbour misidentified specimens to some degree (Nilsson et al., 2006) and have a narrower coverage of plant taxa and specific barcodes. For *Taxus*, BOLD (accessed 23 April 2018) has 429 public specimen records representing 12 *Taxus* species (including some synonyms), plus two hybrids and four varieties, of which 416 are mined from GenBank, NCBI; all sequences are *rbcL*, *matK* or ITS2 sequences. If using only BOLD or GenBank sequences for barcoding reference, the potential exists for these to cause confusion or even incorrect identification according to our results, but the availability of our reference library should fix this problem. Nevertheless, it is still feasible to use GenBank and BOLD to determine whether the unknown belongs to *Taxus*.

Once a sample is certainly known to be *Taxus*, only *trnL-trnF* and ITS need to be used. If the *trnL-trnF* haplotype does not immediately match one of the 73 species-specific haplotypes detected here, NJ analysis can be used to place them within a species, which may be further supported by ITS, and the distribution map can be used to further collaborate the result.

5 | CONCLUSIONS

In the present study, three data sets, with a total of 4,151 individuals representing all the 15 currently known *Taxus* species worldwide, were used to determine the ideal DNA barcode and construct a species identification system. Five data analysis methods (sequence similarity method, PTP, NJ, ML and ABGD) were tested for species discrimination power, and based on our results, we recommend the tree-based NJ method for adoption as the standard method for forensic identification. Based on the performance of single barcodes and their combinations, we recommend *trnL-trnF* as the best single DNA barcode for *Taxus* and *trnL-trnF* + ITS as the best combined barcode. By comparing three data sets, the results indicate that the success of a DNA barcode library construction depends on adequate sampling of species within the genus, and both populations and individuals within each species across its distribution range. Moreover, the level of sampling required for adequate coverage may differ between chloroplast and nuclear barcode markers. This study has constructed a comprehensive DNA barcode reference library based on *trnL-trnF* and ITS for *Taxus* across the world, plus a global distribution map for the genus. Together, these form a standard identification system that will aid species identification for unknown *Taxus* samples. The identification system developed here successfully identified two unknown forensic samples to the species level, pinpointing the location for one of them, and identified both parents of a third unknown sample that was apparently of hybrid origin. Therefore, this system can determine both species and hybrids and can in some cases greatly narrow down the geographical location. Our work will serve as an effective tool for species identification for IUCN- and CITES-listed species. This will in turn reinforce the objectives of international treaties, strengthen national forestry management and help enforce conservation laws designed to curb

the increasing threat of illegal exploration and illicit trade, all of which will partly mitigate the extinction risk of species.

ACKNOWLEDGEMENTS

We are grateful to Dr. Zeng-Yuan Wu, Dr. Jim Provan, Dr. De-Quan Zhang, Xue-Wen Liu, Chao-Nan Fu and other colleagues for collecting samples, laboratory work and data analysis. We thank Dr. Jeremy deWaard and three anonymous reviewers for valuable comments and insights. This study was supported by the National Key Basic Research Program of China (2014CB954100), the National Natural Science Foundation of China (41571059, 31200182 and 31370252), the Interdisciplinary Research Project of Kunming Institute of Botany (KIB2017003) and the Ministry of Science and Technology, China, Basic Research Project (2013FY112600). Jie Liu was supported by the China Scholarship Council for one-year study at the Aberystwyth University, UK. Laboratory work was performed at the Laboratory of Molecular Biology at the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences. The Royal Botanic Garden Edinburgh is supported by the Rural and Environment Research and Analysis Directorate (RERAD).

AUTHORS CONTRIBUTION

Funding was obtained by J.L., L.M.G. and D.Z.L.; research was conceived and designed by J.L., L.M.G. and D.Z.L.; the samples were collected by J.L., L.M.G., M.M., D.Z.L. and J.N.W.; advise on study design and data analysis was provided by R.M. and M.M.; the molecular laboratory work was carried out and data were analysed by J.L., G.F.Z., L.J.Y., Y.H.L., J.B.Y. and J.N.W.; the first draft of the manuscript with critical input from R.M. was written by J.L.; and revisions were performed by all authors.

DATA ACCESSIBILITY

All DNA sequences have been deposited in GenBank (Supporting information: Tables S1, S3), and the haplotype-based *trnL-trnF* and ITS DNA reference libraries together with unknowns' identification matrices were deposited at the website (https://pan.baidu.com/s/1KfODaFF1WVoR9IMwFSBVzQ).

ORCID

Jie Liu http://orcid.org/0000-0003-4356-1943
De-Zhu Li http://orcid.org/0000-0002-4990-724X
Lian-Ming Gao http://orcid.org/0000-0001-9047-2658

REFERENCES

Avise, J. C. (2000). *Phylogeography: The history and formation of species*. Cambridge, MA: Harvard University Press.

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*(5), 851–869. https://doi.org/10.1093/sysbio/sys037

Birky, C. W., Maruyama, T., & Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. *Genetics*, *103*(3), 513–527.

Bittencourt-Silva, G. B., Lawson, L. P., Tolley, K. A., Portik, D. M., Barratt, C. D., Nagel, P., & Loader, S. P. (2017). Impact of species delimitation and sampling on niche models and phylogeographical inference: A case study of the East African reed frog *Hyperolius substriatus* Ahl, 1931. *Molecular Phylogenetics and Evolution*, *114*, 261–270. https://doi.org/10.1016/j.ympev.2017.06.022

Blagoev, G. A., deWaard, J. R., Ratnasingham, S., deWaard, S. L., Lu, L., Robertson, J., … Hebert, P. D. (2016). Untangling taxonomy: A DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources*, *16*(1), 325–341. https://doi.org/10.1111/1755-0998.12444

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *359*(1444), 669–679. https://doi.org/10.1098/rstb.2003.1447

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*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*(5), e1400253. https://doi.org/10.1126/sciadv.1400253

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Chen, S. L., Yao, H., Han, J. P., Liu, C., Song, J. Y., Shi, L. C., … Leon, C. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, *5*(1), e8613. https://doi.org/10.1371/journal.pone.0008613

CITES (2007) Checklist of CITES species. Retrieved from http://checklist.cites.org/#/en/search/output_layout=alphabetical&level_of_listing=0&show_synonyms=1&show_author=1&show_english=1&show_spanish=1&show_french=1&scientific_name=taxus&page=1&per_page=20

Collins, D., Mill, R. R., & Möller, M. (2003). Species separation of *Taxus baccata*, *T. canadensis*, and *T. cuspidata* (Taxaceae) and origins of their reputed hybrids inferred from RAPD and cpDNA data. *American Journal of Botany*, *90*(2), 175–182. https://doi.org/10.3732/ajb.90.2.175

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). JMODELTEST 2: More models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 772. https://doi.org/10.1038/nmeth.2109

deWaard, J. R., Hebert, P. D. N., & Humble, L. M. (2011). A comprehensive DNA barcode library for the looper moths (Lepidoptera: Geometridae) of British Columbia, Canada. *PLoS ONE*, *6*(3), e18290. https://doi.org/10.1371/journal.pone.0018290

Dormontt, E. E., Boner, M., Braun, B., Breulmann, G., Degen, B., Espinoza, E., … Lowe, A. J. (2015). Forensic timber identification: It's time to integrate disciplines to combat illegal logging. *Biological Conservation*, *191*, 790–798. https://doi.org/10.1016/j.biocon.2015.06.038

Ekrem, T., Willassen, E., & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, *43*(2), 530–542. https://doi.org/10.1016/j.ympev.2006.11.021

Farjon, A. (2010). *A handbook of the world's conifers*. Leiden, The Netherlands: Brill.

Farjon, A., & Filer, D. (2013). *An atlas of the world's conifers: An analysis of their distribution, biogeography, diversity and conservation status*. Leiden, The Netherlands: Brill.

Ferri, G., Corradini, B., Ferrari, F., Santunione, A. L., Palazzoli, F., & Alu', M. (2015). Forensic botany II, DNA barcode for land plants: Which markers after the international agreement? *Forensic Science International: Genetics*, *15*, 131–136. https://doi.org/10.1016/j.fsigen.2014.10.005

Fu, L. G., Li, N., & Mill, R. R. (1999). Taxaceae. In Z. Y. Wu, & R. H. Peter (Eds.), *Floral of China* (pp. 89–96). Beijing, and Missouri Botanical Garden Press, St. Louis, MO: Science Press.

Gao, L. M., Möller, M., Zhang, X. M., Hollingsworth, M. L., Liu, J., Mill, R. R., … Li, D. Z. (2007). High variation and strong phylogeographic pattern among cpDNA haplotypes in *Taxus wallichiana* (Taxaceae) in China and North Vietnam. *Molecular Ecology*, *16*(22), 4684–4698. https://doi.org/10.1111/j.1365-294X.2007.03537.x

Godfray, H. C. J. (2002). Challenges for taxonomy. *Nature*, *417*(6884), 17–19. https://doi.org/10.1038/417017a

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hewitt, G. M. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, *405*(6789), 907–913.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1016/j.ympev.2006.11.021

Hoffman, M. H. A. (2004) Cultivar classification of *Taxus* L. (Taxaceae). In: *Fourth International Symposium on Taxonomy of Cultivated Plants*, pp. 91–96.

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE*, *6*(5), e19254. https://doi.org/10.1371/journal.pone.0019254

Hollingsworth, P. M., Li, D. Z., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150338. https://doi.org/10.1098/rstb.2015.0338

Hou, G., Chen, W. T., Lu, H. S., Cheng, F., & Xie, S. G. (2018). Developing a DNA barcode library for perciform fishes in the South China Sea: Species identification, accuracy and cryptic diversity. *Molecular Ecology Resources*, *18*(1), 137–146. https://doi.org/10.1111/1755-0998.12718

Itokawa, H., & Lee, K.-H. (2003). Taxus: The genus *Taxus*. In R. Hardman (Ed.), *Medicinal and aromatic plants – industrial profiles*. New York, NY: Taylor & Francis.

IUCN (2017) The IUCN red list of threatened species. Retrieved from http://www.iucnredlist.org. Version 2017-2

Joly, S., Davies, T. J., Archambault, A., Bruneau, A., Derry, A., Kembel, S. W., … Wheeler, T. A. (2014). Ecology in the age of DNA barcoding: The resource, the promise, and the challenges ahead. *Molecular Ecology Resources*, *14*(2), 221–232. https://doi.org/10.1111/1755-0998.12173

Kozyrenko, M. M., Artyukova, E. V., & Chubar, E. A. (2017). Genetic diversity and population structure of *Taxus cuspidata* Sieb. et Zucc. ex Endl. (Taxaceae) in Russia according to data of the nucleotide polymorphism of intergenic spacers of the chloroplast genome. *Russian Journal of Genetics*, *53*(8), 865–874. https://doi.org/10.1134/s1022795417070079

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., … Wilting, A. (2013). The importance of correcting for sampling bias in MAXENT species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Kress, W. J. (2017). Plant DNA barcodes: Applications today and in the future. *Journal of Systematics and Evolution*, *55*(4), 291–307. https://doi.org/10.1111/jse.12254

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*(23), 8369–8374. https://doi.org/10.1073/pnas.0503123102

Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., Liu, J. Q., … Duan, G. W. (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*(49), 19641–19646. https://doi.org/10.1073/pnas.1104551108

Librado, P., & Rozas, J. (2009). DNASP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*(11), 1451–1452. https://doi.org/10.1093/bioinformatics/btp187

Little, D. P., & Stevenson, D. W. (2007). A comparison of algorithms for the identification of specimens using DNA barcodes: Examples from gymnosperms. *Cladistics*, *23*(1), 1–21. https://doi.org/10.1111/j.1096-0031.2006.00126.x

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*(3), 385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x

Liu, J., & Gao, L. M. (2011). Comparative analysis of three different methods of total DNA extraction used in *Taxus*. *Guihaia*, *31*(2), 244–249. https://doi.org/10.3969/i.issn.1000-3142.2011.03.020

Liu, J., Möller, M., Gao, L. M., Zhang, D. Q., & Li, D. Z. (2011). DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. *Molecular Ecology Resources*, *11*(1), 89–100. https://doi.org/10.1111/j.1755-0998.2010.02907.x

Liu, J., Möller, M., Provan, J., Gao, L. M., Poudel, R. C., & Li, D. Z. (2013). Geological and ecological factors drive cryptic speciation of yews in a biodiversity hotspot. *New Phytologist*, *199*(4), 1093–1108. https://doi.org/10.1111/nph.12336

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*(7), 8740–8751. https://doi.org/10.3390/ijms13078740

Mayol, M., Riba, M., Gonzalez-Martinez, S. C., Bagnoli, F., de Beaulieu, J. L., Berganzo, E., … Vendramin, G. G. (2015). Adapting through glacial cycles: Insights from a long-lived tree (*Taxus baccata*). *New Phytologist*, *208*(3), 973–986. https://doi.org/10.1111/nph.13496

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, *55*(5), 715–728. https://doi.org/10.1080/10635150600969864

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MAXENT for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, *3*(12), 2229–2238. https://doi.org/10.1371/journal.pbio.0030422

Möller, M., Gao, L. M., Mill, R. R., Liu, J., Zhang, D. Q., Poudel, R. C., & Li, D. Z. (2013). A multidisciplinary approach reveals hidden taxonomic diversity in the morphologically challenging *Taxus wallichiana* complex. *Taxon*, *62*(6), 1161–1177. https://doi.org/10.12705/626.9

Morinière, J., Hendrich, L., Balke, M., Beermann, A. J., König, T., Hess, M., … Haszprunar, G. (2017). A DNA barcode library for Germany's mayflies, stoneflies and caddisflies (Ephemeroptera, Plecoptera and Trichoptera). *Molecular Ecology Resources*, *17*(6), 1293–1307. https://doi.org/10.1111/1755-0998.12683

Moritz, C., & Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLoS Biology*, *2*(10), 1529–1531. https://doi.org/10.1371/journal.pbio.0020354

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K. H., & Koljalg, U. (2006). Taxonomic reliability of DNA sequences in

public sequence databases: A fungal perspective. *PLoS ONE*, *1*(1), e59. https://doi.org/10.1371/journal.pone.0000059

Ogden, R., & Linacre, A. (2015). Wildlife forensic science: A review of genetic geographic origin assignment. *Forensic Science International: Genetics*, *18*, 152–159. https://doi.org/10.1016/j.fsigen.2015.02.008

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Poudel, R. C., Möller, M., Gao, L. M., Ahrends, A., Baral, S. R., Liu, J., … Li, D. Z. (2012). Using morphological, molecular and climatic data to delimitate yews along the Hindu Kush-Himalaya and adjacent regions. *PLoS ONE*, *7*(10), e46873. https://doi.org/10.1371/journal.pone.0046873

Poudel, R. C., Möller, M., Li, D. Z., Shah, A., & Gao, L. M. (2014). Genetic diversity, demographical history and conservation aspects of the endangered yew tree *Taxus contorta* (syn. *Taxus fuana*) in Pakistan. *Tree Genetics & Genomes*, *10*(3), 653–665. https://doi.org/10.1007/s11295-014-0711-7

Poudel, R. C., Möller, M., Liu, J., Gao, L. M., Baral, S. R., & Li, D. Z. (2014). Low genetic diversity and high inbreeding of the endangered yews in Central Himalaya: Implications for conservation of their highly fragmented populations. *Diversity and Distributions*, *20*(11), 1270–1284. https://doi.org/10.1111/ddi.12237

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), 1864–1877. https://doi.org/10.1111/j.1365-294X.2011.05239.x

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rachmat, H. H., Subiakto, A., & Kamiya, K. (2016). Genetic diversity and conservation strategy considerations for highly valuable medicinal tree of *Taxus sumatrana* in Indonesia. *Biodiversitas Journal of Biological Diversity*, *17*(2), 487–491. https://doi.org/10.13057/biodiv/d170213

Rachmayanti, Y., Leinemann, L., Gailing, O., & Finkeldey, R. (2009). DNA from processed and unprocessed wood: Factors influencing the isolation success. *Forensic Science International: Genetics*, *3*(3), 185–192. https://doi.org/10.1016/j.fsigen.2009.01.002

Sajwan, B. S., & Prakash, K. C. (2007). Conservation of medicinal plants: Conventional and contemporary strategies, regulations and executions. *Indian Forester*, *133*(4), 484–495.

Sandionigi, A., Galimberti, A., Labra, M., Ferri, E., Panunzi, E., De Mattia, F., & Casiraghi, M. (2012). Analytical approaches for DNA barcoding data–how to find a way for plants? *Plant Biosystems*, *146*(4), 805–813. https://doi.org/10.1080/11263504.2012.740084

Schippmann, U. (2001). *Medicinal plants significant trade study*. Bonn, Germany: German Federal Agency for Nature Conservation.

Spjut, R. W. (2007). Taxonomy and nomenclature of *Taxus* (Taxaceae). *Journal of the Botanical Research Institute of Texas*, *1*(1), 203–289.

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAXML web servers. *Systematic Biology*, *57*(5), 758–771. https://doi.org/10.1080/10635150802429642

State Forestry Administration and Ministry of Agriculture P.R. China (1999) List of national key protected wild species of China. Retrieved from http://www.gov.cn/gongbao/content/2000/content_60072.htm

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*(10), 2731–2739. https://doi.org/10.1093/molbev/msr121

Tang, W. J. (2010). Chenzhou Intermediate People's Court: "New series of cases", 34 people were sentenced to imprisonment. *China Trial*, *10*(51).

Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. (2017). Future threats to biodiversity and pathways to their prevention. *Nature*, *546*(7656), 73. https://doi.org/10.1038/nature22900

Tosh, J., James, K., Rumsey, F., Crookshank, A., Dyer, R., & Hopkins, D. (2016). Is DNA barcoding child's play? Science education and the utility of DNA barcoding for the discrimination of UK tree species. *Botanical Journal of the Linnean Society*, *181*(4), 711–722. https://doi.org/10.1111/boj.12449

Vaidya, G., Lohman, D. J., & Meier, R. (2011). SEQUENCEMATRIX: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*, *27*(2), 171–180. https://doi.org/10.1111/j.1096-0031.2010.00329.x

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*(2), 110–117. https://doi.org/10.1016/j.tree.2008.09.011

van Velzen, R., Weitschek, E., Felici, G., & Bakker, F. T. (2012). DNA barcoding of recently diverged species: Relative performance of matching methods. *PLoS ONE*, *7*(1), e30490. https://doi.org/10.1371/journal.pone.0030490

Virgilio, M., Backeljau, T., Nevado, B., & De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics*, *11*(1), 206. https://doi.org/10.1186/1471-2105-11-206

Wickham, H. (2009). GGPLOT2: *Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Yan, L. J., Liu, J., Möller, M., Zhang, L., Zhang, X. M., Li, D. Z., & Gao, L. M. (2015). DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya–Hengduan Mountains. *Molecular Ecology Resources*, *15*(4), 932–944. https://doi.org/10.1111/1755-0998.12353

Zhang, A. B., He, L. J., Crozier, R. H., Muster, C., & Zhu, C. D. (2010). Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution*, *54*(3), 1035–1039. https://doi.org/10.1016/j.ympev.2009.09.014

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, *29*(22), 2869–2876. https://doi.org/10.1093/bioinformatics/btt499

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Liu J, Milne RI, Möller M, et al. Integrating a comprehensive DNA barcode reference library with a global map of yews (*Taxus* L.) for forensic identification. *Mol Ecol Resour*. 2018;18:1115–1131. https://doi.org/10.1111/1755-0998.12903