Planta (2019) 249:1617–1625
https://doi.org/10.1007/s00425-019-03116-3

ORIGINAL ARTICLE

Machine learning approaches outperform distance- and tree-based methods for DNA barcoding of *Pterocarpus* wood

Tuo He[1,2,3,4] · Lichao Jiao[1,2] · Alex C. Wiedenhoeft[3,4,5,6] · Yafang Yin[1,2]

Received: 21 September 2018 / Accepted: 20 February 2019 / Published online: 1 March 2019

Abstract

Main conclusion **Machine-learning approaches (MLAs) for DNA barcoding outperform distance- and tree-based methods on identification accuracy and cost-effectiveness to arrive at species-level identification of wood.**

DNA barcoding is a promising tool to combat illegal logging and associated trade, and the development of reliable and efficient analytical methods is essential for its extensive application in the trade of wood and in the forensics of natural materials more broadly. In this study, 120 DNA sequences of four barcodes (ITS2, *matK*, *ndhF-rpl32,* and *rbcL*) generated in our previous study and 85 downloaded from National Center for Biotechnology Information (NCBI) were collected to establish a reference data set for six commercial *Pterocarpus* woods. MLAs (BLOG, BP-neural network, SMO and J48) were compared with distance- (TaxonDNA) and tree-based (NJ tree) methods based on identification accuracy and cost-effectiveness across these six species, and also were applied to discriminate the CITES-listed species *Pterocarpus santalinus* from its anatomically similar species *P. tinctorius* for forensic identification. MLAs provided higher identification accuracy (30.8–100%) than distance- (15.1–97.4%) and tree-based methods (11.1–87.5%), with SMO performing the best among the machine learning classifiers. The two-locus combination ITS2 + *matK* when using SMO classifier exhibited the highest resolution (100%) with the fewest barcodes for discriminating the six *Pterocarpus* species. The CITES-listed species *P. santalinus* was discriminated successfully from *P. tinctorius* using MLAs with a single barcode, *ndhF-rpl32*. This study shows that MLAs provided higher identification accuracy and cost-effectiveness for forensic application over other analytical methods in DNA barcoding of *Pterocarpus* wood.

Keywords DNA barcoding · Forensic wood identification · Identification accuracy · Machine learning approaches (MLAs) · *Pterocarpus* · SMO classifier

Abbreviations

BLOG	Barcoding with logic
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
MLAs	Machine learning approaches
NCBI	National Center for Biotechnology Information
NJ	Neighbor Joining
SMO	Sequential Minimal Optimization

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00425-019-03116-3) contains supplementary material, which is available to authorized users.

✉ Yafang Yin
yafang@caf.ac.cn

[1] Department of Wood Anatomy and Utilization, Chinese Research Institute of Wood Industry, Chinese Academy of Forestry, Beijing 100091, China

[2] Wood Collections (WOODPEDIA), Chinese Academy of Forestry, Beijing 100091, China

[3] Forest Products Laboratory, Center for Wood Anatomy Research, USDA Forest Service, Madison, WI 53726, USA

[4] Department of Botany, University of Wisconsin, Madison, WI 53706, USA

[5] Department of Forestry and National Resources, Purdue University, West Lafayette, IN 47907, USA

[6] Ciências Biológicas (Botânica), Univesidade Estadual Paulista, Botucatu, São Paulo, Brazil

Introduction

Tropical forests harbor more than half of the world's plant and wild animal species, and along with temperate forests are the world's sources of timbers, forest products, and other ecosystem services, all of which are threatened by forest loss (Saatchi et al. 2011; Lewis et al. 2015). Deforestation thus represents a massive threat to global biodiversity, especially preventable deforestation, whether due to land conversion or illegal logging. International efforts to prohibit or limit the trade of endangered species have been made to combat illegal logging and associated trade (Dormontt et al. 2015; Ng et al. 2016; Brancalion et al. 2018), usually emphasizing the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), which lists species in one of the three appendices depending on the degree of protection required. By 2017, more than 500 tree species were listed in the CITES appendices.

Among the species protected by the CITES Appendix II are *Pterocarpus santalinus* and *Pterocarpus erinaceus*, two of approximately 70 species in the pantropical genus *Pterocarpus* Jacq. (Leguminosae). In addition, a number of other species in this genus are over-harvested for their wood, which is prized for furniture, traditional medicine, and artisanal crafts (Saslis-Lagoudakis et al. 2011). The International Union for Conservation of Nature (IUCN) has listed *P. santalinus* as endangered, *P. indicus* as vulnerable, and *P. angolensis* as near threatened (IUCN 2017). Furthermore, *P. tinctorius* from Africa has extremely similar wood anatomical features to *P. santalinus*, and has been used as a substitute or adulterant for *P. santalinus* in the timber trade because of the much lower cost of *P. tinctorius*. Whether the driving force for wood identification is conservation, law enforcement, or to prevent fraud, species-level identification is critically needed, but traditional wood anatomical identification is generally only accurate to the genus level (Gasson 2011).

DNA barcoding is a molecular approach, based on one or more short genetic markers from a standard part of a genome, used to achieve species-level identifications without the need for expert taxonomic knowledge (Hebert et al. 2003; Kress et al. 2005). For land plants, the combination of two chloroplast regions *matK* and *rbcL* was recommended as the universal DNA barcodes (CBOL Plant Working Group 2009; Pang et al. 2010). In addition, the nuclear internal transcribed spacer ITS2 and the plastid *psbA-trnH* were proposed as complementary barcodes for seed plants (Chen et al. 2010; Pang et al. 2010; Yao et al. 2010). The main challenge for applying DNA barcoding to wood specimens is the extraction and isolation of quality DNA, because DNA can become severely degraded after being stored for a long time or having undergone industrial processes (Jiao et al. 2014, 2015; Wiedenhoeft 2014; Yu et al. 2017; Zeng et al. 2018). One way to overcome this problem is to choose comparatively short markers that are more likely to remain intact, and public nucleotide sequence databases provide a large number of reference sequences from which to build a sequence reference library for DNA barcoding analysis (Hajibabaei et al. 2006; Ekrema et al. 2007; Hendrich et al. 2015; Xu et al. 2015a; NCBI Resource Coordinators 2016).

Along with a comprehensive reference library, reliable, and effective analytical methods to conduct DNA barcoding analyses are crucial to the broadest implementation of DNA barcoding methods, especially in conservation and law enforcement contexts. Various analytical methods of DNA barcoding have been proposed to match DNA sequences from unidentified specimens to a reference library for species identification (Meier et al. 2006; Rach et al. 2008; Sarkar et al. 2008; Damm et al. 2010; Tanabe and Toju 2013). Distance-based methods convert DNA sequences into genetic distances, and define a similarity threshold below which a DNA barcode is assigned to a given species (Zou et al. 2011). However, as coalescent depth may vary among species, a global overlap between intra- and interspecific distances might not interfere with identification success (McArdle and Anderson 2001; Collins and Cruickshank 2012; Srivathsan and Meier 2012). Tree-based methods assign queried DNA sequences to a given species based on their membership of clades in a phylogenetic tree, nevertheless, in the situation of species-level paraphyly or incomplete lineage sorting, they may lead to incorrect or ambiguous identifications (Lowenstein et al. 2009; Yassin et al. 2010).

Machine learning aims to build algorithms that can receive input data and then conduct statistical analyses to predict an output value within an acceptable range without specific step-by-step programming (Goldberg and Holland 1988; MacLeod et al. 2010; Jordan and Mitchell 2015). For species discrimination, given a reference data set composed of DNA barcode sequences of known species (training set), MLAs can place unknown specimens (test set) into a species class present in the reference library (Bertolazzi et al. 2009), and have been applied to issues of species classification (Velzen et al. 2012; Zhang et al. 2012; Weitschek et al. 2013, 2014; Hartvig et al. 2015; Libbrecht and Noble 2015; Delgado-Serrano et al. 2016; More et al. 2016; Li et al. 2017; He et al. 2018). However, MLAs for DNA barcoding have not been widely compared to distance- and tree-based methods with regard to the criteria that are likely to influence the adoption and regular application of these techniques forensically, specifically identification accuracy and cost-effectiveness (Little and Stevenson 2007; Ross et al. 2008; Collins et al. 2012).

In this study, 120 DNA sequences of four barcodes (ITS2, *matK*, *ndhF-rpl32,* and *rbcL*) generated in our previous study and 85 downloaded from NCBI were collected to establish a reference data set for six commercial *Pterocarpus* woods. MLAs (BLOG, BP-neural network, SMO and J48) were compared with distance- and tree-based methods based on identification accuracy and cost-effectiveness across these six species, and also were applied to discriminate the CITES-listed species *P. santalinus* from its anatomically similar species *P. tinctorius* for implementation in forensic wood identification.

Materials and methods

Data set construction

A total of 120 sequences generated in our previous study (Jiao et al. 2018) and 85 sequences downloaded from NCBI representing four DNA barcodes, ITS2, *matK*, *ndhF-rpl32* and *rbcL* (Supplementary Table S1) were assembled to establish the reference data set for DNA barcoding analysis of six *Pterocarpus* timber species, i.e., *Pterocarpus angolensis* DC., *Pterocarpus indicus* Willd., *Pterocarpus macrocarpus* Kurz, *Pterocarpus santalinus* L.f., *Pterocarpus soyauxii* Taub. and *Pterocarpus tinctorius* Welw. Sequences were aligned using Clustal X2.0 (UCD Conway Institute, Dublin, Ireland) with a final manual adjustment in BioEdit v.7.0 (Ibis Therapeutics, Carlsbad, CA, USA) and then combined to multivariable combinations manually.

Among the six wood species and four barcodes, a total of 205 unique sequences were compiled, 120 from Jiao et al. 2018, and 85 from NCBI (Table 1). From these we established 162 single barcodes and 373 combined barcodes as our reference data set, and for our test data set from non-vouchered wood specimens 43 single barcodes and 96 combined barcodes.

Machine-learning analysis

DNA sequences were analyzed using MLAs of the type barcoding with logic (BLOG) software version 2.4 (National Research Council, Roman, Italy) with the following input parameters for feature extraction: a maximum number of 50 features chosen (BETA = 50), a maximum of 1000 iterations (GRASPITER = 1000), and a maximum time of 120 s for analysis (GRASPSECS = 120) (Velzen et al. 2012). BLOG reported the successful classification rates for both the training set and the test set, and generated logic rules to discriminate the six *Pterocarpus* species in terms of diagnostic characters.

BP-neural network analysis was conducted using "BarcodingR", which is an R package developed by Zhang et al. (2008, 2017). The function "*optimize.kmer*" was ran to calculate the optimal *k*-mer length for every barcode and combination, and then used to conduct the BP-based analysis using the function "bbsik" in R version 3.4.2. The confusion matrix of six *Pterocarpus* species was output showing the identification accuracy.

Table 1 Number of DNA sequences of four barcodes and their combinations in the reference and query data set

Single barcodes and combinations	Reference data set		Query data set
	NCBI	Jiao et al. (2018)	Jiao et al. (2018)
ITS2	22	18	8
matK	20	20	13
ndhF-rpl32	14	22	13
rbcL	29	17	9
ITS2 + *matK*	20	18	8
ITS2 + *ndhF-rpl32*	14	18	8
ITS2 + *rbcL*	22	17	8
matK + *ndhF-rpl32*	14	20	13
matK + *rbcL*	20	17	9
ndhF-rpl32 + *rbcL*	14	17	9
ITS2 + *matK* + *ndhF-rpl32*	14	18	8
ITS2 + *matK* + *rbcL*	20	17	8
ITS2 + *ndhF-rpl32* + *rbcL*	14	17	8
matK + *ndhF-rpl32* + *rbcL*	14	17	9
ITS2 + *matK* + *ndhF-rpl32* + *rbcL*	14	17	8
Total	535		139

Sequential Minimal Optimization (SMO), the implementation of Support Vector Machine (SVM) in Waikato Environment for Knowledge Analysis (WEKA) workbench (The University of Waikato, Hamilton, Waikato, New Zealand) was employed to analyze our DNA barcodes, using reference and query data sets for training and testing, respectively. The linear kernel was used to compute attribute weights between every two species and the number of kernel evaluations was calculated for the binary SMO. The SMO separated all the reference species in the trained model, and this model was then used to identify the query sequences.

J48, the implementation of a decision tree algorithm in WEKA, was applied to train rules based on the reference data set, and then, the resultant rules were used to identify the query sequences. For the training process, J48 chooses the attributes that can effectively split the set of training data into subsets, and then, the attributes with the highest normalized information gain are used to make the decision to separate different species (Patel and Upadhyay 2012).

The ".fasta" files of barcode sequences were converted to ".arff" format using "Fasta2Weka" programme

provided by WEKA for MLAs. The reference data sets were input as training sets, and the query data sets were input as test sets. The MLAs programmes output the identification accuracy and confusion matrix. The successful identification rates of MLAs were defined as the percentage of the correctly classified sequences in the query data set, as were distance-based and tree-based analysis.

Distance-based analysis

The Kimura two-parameter (K2P) distances between all sequence pairs were calculated in TaxonDNA 1.7.8 (National University of Singapore, Singapore), and then, the "best match" and "best close match" functions in TaxonDNA were applied to test all four barcodes and their combinations under the K2P distance model (Meier et al. 2006).

Tree-based analysis

Phylogenetic analysis was conducted in MEGA 5.05 (Pennsylvania State University, State College, PA, USA) with the p-distance model. Unrooted Neighbor Joining (NJ) trees were constructed with 1000 bootstrap replications and only clades that appeared in > 50% of the trees were retained. Specimens were considered successfully identified when the query sequence was found within a cluster consisting exclusively of two or more reference sequences (Yan et al. 2015).

Results

Identification accuracy and cost-effectiveness of different analytical methods

The identification success rates of MLAs, distance- and tree-based methods based on the four barcodes and their combinations are listed in Table 2. For BLOG analysis, the identification accuracy ranged from 30.8% (*matK*) to 100% (ITS2 + *matK* + *ndhF-rpl32* + *rbcL*), and a set of logic rules in terms of diagnostic loci were output for discrimination of six *Pterocarpus* species (Table S2). The identification success rates of BP-neural network based on four barcodes and their combinations were over 50% (Supplementary Fig. S1), of which the three-locus combination ITS2 + *matK* + *rbcL* achieved the highest accuracy (Fig. 1). SMO and J48 ran by WEKA showed similar identification accuracy for all the barcodes and combinations. The three-locus combinations ITS2 + *matK* + *ndhF-rpl32* and ITS2 + *matK* + *rbcL*, and the four-locus combination ITS2 + *matK* + *ndhF-rpl32* + *rbcL* provided the best performance (100%) for discriminating the six *Pterocarpus* species when using SMO and J48. Furthermore, the two-locus combination ITS2 + *matK* succeeded in separating six *Pterocarpus* species when running SMO programme, and the criterions for assessing of SMO classifier achieved the highest accuracy (Supplementary Fig. S2). J48 output the logic rules to decide the classifying species by the diagnostic loci in shape of flowchart-like structure (Supplementary Fig. S3), and the decision tree generated

Table 2 Identification success rates of MLAs, distance- and tree-based methods based on four barcodes and their combinations

Single barcodes and combinations	BLOG (%)	BP-neural network (%)	WEKA (%)		TaxonDNA (%)		NJ tree (%)
			SMO	J48	Best match	Best close match	
ITS2	87.5	87.5	87.5	87.5	83.3	83.3	87.5
matK	30.8	64.2	84.6	76.9	15.1	15.1	15.4
ndhF-rpl32	30.8	61.5	46.2	46.2	30.6	30.6	15.4
rbcL	55.6	54.6	55.6	44.4	18.2	18.2	11.1
ITS2 + *matK*	97.5	98.7	100	87.5	91.2	55.9	75
ITS2 + *ndhF-rpl32*	87.5	97.8	87.5	87.5	80	54.3	75
ITS2 + *rbcL*	85.7	95.2	85.7	85.7	71.4	45.7	71.4
matK + *ndhF-rpl32*	84.6	89.4	84.6	84.6	76.5	61.8	64.3
matK + *rbcL*	74.4	83.7	88.9	88.9	35	30	44.4
ndhF-rpl32 + *rbcL*	75.6	82.1	82.6	66.7	61.8	44.1	55.6
ITS2 + *matK* + *ndhF-rpl32*	87.5	97.4	100	100	97.4	97.4	85.7
ITS2 + *matK* + *rbcL*	92.7	100	100	100	95.1	90.2	87.5
ITS2 + *ndhF-rpl32* + *rbcL*	85.7	97.4	83.3	83.3	79.5	76.9	71.4
matK + *ndhF-rpl32* + *rbcL*	88.9	95.3	88.9	88.9	72.1	69.8	77.8
ITS2 + *matK* + *ndhF-rpl32* + *rbcL*	100	97.4	100	100	94.7	94.7	87.5

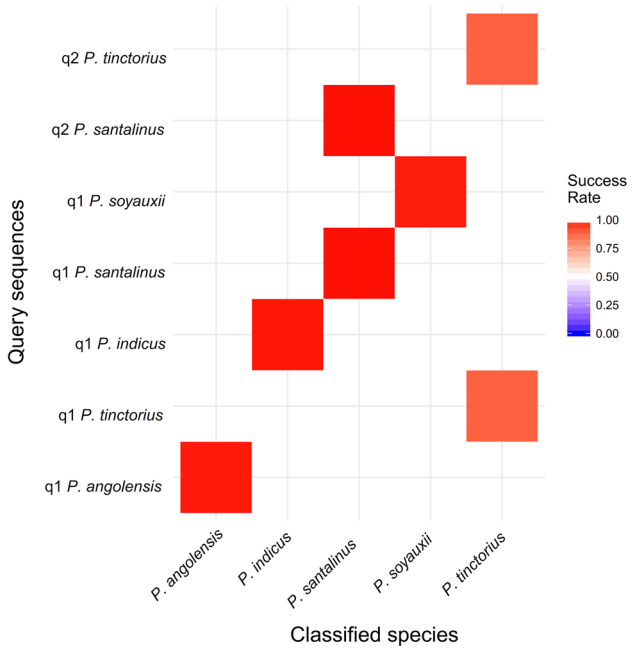

Fig. 1 Confusion matrix of BP-neural network generated by ITS2 + *matK* + *rbcL* showing classification results of the query sequences

by ITS2 + *matK* + *rbcL* is provided as Fig. 2. Both "best match" and "best close match" functions of TaxonDNA provided the identification accuracy ranging from 15.1 to 97.4%, which is slightly higher than NJ tree (11.1–87.5%). However, TaxonDNA (Supplementary Fig. S4) and NJ tree (Supplementary Fig. S5) methods failed to discriminate the six *Pterocarpus* species.

Comparison of different analytical methods on cost-effectiveness, BLOG requires four-locus combination to discriminate the six *Pterocarpus* species successfully. The approaches of BP-neural network and J48 need at least three-locus combination to arrive at the species discrimination. Instead, the SMO classifier could separate the six *Pterocarpus* species with only two-locus combination, ITS2 + *matK*. Our previous study (Jiao et al. 2018) demonstrated the three-locus combination (*matK* + *ndhF-rpl32* + ITS2) was necessary to separate these six *Pterocarpus* species when using TaxonDNA and NJ tree methods. The results provided here showed that the MLAs type of SMO could reduce the cost of candidate barcodes for these species.

Species discrimination between *P. santalinus* and *P. tinctorius* using MLAs

The results of discrimination between *P. santalinus* and *P. tinctorius* using MLAs are shown in Fig. 3. BLOG showed that the two species could be separated based on the diagnostic locus of position 51, namely, "if position 51 is G, then the species is *P. santalinus*; if position 51 is T, then the species is *P. tinctorius*". The results of the BP-neural network indicated that 4 query sequences of *P. santalinus* and 2 query sequences of *P. tinctorius* were successfully identified with high probability (0.977 and 0.952, respectively). SMO provided variable hyperplanes for separating the two species with normalized attribute weights, of which position 51 presented the optimal hyperplane for separating 18 sequences of *P. santalinus* from 15 sequences of *P. tinctorius*. J48 also presented the logic decisions to recognize the two species based on position 51 of the sequences, which was consistent with the result of BLOG and SMO. In this regard, any of the MLAs but the BP-neural network correctly identified a simple, single-position, species-specific character for species identification.

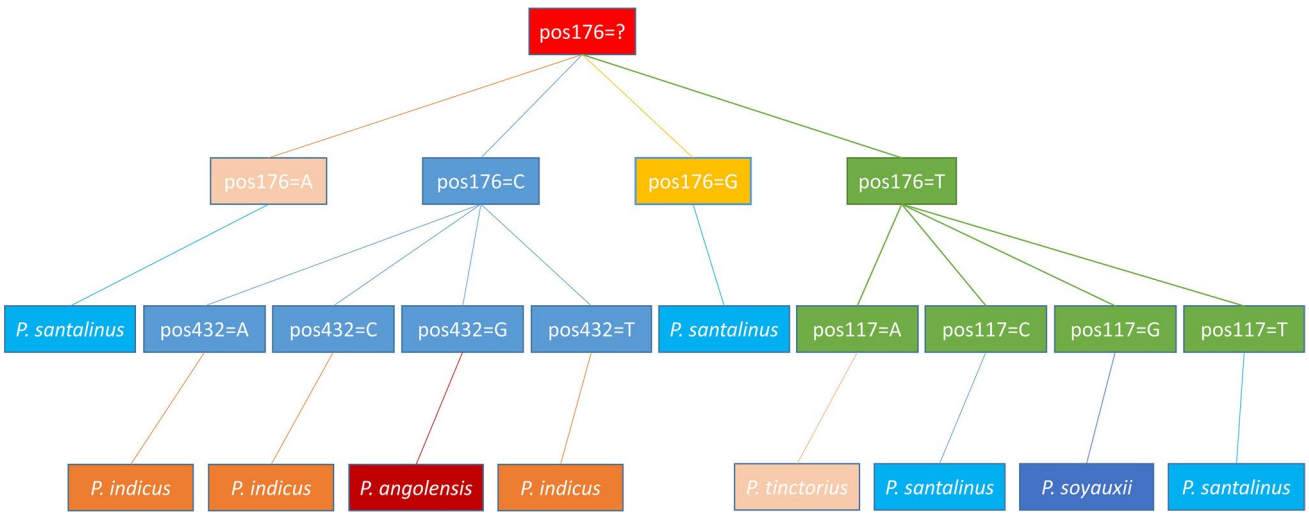

Fig. 2 Decision tree for discrimination of six *Pterocarpus* species produced by ITS2 + *matK* + *rbcL*

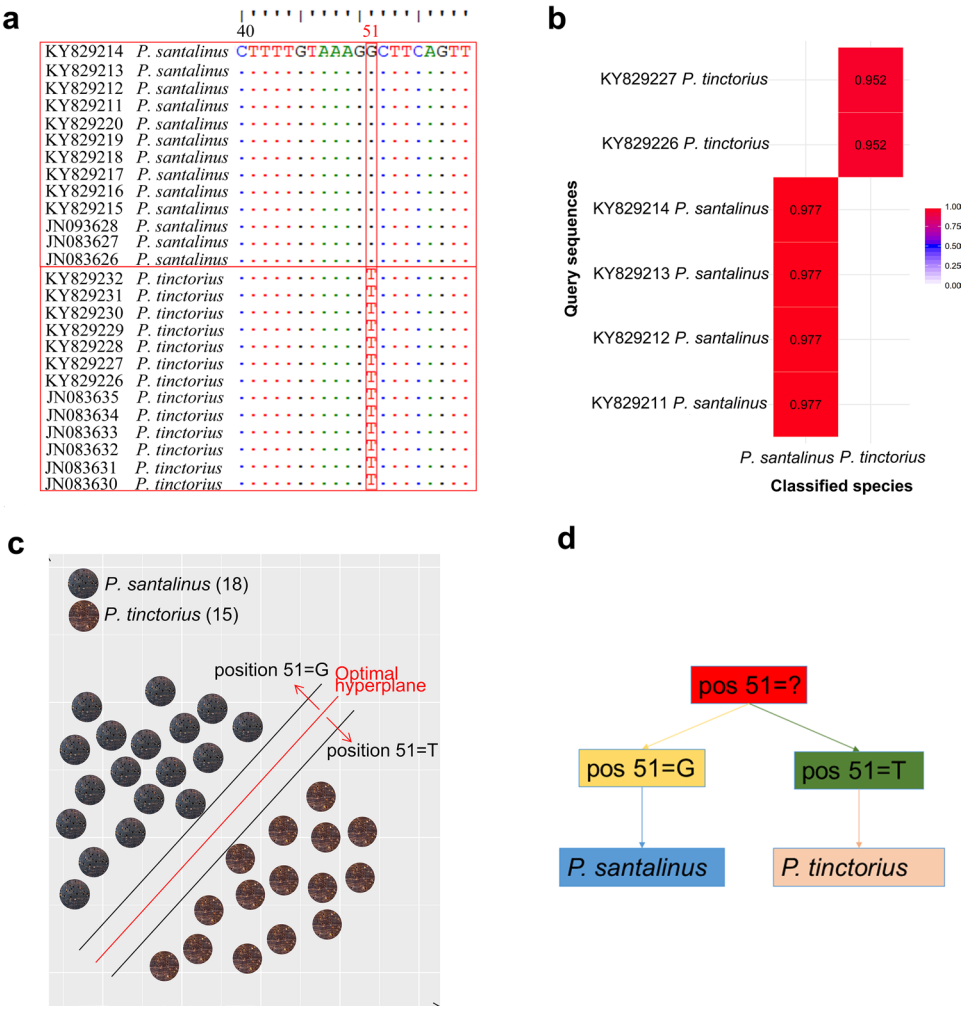

Fig. 3 Results of discrimination between *P. santalinus* and *P. tinctorius* using MLAs. **a** BLOG. **b** BP-neural network. **c** SMO. **d** J48

Discussion

Comparison of different analytical methods based on identification accuracy and cost-effectiveness

The results showed that MLAs exhibit higher identification accuracy than distance- and tree-based methods. The MLAs achieved the highest resolution with various multiple-locus combinations, while both the distance- and tree-based methods failed to provide a 100% accuracy, even using any combination of the four barcodes. Specifically, the six *Pterocarpus* species could be separated with 100% accuracy by the SMO classifier using the two-locus combination ITS2 + *matK*. The main objective of the SMO is to maximize the margin, which is the distance between the hyperplane and the closest vectors to it from both classes (Nalepa and Kawulok 2018). For cases where no linear separation is possible, the SMO can work in combination with the "kernel function", which can automatically realize a non-linear mapping to a feature space. Hartvig et al. (2015) tested three markers (ITS, *matK* and *rbcL*) on thirty-one *Dalbergia* species and compared their discrimination ability with distance-based, tree-based and machine learning methods, showing that TaxonDNA and SMO classifier gave the highest correct identification rates. In addition, our previous study (He et al. 2018) established a reference data set with DNA sequences of selected barcodes (ITS2, *matK, trnH-psbA* and *trnL*) and their combinations for DNA barcoding analysis of eight *Dalbergia* species and investigated the efficiency of machine learning approaches for wood species identification, and the results indicated that the SMO classifier performed the best on discrimination of eight endangered *Dalbergia* timber species and identification of non-vouchered wood specimens.

In our prior work on the DNA barcoding of *Pterocarpus* wood, we showed that the three-locus combination *matK* + *ndhF-rpl32* + ITS2 was the best barcode combination when species discrimination was done using distance-based or tree-based methods (Jiao et al. 2018). In this expanded work here, we found that 100% accuracy could be attained with only two barcodes in combination, ITS2 + *matK*, when species discrimination was done using the SMO classifier, which reduces the downstream cost of

barcoding these woods, as only two loci must be amplified, sequenced, and analyzed. This result is significantly meaningful to the DNA barcoding of wood, because it is still challenging to extract high-quality and quantity DNA from wood tissues, especially for the dry and aged wood in the trade (Jiao et al. 2014, 2018). Overall, barcode combinations involving ITS2 achieved better performance for species discrimination in this study, as well as in previous studies, due to the relatively strong discrimination power of ITS2 as a single barcode (Yao et al. 2010; Han et al. 2013, 2016; Xu et al. 2015b; Parveen et al. 2017; Yu et al. 2018). The two-locus combination ITS2 + *matK* is sufficient for discriminating the six *Pterocarpus* species based on the MLAs, and outperforms the three-locus combination of *matK* + *ndhF-rpl32* + ITS2 proposed by our previous study. The use of SMO classifier in this study reduced the required barcodes number from three to two when discriminating the six *Pterocarpus* species based on the current database. This is a particularly important result considering the broad use of these markers in phylogenetic research and thus their availability in sequence databases for a wide range of taxa. Once a comprehensive and reliable reference library of DNA barcode sequences for endangered species was established with the boom of molecular data, MLAs are promising tools to deal with large scale of data and significantly reduce the human labor and cost for DNA barcoding of wood.

Species discrimination between *P. santalinus* and *P. tinctorius* using MLAs

P. santalinus, a CITES-listed species, exhibits extremely similar wood anatomical features with *P. tinctorius*, which is not a CITES-listed species and has much lower economic value. Consequently, showing that the mini DNA barcode, *ndhF-rpl32* can discriminate *P. santalinus* from *P. tinctorius* is critical for the implementation of CITES enforcement and the monitoring of timber trade. Based on our sequences, separation could be achieved without any special computation and a human evaluating the sequence at position 51 (Fig. 3). MLAs output reliable and readable identification results (Fig. 3) with most cost-effectiveness mini barcode, *ndhF-rpl32,* for users to separate *P. santalinus* from *P. tinctorius* in the forensic wood identification. Another possible method, which would preclude the need for sequencing the barcode after PCR, would be to determine if a restriction enzyme would recognize the sequence including position 51 in one species and not in the other. In such a case, amplifying *ndhF-rpl32*, digesting with the targeted endonuclease, and then running a gel would provide species-level identification (Little 2014).

Application of MLAs for DNA barcoding in the conservation of *Pterocarpus* wood

Biodiversity conservation has rapidly become an international concern in part due to the sharp increase of illegal logging and over-exploitation of forest resources around the world (Lowe et al. 2016), and illegal logging imposes huge threats to the conversation of *Pterocarpus* wood. The application of DNA barcoding to identify the species of internationally traded timber has attracted increasing interest as a potential part of global systems to support sustainable forestry and especially to reduce illegal logging (Hartvig et al. 2015; Hassold et al. 2016; Yu et al. 2017; He et al. 2018; Jiao et al. 2018). In this study, MLAs provided higher identification accuracy than distance- and tree-based methods. In addition, MLAs could discriminate the six *Pterocarpus* wood with only two-locus combination ITS2 + *matK*, and separate CITES-listed species *P. santalinus* from its anatomically similar species *P. tinctorius* with mini barcode *ndhF-rpl32*, based on the reference data set we established, which verified the feasibility of MLAs in the application of conservation of *Pterocarpus* wood to combat illegal logging. For the broad forensic application of DNA barcoding, an international scientific DNA barcode reference library for endangered timber species based on the world's xylaria should be established (Robinson and Sinovas 2018). With such a reference library, MLAs could provide strong discrimination power and enhance the application of DNA barcoding as a tool for conservation of endangered timber species, and thus help to combat illegal logging and protect forests for future generations.

Conclusion

Two considerations for the application of DNA barcoding analytical methods in a forensic or conservation context are identification accuracy for testing unknown specimens and cost-effectiveness of candidate barcodes. The results provided in this study demonstrated that MLAs showed higher identification accuracy on species discrimination of six *Pterocarpus* species over TaxonDNA and NJ tree methods; specifically, SMO classifier exhibited the best performance with variable barcode combinations. In addition, only two-locus is necessary for combination (ITS2 + *matK*) to separate these six *Pterocarpus* species successfully, which is more efficient than the three-locus combination provided in our previous study and reduces the downstream cost of the DNA barcodes of these wood in this study. This study shows that MLAs provide higher identification accuracy and cost-effectiveness over other analytical methods to arrive at a species-level identification of wood. The results demonstrated in this study verified MLAs as a reliable and efficient

tool for the DNA barcoding of wood, which can be applied in combatting illegal logging and conservation of endangered species broadly.

Author contribution statement TH, LJ, and YY conceived and designed the research. TH, LJ, AW, and YY analyzed the data. TH, AW, LJ, and YY wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements This work was financially supported by National Natural Science Foundation of China (Grant No. 31600451), the National High-level Talent for Special Support Program of China (Grant No. W02020331), and the China Scholarship Council (Grant No. 2017-3109). We express our gratitude to Professor Xiaomei Jiang, Dr. Min Yu, Dr. Bo Liu and Dr. Prabu Ravindran for their assistance and suggestions on this study. We thank Sarah Friedrich for her help with the figure works.

References

Bertolazzi P, Felici G, Weitschek E (2009) Learning to classify species with barcodes. BMC Bioinform 10(14):S7

Brancalion PHS, Almeida DRA, Vidal E, Molin PG, Sontag VE, Souza SEXF, Schulze M (2018) Fake legal logging in the Brazilian Amazon. Sci Adv 4(8):aat1192

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proc Natl Acad Sci USA 106(31):12794–12797

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS One 5:e8613

Collins RA, Cruickshank RH (2012) The seven deadly sins of DNA barcoding. Mol Ecol Resour 13(6):969–975

Collins RA, Boykin LM, Cruickshank RH, Armstrong KF (2012) Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. Methods Ecol Evol 3(3):457–465

Damm S, Schierwater B, Hadrys H (2010) An integrative approach to species discovery in odonates: from character-based DNA barcoding to ecology. Mol Ecol 19(18):3881–3893

Delgado-Serrano L, Restrepo S, Bustos JR, Zambrano MM, Anzola JM (2016) Mycofier: a new machine learning-based classifier for fungal ITS sequences. BMC Res Notes 9(1):402

Dormontt EE, Boner M, Braun B, Breulmann G, Degen B, Espinoza E, Gardner S, Guillery P, Hermanson JC, Koch G, Lee SL, Kanashiro M, Rimbawanto A, Thomas D, Wiedenhoelft AC, Yin Y, Zahnen J, Lowe AJ (2015) Forensic timber identification: it's time to integrate disciplines to combat illegal logging. Biol Conserv 191:790–798

Ekrema T, Willassen E, Stura E (2007) A comprehensive DNA sequence library is essential for identification with DNA barcodes. Mol Phylogenet Evol 43(2):530–542

Gasson P (2011) How precise can wood identification be? Wood anatomy's role in support of the legal timber trade, especially CITES. IAWA J 32(2):137–154

Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning. Mach Learn 3(2):95–99

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitefield JB, Hebert PDN (2006) A minimalist barcode can identify a specimen whose DNA is degraded. Mol Ecol Resour 6(4):959–964

Han J, Zhu Y, Chen X, Liao B, Yao H, Song J, Chen S, Meng F (2013) The short ITS2 sequence serves as an efficient taxonomic sequence tag in comparison with the full-length ITS. BioMed Res Intl 2013:741476

Han Y, Duan D, Ma X, Jia Y, Liu Z, Zhao G, Li Z (2016) Efficient identification of the forest tree species in Aceraceae using DNA barcodes. Front Plant Sci 7:1707

Hartvig I, Czako M, Kjaer ED, Nielsen LR, Theilade I (2015) The use of DNA barcoding in identification and conservation of rosewood (*Dalbergia* spp.). PLoS One 10:e0138231

Hassold S, Lowry PP II, Bauert MR, Razafintsalama A, Ramamonjisoa L, Widmer A (2016) DNA barcoding of Malagasy rosewoods: towards a molecular identification of CITES-listed *Dalbergia* species. PLoS One 11:e0157881

He T, Jiao L, Yu M, Guo J, Jiang X, Yin Y (2018) DNA barcoding authentication for the wood of eight endangered *Dalbergia* timber species using machine learning approaches. Holzforschung. https://doi.org/10.1515/hf-2018-0076

Hebert PDN, Cywinska A, Ball SL, Dewaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc B Biol Sci 270(1512):313–321

Hendrich L, Morinière J, Haszprunar G, Hebert PDN, Hausman A, Köhler F, Balke M (2015) A comprehensive DNA barcode database for Central European beetles with a focus on Germany: adding more than 3500 identified species to BOLD. Mol Ecol Resour 15(4):795–818

IUCN Red List of Threatened Species (2017) http://www.iucnredlist.org/. Accessed 5 Feb 2018

Jiao L, Yin Y, Cheng Y, Jiang X (2014) DNA barcoding for identification of the endangered species *Aquilaria sinensis*: comparison of data from heated or aged samples. Holzforschung 68(4):487–494

Jiao L, Liu X, Jiang X, Yin Y (2015) Extraction and amplification of DNA from aged and archaeological *Populus euphratica* wood for species identification. Holzforschung 69(8):925–931

Jiao L, Yu M, Wiedenhoeft AC, He T, Li J, Liu B, Jiang X, Yin Y (2018) DNA barcode authentication and library development for the wood of six commercial *Pterocarpus* species: the critical role of xylarium specimens. Sci Rep 8(1):1945

Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 102(23):8369–8374

Lewis SL, Edwards DP, Galbraith D (2015) Increasing human dominance of tropical forests. Science 349(6250):827–832

Li J, Cui Y, Jiang J, Yu J, Niu L, Deng J, Shen F, Zhang L, Yue B, Li J (2017) Applying DNA barcoding to conservation practice: a case study of endangered birds and large mammals in China. BioL Conserv 26(3):653–668

Libbrecht MW, Nobble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321–332

Little DP (2014) A DNA mini-barcode for land plants. Mol Ecol Resour 14(3):437–446

Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. Cladistics 3(1):1–21

Lowe AJ, Dormontt EE, Bowie MJ, Degen B, Gardner S, Thomas D, Clarke C, Rimbawanto A, Wiedenhoeft AC, Yin Y, Sasaki N (2016) Opportunities for improved transparency in the timber trade through scientific verification. Bioscience 66(11):990–998

Lowenstein JH, Amato G, Kolokotronis SO (2009) The real *maccoyii*: identification tuna sushi with DNA barcodes-contrasting characteristic attributes and genetic distances. PLoS One 4:e7866

MacLeod N, Benfield M, Culverhouse P (2010) Time to automate identification. Nature 467(7312):154–155

McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82(1):290–297

Meier R, Shiyang K, Vaidya G, Peter KLN (2006) DNA Barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst Biol 55(5):715–728

More RP, Mane RC, Purohit HJ (2016) *MatK*-QR classifier: a patterns based approach for plant species identification. BioData Min 9(1):39

Nalepa J, Kawulok M (2018) Selecting training sets for support vector machine: a review. Artif Intell Rev 6:1–44

NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:7–19

Ng KKS, Lee SL, Tnah LH, Nurul-Farhanah Z, Ng CH, Lee CT, Tani N, Diway B, Lai PS, Khoo E (2016) Forensic timber identification: a case study of a CITES listed species, *Gonystylus bancanus* (Thymelaeaceae). Forensic Sci Int Genet 23:197–209

Pang X, Song J, Zhu Y, Xu H, Huang L, Chen S (2010) Appling plant DNA barcodes for Rosaceae species identification. Cladistics 27(2):165–170

Parveen I, Singh HK, Malik S, Raghuvanshi S, Babbar SB (2017) Evaluating five different loci (*rbcL*, *rpoB*, *rpoC1*, *matK*, and ITS) for DNA barcoding of Indian orchids. Genome 60(8):665–671

Patel N, Upadhyay S (2012) Study of various decision tree pruning methods with their empirical comparison in WEKA. Intl J Comput Appl 60(12):20–25

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proc R Soc B 275(1632):237–247

Robinson JE, Sinovas P (2018) Challenges of analyzing the global trade in CITES-listed wildlife. Conserv Biol 32(5):1203–1206

Ross HA, Murugan S, Li WL (2008) Testing the reliability of genetic methods of species identification via simulation. Syst Biol 57(2):216–230

Saatchi SS, Harris NL, Brown S, Lefsky M, Mitchard ETA, Salas W, Zutta BR, Buermann W, Lewis SL, Hagen S, Petrova S, White L, Silman M, Morel A (2011) Benchmark map of forest carbon stocks in tropical regions across three continents. Proc Natl Acad Sci USA 108(24):9899–9904

Sarkar IN, Planet PL, Desalle R (2008) CAOS software for use in character-based DNA barcoding. Mol Ecol Resour 8(6):1256–1259

Saslis-Lagoudakis CH, Klitgaard BB, Forest F, Francis L, Savolainen V, Williamson EM, Hawkins JA (2011) The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: an example from *Pterocarpus* (Leguminosae). PLoS One 6:e22275

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

Tanabe AS, Toju H (2013) Two new computational methods for universal DNA barcoding: a benchmark using barcode sequences of bacteria, archaea, animals, fungi and land plants. PLoS One 8:e76910

Velzen RV, Weitschek E, Felici G, Bakker FT (2012) DNA barcoding of recently diverged species: relative performance of matching methods. PLoS One 7:e30490

Weitschek E, Velzen R, Felici G, Bertolazzi P (2013) BLOG 2.0: a software system for character-based species classification with DNA barcode sequences. What it does, how to use it? Mol Ecol Resour 13(6):1043–1046

Weitschek E, Fiscon G, Felici G (2014) Supervised DNA barcodes species classification: analysis, comparisons and results. BioData Min 7:4

Wiedenhoeft AC (2014) Curating xylaria. In: Salick J, Konchor K, Nesbitt M (eds) Curating biocultural collections. A handbook. Kew Publishing, London, pp 127–134

Xu C, Dong W, Shi S, Cheng T, Li C, Liu Y, Wu P, Wu H, Gao P, Zhou S (2015a) Accelerating plant DNA barcode reference library construction using herbarium specimens: improved experimental techniques. Mol Ecol Resour 15(6):1366–1374

Xu S, Li D, Li J, Xiang X, Jin W, Huang W, Jin X, Huang L (2015b) Evaluation of the DNA barcodes in *Dendrobium* (Orchidaceae) from mainland Asia. PLoS One 10:e0115168

Yan L, Liu J, Möller M, Zhang L, Zhang X, Li D, Gao L (2015) DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya-Hengduan Mountains. Mol Ecol Resour 15(4):932–944

Yao H, Song J, Chang L, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS One 5:e13102

Yassin A, Markow TA, Narechania A, O'Grady PM, DeSallea R (2010) The genus *Drosophila* as a model for testing tree- and character-based methods of species identification using DNA barcoding. Mol Phylogent Evol 57(2):509–517

Yu M, Jiao L, Guo J, Wiedenhoeft AC, He T, Jiang X, Yin Y (2017) DNA barcoding of vouchered xylarium wood specimens of nine endangered *Dalbergia* species. Planta 246(6):1165–1176

Yu N, Wei Y, Zhang X, Zhu N, Wang Y, Zhu Y, Zhang H, Li F, Yang L, Sun J, Sun A (2018) Barcode ITS2: a useful tool for identifying *Trachelospermum jasminoides* and a good monitor for medicine market. Sci Rep 7:5037

Zeng C, Hollingsworth PM, Yang J, He Z, Zhang Z, Li D, Yang J (2018) Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods 14:43

Zhang AB, Sikes DS, Muster C, Li SQ (2008) Inferring species membership using DNA sequences with back-propagation neural network. Syst Biol 57(2):202–215

Zhang A, Muster C, Liang H, Zhu C, Crozier R, Wan P, Feng J (2012) A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. Mol Ecol 21(8):1848–1863

Zhang AB, Hao MD, Yang CQ, Shi ZY (2017) BarcodingR: an integrated R package for species identification using DNA barcodes. Methods Ecol Evol 8(5):627–637

Zou S, Li Q, Kong L, Yu H, Zheng X (2011) Comparing the usefulness of distance, mornophyly and character-based DNA barcoding methods in species identification: a case study of *Neogastropoda*. PLoS One 6:e26619