Botanical Journal of the Linnean Society, 2022, **199**, 741–753. With 3 figures.

Do taxon-specific DNA barcodes improve species discrimination relative to universal barcodes in Lauraceae?

ZHI-FANG LIU[1,2,8], HUI MA[2], XIAO-YAN ZHANG[2,3], XIU-QIN CI[2,4], LANG LI[2,4], JIAN-LIN HU[2,3], CAN-YU ZHANG[2,3], JIAN-HUA XIAO[2,3], HIS-WEN LI[5], JOHN G. CONRAN[6], ALEX D. TWYFORD[7,8,*], PETER M. HOLLINGSWORTH[8,*] and JIE LI[2,4,*]

[1]*Key Laboratory of Chemical Biology (Ministry of Education), School of Pharmaceutical Sciences, Cheeloo College of Medicine, Shandong University, Jinan, 250012, China*
[2]*Plant Phylogenetics and Conservation Group, Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming, 650223, China*
[3]*University of Chinese Academy of Sciences, Beijing, 100049, China*
[4]*Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Mengla, 666303, China*
[5]*Herbarium (KUN), Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China*
[6]*Australian Centre for Evolutionary Biology and Biodiversity & Sprigg Geobiology Centre, School of Biological Sciences, University of Adelaide, Adelaide, 5005, Australia*
[7]*Institute of Evolutionary Biology, Ashworth Laboratories, The University of Edinburgh, Edinburgh, EH9 3FL, United Kingdom*
[8]*Genetics and Conservation Section, Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, United Kingdom*

Received 15 May 2021; revised 12 September 2021; accepted for publication 8 October 2021

The aim of DNA barcoding is to enable fast and accurate species identification. However, universal plant DNA barcodes often do not provide species-level discrimination, especially in taxonomically complex groups. Here we use Lauraceae for the design and evaluation of DNA barcoding strategies, considering: (1) the efficacy of taxon-specific DNA barcode regions compared with universal barcodes for species discrimination; and (2) how the extent of intra- and interspecific sampling affects species discrimination rates. To address these areas, we targeted the highly polymorphic, taxon-specific barcode regions *ycf1* + *ndhH–rps15* + *trnL–ycf2* for Lauraceae and compared them against the suite of standard plastid loci used for DNA barcoding (*rbcL* + *matK* + *trnH–psbA*) and the standard nuclear barcode ITS. The highest discrimination success came from nrDNA ITS, whereas the plastid regions (*rbcL* + *matK* + *trnH–psbA*) and the taxon-specific regions (*ycf1* + *ndhH–rps15* + *trnL–ycf2*) showed limited and inconsistent resolution. These results highlight that taxon-specific plastid barcodes may provide limited gains in discriminatory power in complex, closely related groups like Lauraceae. Moreover, our study showed that species discrimination greatly depends on the taxon sampling scheme, with relatively lower species discrimination observed where there is more comprehensive intra- and interspecific sampling. The outstanding challenge for plant DNA barcoding is the development of assays that allow routine low-cost access to large numbers of nuclear markers to facilitate the sequencing of large numbers of individuals.

ADDITIONAL KEYWORDS: DNA barcoding – ITS – phylogenetics – plastid DNA – specific barcode design.

INTRODUCTION

The goal of DNA barcoding is to tell species apart using DNA sequencing. Three plastid regions (*rbcL*, *matK* and *trnH–psbA*) plus the nuclear ribosomal DNA internal transcribed spacer region ITS, or their combination, are widely used as standard barcodes for plants (CBOL Plant Working Group, 2009; China Plant BOL Group, 2011). However, for complex and closely related taxa, these regions often lack adequate variation (Li *et al.*, 2015; Wang *et al.*, 2018), resulting

*Corresponding authors. E-mail: PHollingsworth@rbge.org.uk; jieli@xtbg.ac.cn; alex.twyford@ed.ac.uk

in low species discrimination rates. To overcome this problem, more variable sequence regions may be used to supplement the standard plant barcodes and provide more informative characters to help distinguish taxa. However, although this approach can be promising, these high-variability plastid regions can still fail to discriminate taxa if the plastid genome does not track species boundaries (Percy *et al.*, 2014), and careful testing is required prior to their wider deployment.

DNA barcoding studies vary dramatically in their taxonomic scope and number of individuals and species sampled, ranging from a few to thousands of individuals (Table 1). Broad-scale DNA barcoding studies with a wide phylogenetic scope and limited sampling of sister species, such as floristic surveys or studies comparing the relative performance of barcode markers, often find large genetic distances and reasonable taxon discrimination (CBOL Plant Working Group, 2009; China Plant BOL Group, 2011). In contrast, in monographic studies when the focus is on a specific family, genus or group of species, discrimination rates are often much lower, e.g. in Lauraceae (Liu *et al.*, 2017) and *Rhododendron* L. (Yan *et al.*, 2015) (Table 1). Numerous studies have found that the standard barcodes often do not provide species level resolution (e.g. Lahaye *et al.*, 2008; Little, 2014; Pei *et al.*, 2015; Yan *et al.*, 2015; Liu *et al.*, 2017), with these universal barcodes limited in their ability to correctly identify related species due to low levels of sequence variability (Pei *et al.*, 2015). The success of DNA barcodes in discriminating related species may also be contingent upon the evolutionary and/or biogeographic histories of the taxonomic group tested, with particular challenges in recent species radiations and groups characterized by hybridization.

To overcome these resolution issues, researchers have suggested designing taxon-specific barcodes to improve discrimination in a given group (Takahiro *et al.*, 1999; Hadziavdic *et al.*, 2014; Shen *et al.*, 2019; Chen *et al.*, 2020). For example, Chen *et al.* (2020) used *ycf1* and *psbM–psbD* to identify species of *Fritillaria* L., with a success rate of 87.5%, equivalent to the universal barcode *matK* (87.5%) but preferable to ITS (62.5%), *rbcL* (62.5%) and *trnH–psbA* (25%).

Four of the dominant elements of the subtropical evergreen broadleaved forests in East Asia are Theaceae, Fagaceae, Magnoliaceae and Lauraceae (Fang & Yod, 1989), all of which are species-rich, morphologically complex groups where species are hard to identify, especially when sterile (Nie *et al.*, 2008; Oh & Manos, 2008; Zhu *et al.*, 2016; Yu *et al.*, 2017; Zheng *et al.*, 2018). Here we use Lauraceae as a representative group to test the efficacy of taxon-specific DNA barcodes for increasing species discrimination.

Lauraceae comprise *c.* 50 genera and *c.* 2500–3000 species (Rohwer 1993; van der Werff & Richter, 1996). In China, Lauraceae are well represented, with 25 genera and 445 species (Li *et al.*, 2008; Yang & Liu, 2015). Many species of Lauraceae are important economically as sources of medicine, timber, nutritious fruits, spices and perfumes, and ecologically as dominant canopy species in tropical and subtropical forests (Kostermans, 1957; van der Werff & Richter, 1996; Li *et al.*, 2008). Important spice and fruit species include avocado (*Persea americana* Mill.), bay laurel (*Laurus nobilis* L.), camphor tree or camphor laurel [*Cinnamomum camphora* (L.) J.Presl], cassia [*Cinnamomum cassia* (L.) D.Don] and cinnamon (several species of *Cinnamomum* Spreng.) (Chaw *et al.*, 2019).

However, despite their importance, taxonomic identification in Lauraceae remains challenging, and species richness and local estimates of species diversity remain poorly known (van der Werff & Richter, 1996). Moreover, morphological similarity of vegetative specimens and intra-taxon variability are major causes of taxonomic confusion. DNA barcoding has substantial promise for resolving issues with identifying vegetative specimens, although previous research has shown low single locus discrimination rates (*rbcL* 8.2%, *matK* 9.3%, *trnH–psbA* 9.5%, ITS 57.5%) (Liu *et al.*, 2017). This study investigates whether the application of variable taxon-specific DNA barcodes can improve taxon discrimination in Lauraceae. We also explore how sample density impacts on levels of inferred species discrimination.

MATERIAL AND METHODS

Sampling

Our sampling strategy was designed to have dense taxon sampling, such that pairs of closely related species are included, and broad geographical sampling of the family. Samples were selected based on the *Flora of China* (Li *et al.*, 2008), previous Chinese Lauraceae research (Liu *et al.*, 2017) and Lauraceae specimens collected by the Plant Phylogenetics & Conservation Group (details in Supporting Information, Table S1). This resulted in a total of 257 samples from 206 species (one to five individuals per species), representing 24 out of the 25 currently recognized genera of Lauraceae in China and 55 of the global total. These samples were distributed across nine provinces in China (Guangdong, Guangxi, Hainan, Hubei, Hunan, Sichuan, Xizang, Yunnan and Zhejiang) and sites in Indonesia, Japan, Laos and Myanmar, representing much of the East Asian diversity for the family (Supporting Information, Table S1; Fig. S1). *Calycanthus chinensis* Cheng & S.Y.Chang and *Calycanthus fertilis* Walter (Calycanthaceae) were selected as outgroups, based on evolutionary relationships in APG IV (APG IV, 2016),

Table 1. Summary of representative studies comparing candidate DNA barcodes in plants. The column 'Discrimination success' reports the percentage of species discriminated using the barcode combination considered optimal from the original study

Taxon	Number of samples/species	Candidate barcodes	Discrimination success (%)	Reference(s)
Land plants	96/96	*matK*, *rbcL*–a, *rpoB2*, *rpoC1*, *trnH–psbA*, *accD*, *ycf5*, *ndhJ*, ITS1	*trnH–psbA* (79.1)	Kress & Erickson, 2007
Land plants	907/550	*atpF–atpH*, *matK*, *rbcL*, *rpoB*, *rpoC1*, *psbK–psbI*, *trnH–psbA*	*rbcL* + *matK* (72)	CBOL Plant Working Group, 2009
Land plants	98/39	*rpoC1*, *rpoB*, *rbcL*, *matK*, *trnH–psbA*, *atpF–atpH*, *psbK–psbI*	*rbcL* + *trnH–psbA* + *matK* or *rpoC1* + *rbcL* + *matK* (*c.* 60)	Hollingsworth *et al.*, 2009
Land plants	490/420	*matK*, *rbcLb*, *trnH–psbA*, *ycf1*a, *ycf1*b	*ycf1*b (71.87)	Dong *et al.*, 2015
Seed plants	6286/1757	*rbcL*, *matK*, *trnH–psbA*, ITS	*rbcL* + *matK* + *trnH–psbA* + ITS (82.8)	China Plant BOL Group, 2011
Lauraceae	409/133	*matK*, *rbcL*, *trnH–psbA*, ITS2, ITS	ITS (57.5)	Liu *et al.*, 2017
Meliaceae	33/22	*rpoC1*, *rpoB*, *accD*, *psbB*, *psbN*, *psbT*, *trnS–trnG*, ITS	ITS (66.67)	Muellner *et al.*, 2011
Myristicaceae	40/8	*accD*, *matK*, *trnH–psbA*, *rbcL*, *rpoB*, *rpoC1*, UPA	*matK* + *trnH–psbA* (94.7)	Newmaster *et al.*, 2008
Rhododendron	531/173	ITS, *rbcL*, *matK*, *psbA–trnH*	ITS + *psbA–trnH* + *matK* or ITS + *psbA–trnH* + *matK*+*rbcL* (41.98)	Yan *et al.*, 2015
Primula	227/66	*rbcL*, *matK*, *trnH–psbA*, ITS, ITS2	*rbcL* + *matK* + *trnH–psbA* + ITS (68.75)	Yan *et al.*, 2015
Adiantum	154/33	*rbcL*, *matK*, *psbA–trnH*, *trnL–F*, *rps4–trnS*, ITS, *pgiC*, *gapC*, LEAFY, ITS2, IBR3_2, DET1, SQD1_1	*trnH–psbA* (75)	Wang *et al.*, 2016
Aspalathus	133/51	ITS, *psbA–trnH*	ITS (84)	Edward *et al.*, 2008
Protea	88/85	*rps16*, *ncp*GS	*rps16* & *ncp*GS (> 95)	Chase *et al.*, 2005
Curcuma	96/44	*rbcL*, *matK*, *trnH-psbA*, *trnL*–F, ITS2	ITS2 (46.7)	Chen *et al.*, 2015
Dalbergia	50/9	ITS2, *matK*, *trnL, trnH–psbA*, *trnV–trnM1*, *trnV–trnM2*, *trnC–petN*, *trnS–trnG*	ITS2 + *trnH–psbA* (100)	Yu *et al.*, 2017
Santalum	49/5	*matK*, *psbA–trnH*, *trnK*, *trnL*	Combinations including *psbA–trnH* (100)	Jiao *et al.*, 2019
Pterocarpus	39/6	*matK*, *ndhF–rpl32*, ITS2	*matK* + *ndhF–rpl32* + ITS2 (100)	Jiao *et al.*, 2018

with outgroup data downloaded from GenBank. The samples and vouchers of Lauraceae were identified based on a combination of morphological and molecular evidence, as described previously (Liu *et al.*, 2017), and all vouchers are stored at the Herbarium of Xishuangbanna Tropical Botanical Garden (HITBC) and Kunming Institute of Botany (KUN), Chinese Academy of Sciences, Yunnan, China. Total genomic DNA was extracted from the herbarium specimens by a modified CTAB method (Doyle & Doyle, 1987) or by using a Tiangen DNAsecure Plant Kit (DP320).

Specific DNA barcode design and verification

We designed new primer pairs to target non-standard DNA barcoding regions. It is known that many highly conserved regions exist in plastid genomes of Lauraceae (Song *et al.*, 2016, 2017; Liu *et al.*, 2021),

providing an opportunity to design primer pairs that are anchored in conserved regions, but which span diverse regions, thus amplifying variable sequences across a broad phylogenetic scope.

Primer pairs for the amplification of regions with high divergence were designed using Primer Premier (Singh *et al.*, 1998), which were subsequently checked with Oligo 7 (Rychlik, 2007) following the methodology of Liu *et al.* (2021). In brief, 80 *de novo* plastid genomes were searched for regions that may be suitable as Lauraceae-specific barcodes. To verify the divergence of filtered regions in the plastid genome, we performed multiple sequence alignments of 11 plastid genomes (choosing one species as the representative from genera with more than one species) using mVISTA (http://genome.lbl.gov/vista/index.shtml) (Frazer *et al.*, 2004) in LAGAN mode, the early-diverging species *Cryptocarya hainanensis* Merr. as the reference, then checked the divergence of the candidate output regions in the position of mVISTA (Supporting Information, Fig. S2). The three most consistently variable candidate regions (*ycf1*, *ndhH–rps15* and *trnL–ycf2*) were screened for suitability in this study (Table 2).

To compare our results with the Lauraceae study by Liu *et al.* (2017), we used most of the same species to amplify and verify our three plastid specific markers (Supporting Information, Table S1). PCRs were performed in 25 μL reaction mixtures containing 0.3 μL TaKaRa Taq polymerase (5 U), 2.5 μL 10× PCR buffer, 2.5 μL 25 mM $MgCl_2$, 2.0 μL 2.5 mM dNTPs, 1.0 μL (10 μM) of each primer and 2.0 μL template DNA. For PCR, cycling conditions were as follows: 94 °C, 2 min; 35 cycles of 30 s melting at 94 °C and 45 s annealing at 47 °C (*ycf1*), 54 °C (*ndhH–rps15*) and 60 °C (*trnL–ycf2*), increasing the extension time by 60 s at each cycle; at the end of 35 cycles, 10 min at 72 °C to complete extension and subsequent storage at 4 °C. PCR products were analysed using 1% agarose gel electrophoresis. These PCR results were compared with the results for the standard plastid markers *rbcL*, *matK*, *trnH–psbA* and the nuclear marker ITS (ITS1 + 5.8S + ITS2), most of which were generated in a previous study (Liu *et al.*, 2017) and downloaded from GenBank. However, a few standard barcode sequences were also newly generated (Supporting Information, Table S2). All PCR products were sequenced in both directions at the Beijing Genomics Institute (BGI) with an ABI 3730XL sequencer.

Table 2. The primers used for amplification and sequencing of three Lauraceae-specific barcodes

Lauraceae primers	Sequence 5′–3′
ndhH–rps15F	GAATATTTCCTAATTGTTCTGGT
ndhH–rps15R	AAAGG(G/A)TCTGTTGAATTTCAA
trnL–ycf2F	TGCATCCAGCAGGAATTGAACC
trnL–ycf2R	CTTGGCGGAATTGCCAC(G/A)TATGA
ycf1F	CCACTCCAAA(T/A)ATTTTCTAT
ycf1R	GAAAGAATATACAT(G/A)(G/C)ATA

Sequence editing and alignment

Raw sequences were assembled and edited using Sequencher 4.14 (GeneCodes Corp., Ann Arbor, Michigan, USA) and deposited in GenBank (Supporting Information, Tables S1, S2). Edited sequences were then aligned and adjusted manually using Geneious Prime. All variable sites were confirmed in the original trace files. For the *ycf1* marker, due to the presence of degenerate bases in the reverse primer (Table 2), the success rate of reverse sequences was low and most of the *ycf1* sequence results are therefore from forward (single-read) products. A supermatrix was created by concatenating the aligned sequences of the remaining markers.

DNA barcode comparisons

We compared between the standard DNA barcodes (*rbcL*, *matK*, *trnH–psbA*), our taxon-specific barcodes chosen for variability in Lauraceae (*ycf1*, *ndhH–rps15*, *trnL–ycf2*) and the stand-alone nuclear marker ITS. Three datasets differing by number of samples and occurrence of missing data were generated (details in Supporting Information, Note S1). The first dataset, 'Dataset A', includes all the newly generated sequences and the accessions downloaded from GenBank and merges data from different individual specimens to generate species-level consensuses. Species with missing data are included. For example, *Cinnamomum burmannii* (Nees & T.Nees) Blume in Supporting Information (Table S2) is represented by plastid-specific barcodes generated from the specimen YB02, by standard barcodes generated from the individual CXQ0020 and the ITS sequence is missing. The second dataset, 'Dataset B', includes samples for which all barcodes have been obtained (without missing data) and merges different individuals to generate species-level consensuses. For example, *Actinodaphne cupularis* (Hemsl.) Gamble in Supporting Information (Table S2) is represented by plastid-specific barcodes generated from the individual CXQ0454, plastid standard barcodes generated from the same individual CXQ045 and ITS generated from the individual CXQ0467. The third dataset, 'Dataset C', includes samples for which all barcode sequences have been obtained from the same individual specimen. For example, the sequences of *Actinodaphne forrestii* (Allen) Kosterm. in Supporting Information (Table S2) were all generated from the individual GBOW0216 (Fig. 1; Table 3; Supporting

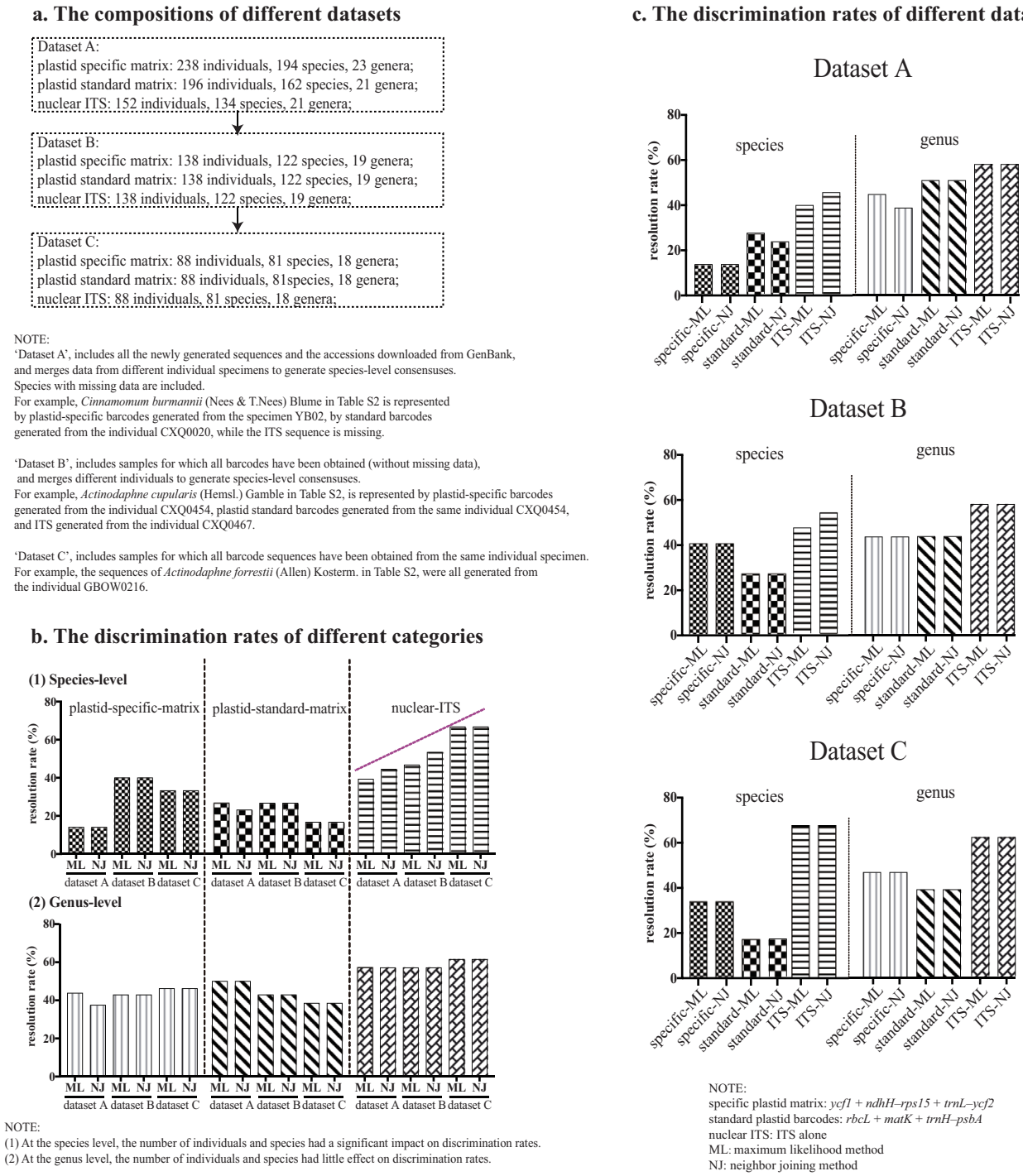

Figure 1. The compositions and performances of different matrices with different datasets. Different matrices are indicated with three different patterns at the species level and three different symbols at the genus level. a. The compositions of different datasets. b. The discrimination rates of different categories. c. The discrimination rates of different datasets.

Information, Table S2). Overall, Dataset A includes the most species, followed by B then C. However, Dataset C faithfully separates out individuals rather than merging them to form a species-level consensus. These three datasets were then compared both horizontally (between different barcodes) (Fig. 1B) and vertically (between different datasets) (Fig. 1C) to evaluate their discrimination abilities.

Data analysis

The utility of different datasets for species identification were investigated using two tree-based approaches to evaluate whether species were recovered as monophyletic with each DNA barcode matrix. These approaches were a maximum likelihood (ML) method using IQ-TREE (Minh *et al.*, 2020) and a neighbour joining (NJ) method using Geneious 11.1.4. The best-fit ML model for each dataset was then determined using ModelFinder (Kalyaanamoorthy *et al.*, 2017) using the option –TEST and a tree search with 1000 bootstrap replicates (Chernomor *et al.*, 2016; Kalyaanamoorthy *et al.*, 2017). The number of species or genera with multiple accessions resolving as monophyletic was recorded, as was the branch support for each node > 50%. One-way ANOVA and Tukey's multiple comparisons tests were used to test for differences in discrimination rates between different methods (ML/NJ) and taxon levels (species/genus).

RESULTS

Barcode universality and sequence characteristics

For the three specific barcodes, 741 sequences from 257 samples, representing 206 species and 24 genera, were obtained. These included 239 sequences for *ycf1*, 250 sequences for *trnL–ycf2* and 252 sequences for *ndhH–rps15* (Table 4; Supporting Information, Table S1). We recovered a sequence for at least one of the three markers, with *ndhH–rps15* showing the highest PCR amplification and sequencing success rates (100% and 98.05%) and shortest sequence length (584 bp). Moderate PCR amplification and sequencing success rates were observed for *trnL–ycf2* (98.05%, 97.28%, respectively) and *ycf1* (97.28%, 93.00%, respectively). In an ideal situation, designing PCR primers for specific taxa under investigation would simply involve identifying highly conserved regions that flank the variable regions, and then choosing non-degenerate. While in the taxa evolving rapidly or highly diverse is difficult to find the highly conserved regions, we need to apply degenerate PCR primers, even though the amplification and sequencing are much harder than non-degenerate, all barcodes still had a high PCR amplification rate (97%) and sequencing success rate (93%) (Table 4).

For the three standard barcodes, 585 sequences from 195 samples, representing 161 species and 21 genera, were obtained based on the recovery of the plastid specific barcodes (Supporting Information, Table S2). For the nuclear ITS, 153 sequences from 153 samples, representing 123 species and 21 genera, were obtained (Supporting Information, Table S2). The recovery of ITS was lower than that of the plastid regions (Supporting Information, Table S2).

Table 3. Comparison of characteristics of different DNA barcoding datasets in Lauraceae

Datasets		Number of taxa	Number of species	Number of genera	Number of sites	Best fit model of ML analysis
A*	Specific matrix	238	194	23	2455	TVM+F+G4
	Standard matrix	196	162	21	1928	K3Pu+F+I+G4
	nrDNA ITS	152	134	21	883	TIM2+F+I+G4
B§	Specific matrix	138	122	19	2440	TVM+F+G4
	Standard matrix	138	122	19	1922	K3Pu+F+I+G4
	nrDNA ITS	138	122	19	861	TIM2+F+I+G4
C*	Specific matrix	88	81	18	2433	TVM+F+G4
	Standard matrix	88	81	18	1900	K3Pu+F+I+G4
	nrDNA ITS	88	81	18	735	TIM2+F+I+G4

*"Dataset A' includes all the newly generated sequences and the accessions downloaded from GenBank, and merges data from different individual specimens to generate species-level consensuses. Species with missing data are included. §"Dataset B' includes samples for which all barcodes have been obtained (without missing data), and merges different individuals to generate species-level consensuses. "Dataset C' includes samples for which all barcode sequences have been obtained from the same individual specimen.

Table 4. PCR amplification and sequencing verification rates of the three Lauraceae-specific barcodes

Barcode regions	*ycf1*	*trnL–ycf2*	*ndhH–rps15*
Successful amplification/sampled individuals	250/257(97.28%)	252/257(98.05%)	257/257(100%)
Successful sequencing/sampled individuals	239/257(93.00%)	250/257(97.28%)	252/257(98.05%)

THE DISCRIMINATION EFFICIENCY OF DIFFERENT BARCODES

Although *trnL–ycf2* has the highest discrimination ability among the three taxon-specific individual barcodes, it could still only discriminate four out of 39 species (with $N \geq 2$ individual sampled per species) (Supporting Information, Fig. S3; Table S1); followed by *ycf1* (three species; Supporting Information, Fig. S4) and *ndhH–rps15* (two species; Supporting Information, Fig. S5). Similarly, the three-barcode combination of *ycf1* + *ndhH–rps15* + *trnL–ycf2*, could still only discriminate five out of 35 species (Fig. 2; Supporting Information, Tables S2, S3). These results suggest that the discrimination ability of the taxon-specific barcodes is limited, and even in combination, they give only marginally better discrimination than any single locus.

The three-barcode combination *rbcL* + *matK* + *trnH–psbA* could only discriminate seven out of 26 species (Supporting Information, Table S3). The universality of nrDNA ITS was lower than that of the plastid regions (*rbcL*, *matK* and *trnH–psbA*) (Supporting Information, Table S2), but it still showed reasonable discriminatory power, with eight out of 18 species in Dataset A with more than one sampled individual being discriminated (Fig. 1B, C; Supporting Information, Table S3). Figure 3 shows the ITS ML tree as the gene tree showing the highest species discrimination.

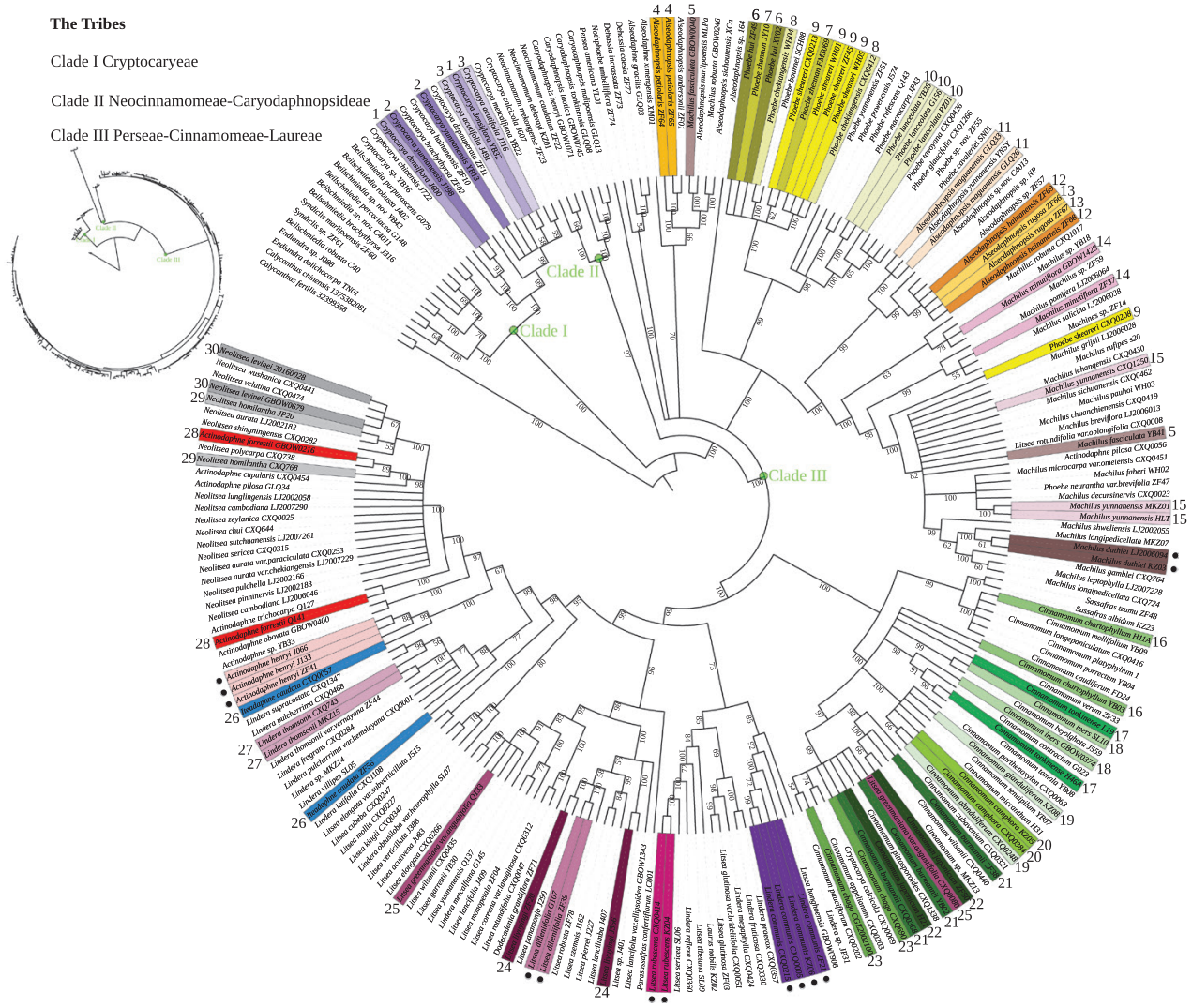

Figure 2. The ML tree of the specific DNA barcode dataset (*ycf1* + *ndhH–rps15* + *trnL–ycf2*) for Lauraceae. This analysis includes 238 individuals. The same colour and number represent the same species with more than one individual. Species successfully discriminated are indicated by a black dot.

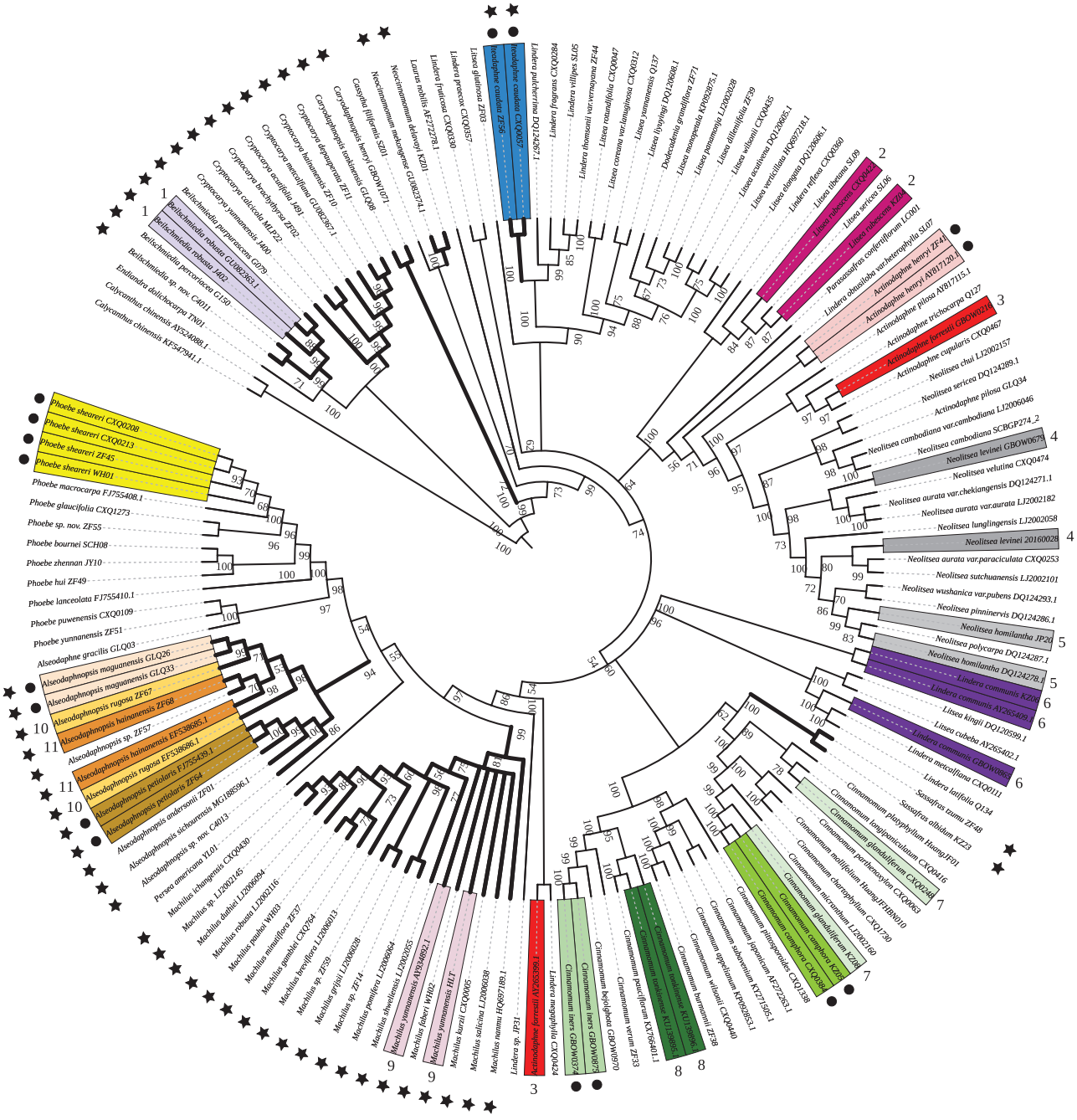

Figure 3. The ML tree of Lauraceae based on an analysis of 152 individuals sequenced with ITS. The same colour represents a given species with more than one individual. Species successfully discriminated are indicated by a black dot; successful genera by a black star. The thicker black lines indicate that the genus clustered is monophyletic.

Comparison of discrimination efficiency of different barcode datasets

The resolution rates of species (14.3–66.7%) and genera (37.5–61.5%) from different methods (NJ and ML) showed little difference between tree building approaches (Fig. 1; Supporting Information, Table S3; Fig. S6). At the species level, the number of individuals and species had a significant impact on species discrimination, with this being particularly clear in ITS where discrimination reduced as sample density increased (Dataset A: *c.* 40% discrimination; B: *c.* 50%; C: *c.* 60%) (Fig. 1), whereas the differences were more

idiosyncratic for the taxon-specific barcode matrices (A: *c.* 15%; B: *c.* 40%; C: *c.* 30%) and the standard plastid matrices (A: *c.* 25%; B: *c.* 30%; C: *c.* 15%). At the genus level, the number of individuals and species had little effect on identification rates, with the nuclear barcode ITS again showing the highest discrimination with *c.* 60% of the genera distinguishable (Fig. 1B, C), whereas both plastid barcodes (specific and standard) showed more limited resolution (35–50%).

Relationships in Lauraceae

Phylogenetic relationships among different barcode matrices were analysed. As the trees obtained from the ML and NJ analyses were almost identical in their topologies, and the *rbcL* + *matK* + *trnH–psbA* + ITS matrix was analysed in Liu *et al.* (2017), only the ML tree based on the *ycf1* + *ndhH–rps15* + *trnL–ycf2* matrix with 238 individuals with bootstrap support (BS) values is discussed here (Fig. 2). The tree contains three principal clades of Lauraceae, with clade 1 (BS = 100%, tribe Cryptocaryeae) strongly supported as the sister with the remaining groups. Clade 2 (BS = 97%, tribes Neocinnamomeae-Carydaphnopsideae) includes just two genera: *Neocinnamomum* H.Liu and *Caryodaphnopsis* Airy Shaw. The remainder, representing tribes Perseae, Cinnamomeae and Laureae, formed clade 3 (BS = 100%).

DISCUSSION

The failure of taxon-specific plastid DNA barcodes

Our study aimed to test whether the use of taxon-specific DNA barcode regions was a more effective approach for telling species of Lauraceae apart compared to universal DNA barcodes. However, although the evaluation of taxon-specific barcodes shows that the amplification and sequencing success rates were high, taxon resolution remains limited with no consistent improvement compared to the standard barcodes (Fig. 1B, C; Supporting Information, Table S3).

Ideally, DNA barcodes should at least satisfy the following criteria: they should (i) possess conserved flanking regions to enable routine amplification across highly divergent taxa; and (ii) have sufficient internal variability to enable species discrimination (Hollingsworth *et al.*, 2009). A prerequisite for the success of DNA barcodes is sufficient time since speciation for mutations and/or drift to lead to a set of genetic characters 'grouping' conspecific individuals together, separate from other species (Hollingsworth *et al.*, 2011). As the species histories are complex and/or where speciation is recent, taxa often show a lack of intraspecific coalescence/shared haplotypes (Hollingsworth *et al.*, 2009). Hence, recently radiating, species-rich taxa are challenging for barcoding identification (Hollingsworth *et al.*, 2009). Chanderbali *et al.* (2001) suggested that Lauraceae possess all these challenging characteristics. We speculate that the early-diverging lineages of Lauraceae have a long generation time and/or slow mutation rates, whereas the terminal Perseae-Cinnamomeae-Laureae clade is more recent in origin and/or a more rapidly evolving group, in which these plastid barcodes do not have sufficient internal variability to enable species discrimination. Species radiations of taxonomically complex groups in Lauraceae are thus a case where we may not expect a clear cut-off between intraspecific variation and interspecific divergence, and DNA barcoding may provide only limited discriminatory power.

The highest species resolution in this study came from using the biparentally inherited nuclear ITS locus, which, depending on the taxon sampling scheme, had a discrimination rate between 38.9–66.7%. In contrast, the predominantly maternally inherited plastid barcodes were less effective, due to a limited number of informative characters and plastid markers not tracking species boundaries. Data from our study thus support the notion that alternative nuclear barcoding solutions should be sought, rather than more intensive investigations of alternative plastid barcodes.

The effect of sampling scheme on species discrimination

The discrimination success rates varied among different datasets with different barcodes, but in general, sampling numbers influenced the percentage of species and genera that are distinguishable (Supporting Information, Fig. S7). For example, Lahaye *et al.* (2008) reported a species discrimination rate of *c.* 90% in orchids; however, this applied to datasets with limited sampling of species from the same genus. When they extended their sampling to a large group of orchids with extensive intrageneric sampling, species discrimination was much lower (Hollingsworth, 2008; Lahaye *et al.*, 2008), indicating the numbers of individuals and species affected discrimination scores. Similarly, Pei *et al.* (2015) reported that broad taxonomic sampling with relatively few co-occurring closely related taxa in a geographically restricted region can give relatively high rates of species discrimination. In the current study, there is marked variation in absolute discriminatory power in the different datasets (A, B, C) and reducing the sample size of ITS, from 152 to 138 individuals, increased species discrimination from *c.* 40% to *c.* 50%, with further reduction to 88

individuals leading to discrimination rates of up to *c.* 67% (Fig. 1; Supporting Information, Table S3). This observation is not unexpected, given that a decrease in sample density decreases the scale of the identification challenge, with fewer congeneric species present to disrupt successful species discrimination.

Relationships among major clades

The ML analysis provided moderate phylogenetic resolution for Lauraceae at both the generic and intrageneric levels (Fig. 2), especially in early-diverging lineages, with Cryptocaryeae, Neocinnamomeae-Carydaphnopsideae and Perseae + Laureae + Cinnamomeae corresponding to our clades 1, 2 and 3, respectively. Cryptocaryeae is the earliest-diverging group of Lauraceae (Chanderbali *et al.*, 2001; Song *et al.*, 2019; Liu *et al.*, 2021). Tribes Neocinnamomeae-Carydaphnopsideae are associated in the present study and have been previously found to have a relatively close relationship (Wang *et al.*, 2010; Li *et al.*, 2016; Liu *et al.*, 2021). The remaining clade (Perseae, Cinnamomeae and Laureae) received strong support, in agreement with previous studies (Chanderbali *et al.*, 2001; Rhower *et al.*, 2009; Liu *et al.*, 2021). However, as in these earlier studies, species relationships in presently accepted tribes are poorly resolved.

CONCLUSION

Although many studies have shown that specific and standard plastid barcodes are useful tools for species identification, some groups, including Lauraceae, pose a challenge for plastid barcoding, where incomplete lineage sorting (Fazekas *et al.*, 2009; Naciri & Linder, 2015; Rendon-Anaya *et al.*, 2019; Chen *et al.*, 2020), high frequencies of introgression (Rieseberg & Soltis, 1991) and the occurrence of selective sweeps (Twyford 2014) may be problematic for telling species apart. Despite the design of taxon-specific barcodes targeting variable regions of the plastid genome in Lauraceae, limited gains in discriminatory power were obtained. Future work should therefore focus on multiple unlinked nuclear DNA regions to improve the discriminatory power of barcoding for such problematic plant groups.

ACKNOWLEDGEMENTS

The authors would like to thank Bing Liu, Yu Song, Jian-Wu Li, Hong-Hu Meng, Jian-Feng Huang and Shu-Li Wang for assistance with sample collection and Jens G. Rohwer for the morphological identification of some species. We are grateful to Kunming Institute of Botany for experiment assistance and Royal Botanic Garden Edinburgh for data analysis assistance.

FUNDING

This study was supported by the National Natural Science Foundation of China (31370245, 31770569, 31970222), the Biodiversity Conservation Program of the Chinese Academy of Sciences (ZSSD-013), the Science and Technology Basic Resources Investigation Program of China: Survey and Germplasm Conservation of plant Species with Extremely small populations in south-west China (2017YF100100) and the 135 programs of the Chinese Academy of Sciences (2017XTBG-T03).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA ACCESSIBILITY

GenBank numbers ITS MZ230384–MZ23230436, *matK* MZ226454–MZ226530, *rbcL* e54MZ226531–MZ226607, *trnH–psbA* MZ226608–226684, *ndhH–rps15* MZ289151–MZ289402, *trnL–ycf2* MZ289403–MZ289652 and *ycf1* MZ289653–MZ289891 are noted in the Supporting Information, Tables S1 and S2.

REFERENCES

APG. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181:** 1–20.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA* **106:** 12794–12797.

Chanderbali AS, van der Werff H, Renner SS. 2001. Phylogeny and historical biogeography of Lauraceae: evidence from the chloroplast and nuclear genomes. *Annals of the Missouri Botanical Garden* **88:** 104–134.

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

Chaw SM, Liu YC, Wu YW, Wang HY, Lin CI, Wu CS, Ke HM, Chang LY, Hsu CY, Yang HT, Sudianto E,

Hsu MH, **Wu KP**, **Wang LN**, **Leebens-Mack JH**, **Tsai IJ. 2019.** Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nature Plants* **5:** 63–73.

Chen J, **Zhao J**, **Erickson DL**, **Xia N**, **Kress WJ. 2015.** Testing DNA barcodes in closely related species of *Curcuma* (Zingiberaceae) from Myanmar and China. *Molecular Ecology Resources* **15:** 337–348.

Chen Q, **Wu X**, **Zhang D. 2020.** Comparison of the abilities of universal, super, and specific DNA barcodes to discriminate among the original species of Fritillariae cirrhosae bulbus and its adulterants. *PLoS One* **15:** e0229181.

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.

Chen YC, **Li Z**, **Zhao YX**, **Gao M**, **Wang JY**, **Liu KW**, **Wang X**, **Wu LW**, **Jiao YL**, **Xu ZL**, **He WG**, **Zhang QY**, **Liang CK**, **Hsiao YY**, **Zhang DY**, **Lan SR**, **Huang L**, **Xu W**, **Tsai WC**, **Liu ZJ**, **Van de Peer Y**, **Wang YD. 2020.** The *Litsea* genome and the evolution of the laurel family. *Nature Communications* **11:** 1675.

Chernomor O, **von Haeseler A**, **Minh BQ. 2016.** Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology* **65:** 997–1008.

China Plant BOL Group. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences, USA* **108:** 19641–19646.

Dong W, **Xu C**, **Li C**, **Sun J**, **Zuo Y**, **Shi S**, **Cheng T**, **Guo J**, **Zhou S. 2015.** *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific Reports* **5:** 8348.

Doyle JJ, **Doyle JL. 1987.** A rapid DNA isolation procedure from small quantities of fresh leaf tissue. *Phytochemistry Bulletin, Botanical Society of America* **19:** 11–15.

Edwards D, **Horn A**, **Taylor D**, **Savolainen V**, **Hawkins JA. 2008.** DNA barcoding of a large genus, *Aspalathus* L. (Fabaceae). *Taxon* **57:** 1317–1327.

Fang J-Y, **Yod K. 1989.** Climate and vegetation in China II. Distribution of main vegetation types and thermal climate. *Ecological Research* **4:** 71–83.

Fazekas AJ, **Kesanakurti PR**, **Burgess KS**, **Percy DM**, **Graham SW**, **Barrett SC**, **Newmaster SG**, **Hajibabaei M**, **Husband BC. 2009.** Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* **9 (S1):** 130–139.

Frazer KA, **Pachter L**, **Poliakov A**, **Rubin EM**, **Dubchak I. 2004.** VISTA: computational tools for comparative genomics. *Nucleic Acids Research* **32:** W273–279.

Hadziavdic K, **Lekang K**, **Lanzen A**, **Jonassen I**, **Thompson EM**, **Troedsson C. 2014.** Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One* **9:** e87624.

Hollingsworth ML, **Andra Clark A**, **Forrest LL**, **Richardson J**, **Pennington RT**, **Long DG**, **Cowan R**, **Chase MW**, **Gaudeul M**, **Hollingsworth PM. 2009.** Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources* **9:** 439–457.

Hollingsworth PM. 2008. DNA barcoding plants in biodiversity hot spots: progress and outstanding questions. *Heredity* **101:** 1–2.

Hollingsworth PM, **Graham SW**, **Little DP. 2011.** Choosing and using a plant DNA barcode. *PLoS One* **6:** e19254.

Jiao L, **He T**, **Dormontt EE**, **Zhang Y**, **Lowe AJ**, **Yin Y. 2019.** Applicability of chloroplast DNA barcodes for wood identification between *Santalum album* and its adulterants. *Holzforschung* **73:** 209–218.

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. *Scientific Reports* **8:** 1945.

Kalyaanamoorthy S, **Minh BQ**, **Wong TKF**, **von Haeseler A**, **Jermiin LS. 2017.** ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14:** 587–589.

Kostermans AJGH. 1957. Lauraceae. *Pengumuman Balai Besar Penjelidikan Kehutanan Indonesia* **57:** 1–64.

Kress WJ, **Erickson DL. 2007.** A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One* **2:** e508.

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

Li H-W, **Li J**, **Huang P-H**, **Wei FN**, **Tsui H-P**, **van der Werff H. 2008.** Lauraceae. In: Wu ZY, Raven PH, Hong DY (Eds.), *Flora of China, Vol. 7. Calycanthaceae-Schisandraceae*. Beijing & St. Louis: Science Press and Missouri Botanical Garden Press, Beijing, China, St. Louis, Missouri, USA, 102–254.

Li L, **Madriñán S**, **Li J. 2016.** Phylogeny and biogeography of *Caryodaphnopsis* (Lauraceae) inferred from low-copy nuclear gene and ITS sequences. *Taxon* **65:** 433–443.

Li X, **Yang Y**, **Henry RJ**, **Rossetto M**, **Wang Y**, **Chen S. 2015.** Plant DNA barcoding: from gene to genome. *Biological Reviews of the Cambridge Philosophical Society* **90:** 157–166.

Little DP. 2014. A DNA mini-barcode for land plants. *Molecular Ecology Resources* **14:** 437–446.

Liu ZF, **Ci XQ**, **Li L**, **Li HW**, **Conran JG**, **Li J. 2017.** DNA barcoding evaluation and implications for phylogenetic relationships in Lauraceae from China. *PLoS One* **12:** e0175788.

Liu Z-F, **Ma H**, **Ci X-Q**, **Li L**, **Song Y**, **Liu B**, **Li H-W**, **Wang S-L**, **Qu X-J**, **Hu J-L**, **Zhang X-Y**, **Conran JG**, **Twyford AD**, **Yang J-B**, **Hollingsworth PM**, **Li J. 2021.** Can plastid genome sequencing be used for species identification in the Lauraceae? *Botanical Journal of the Linnean Society* **197:** 1–14.

Minh BQ, **Schmidt HA**, **Chernomor O**, **Schrempf D**, **Woodhams MD**, **von Haeseler A**, **Lanfear R. 2020.** IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37:** 1530–1534.

Muellner AN, **Schaefer H**, **Lahaye R. 2011.** Evaluation of candidate DNA barcoding loci for economically important

timber species of the mahogany family (Meliaceae). *Molecular Ecology Resources* **11:** 450–460.

Naciri Y, **Linder HP. 2015.** Species delimitation and relationships: the dance of the seven veils. *Taxon* **64:** 3–16.

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

Nie ZL, **Wen J**, **Azuma H**, **Qiu YL**, **Sun H**, **Meng Y**, **Sun WB**, **Zimmer EA. 2008.** Phylogenetic and biogeographic complexity of Magnoliaceae in the Northern Hemisphere inferred from three nuclear data sets. *Molecular Phylogenetics and Evolution* **48:** 1027–1040.

Oh S-H, **Manos PS. 2008.** Molecular phylogenetics and cupule evolution in Fagaceae as inferred from nuclear CRABS CLAW sequences. *Taxon* **57:** 434–451.

Pei N, **Erickson DL**, **Chen B**, **Ge X**, **Mi X**, **Swenson NG**, **Zhang JL**, **Jones FA**, **Huang CL**, **Ye W**, **Hao Z**, **Hsieh CF**, **Lum S**, **Bourg NA**, **Parker JD**, **Zimmerman JK**, **McShea WJ**, **Lopez IC**, **Sun IF**, **Davies SJ**, **Ma K**, **Kress WJ. 2015.** Closely-related taxa influence woody species discrimination via DNA barcoding: evidence from global forest dynamics plots. *Scientific Reports* **5:** 15127.

Percy DM, **Argus GW**, **Cronk QC**, **Fazekas AJ**, **Kesanakurti PR**, **Burgess KS**, **Husband BC**, **Newmaster SG**, **Barrett SC**, **Graham SW. 2014.** Understanding the spectacular failure of DNA barcoding in willows (*Salix*): does this result from a trans-specific selective sweep? *Molecular Ecology* **23:** 4737–4756.

Rieseberg LH, **Soltis DE. 1991.** Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* **5:** 65–84.

Rendón-Anaya M, **Ibarra-Laclette E**, **Méndez-Bravo A**, **Lan T**, **Zheng C**, **Carretero-Paulet L**, **Perez-Torres CA**, **Chacón-López A**, **Hernandez-Guzmán G**, **Chang TH**, **Farr KM**, **Barbazuk WB**, **Chamala S**, **Mutwil M**, **Shivhare D**, **Alvarez-Ponce D**, **Mitter N**, **Hayward A**, **Fletcher S**, **Rozas J**, **Sánchez Gracia A**, **Kuhn D**, **Barrientos-Priego AF**, **Salojärvi J**, **Librado P**, **Sankoff D**, **Herrera-Estrella A**, **Albert VA**, **Herrera-Estrella L. 2019.** The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National Academy of Sciences, USA* **116:** 17081–17089.

Rohwer JG. 1993. Lauraceae. In: Kubitzki K, Rohwer JG, Bittrich V, eds. *The families and genera of vascular plants. Volume 2. Flowering plants. Dicotyledons: magnoliid, hamamelid and caryophyllid families*. Berlin: Springer-Verlag, 366–391.

Rohwer JG, **Li J**, **Rudolph B**, **Schmidt SA**, **van der Werff H**, **Li H-W. 2009.** Is *Persea* (Lauraceae) monophyletic? Evidence from nuclear ribosomal ITS sequences. *Taxon* **58:** 1153–1167.

Rychlik W. 2007. OLIGO 7 primer analysis software. *Methods in Molecular Biology* **402:** 35–60.

Shen Z, **Lu T**, **Zhang Z**, **Cai C**, **Yang J**, **Tian B. 2019.** Authentication of traditional Chinese medicinal herb 'gusuibu' by DNA-based molecular methods. *Industrial Crops and Products* **141:** 111756.

Singh VK, **Mangalam AK**, **Dwivedi S**, **Naik S. 1998.** Primer premier: program for design of degenerate primers from a protein sequence. *BioTechniques* **24:** 318–319.

Song Y, **Yao X**, **Tan Y**, **Gan Y**, **Corlett RT. 2016.** Complete chloroplast genome sequence of the avocado: gene organization, comparative analysis, and phylogenetic relationships with other Lauraceae. *Canadian Journal of Forest Research* **46:** 1293–1301.

Song Y, **Yu WB**, **Tan Y**, **Liu B**, **Yao X**, **Jin J**, **Padmanaba M**, **Yang JB**, **Corlett RT. 2017.** Evolutionary comparisons of the chloroplast genome in Lauraceae and insights into loss events in the magnoliids. *Genome Biology and Evolution* **9:** 2354–2364.

Song Y, **Yu WB**, **Tan YH**, **Jin JJ**, **Wang B**, **Yang JB**, **Liu B**, **Corlett RT. 2019.** Plastid phylogenomics improve phylogenetic resolution in the Lauraceae. *Journal of Systematics and Evolution* **58:** 423–439.

Takahiro M, **Koichi W**, **Ryuichiro T**, **Masafumi F**, **Hiroshi O. 1999.** Distribution of bifidobacterial species in human intestinal microflora examined with 16S rRNA-gene-targeted species-specific primers. *Applied and Environmental Microbiology* **65:** 4506–4512.

Twyford AD. 2014. Testing evolutionary hypotheses for DNA barcoding failure in willows. *Molecular Ecology* **23:** 4674–4676.

Wang FH, **Lu JM**, **Wen J**, **Ebihara A**, **Li DZ. 2016.** Applying DNA barcodes to identify closely related species of ferns: a case study of the Chinese *Adiantum* (Pteridaceae). *PLoS One* **11:** e0160611.

Wang X, **Gussarova G**, **Ruhsam M**, **de Vere N**, **Metherell C**, **Hollingsworth PM**, **Twyford AD. 2018.** DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. *AoB Plants* **10:** ply026.

Wang ZH, **Li J**, **Conran JG**, **Li HW. 2010.** Phylogeny of the Southeast Asian endemic genus *Neocinnamomum* H. Liu (Lauraceae). *Plant Systematics and Evolution* **290:** 173–184.

van der Werff H, **Richter HG. 1996.** Toward an improved classification of Lauraceae. *Annals of the Missouri Botanical Garden* **83:** 409–418.

Yan HF, **Liu YJ**, **Xie XF**, **Zhang CY**, **Hu CM**, **Hao G**, **Ge XJ. 2015.** DNA barcoding evaluation and its taxonomic implications in the species-rich genus *Primula* L. in China. *PLoS One* **10:** e0122903.

Yan LJ, **Liu J**, **Möller M**, **Zhang L**, **Zhang XM**, **Li DZ**, **Gao LM. 2015.** DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya-Hengduan Mountains. *Molecular Ecology Resources* **15:** 932–944.

Yang Y, **Liu B. 2015.** Species catalogue of Lauraceae in China: problems and perspectives. *Biodiversity Science* **23:** 232–236.

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:** 1165–1176.

Zheng W, **Zeng W**, **Tang Y**, **Shi W**, **Cao K. 2018.** Species diversity and biogeographical patterns of Lauraceae and Fagaceae in northern tropical and subtropical regions of China. *Acta Ecologica Sinica* **38:** 8676–8687.

Zhu H, **Chai Y**, **Zhou S**, **Yan L**, **Shi J**, **Yang G. 2016.** Combined community ecology and floristics, a synthetic study on the upper montane evergreen broad-leaved forests in Yunnan, southwestern China. *Plant Diversity* **38:** 295–302.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Map of East Asia with sampling sites of Lauraceae. Sample sites are marked by red dots.
Figure S2. Comparison of 11 plastid genomes using the mVISTA alignment program with *C. hainanensis* as a reference. The *x*-axis represents the early-diverging species of Lauraceae in the plastid phylogenetic relationships. The *y*-axis indicates the average percent identity of sequence similarity in the aligned regions, ranging between 50% and 100%. Genome regions are colour-coded as protein coding, rRNA coding, tRNA coding or conserved non-coding sequences.
Figure S3. ML tree generated using *trnL–ycf2* sequences. Successful species identifications are indicated by black dots.
Figure S4. ML tree generated using *ycf1* sequences. Successful species identifications are indicated by black dots.
Figure S5. ML tree generated using *ndhH–rps15* sequences. Successful species identifications are indicated by black dots.
Figure S6. Species discrimination rates of three barcode matrices (specific plastid matrix: *ycf1* + *ndhH–rps15* + *trnL–ycf2*; standard plastid matrix: *rbcL* + *matK* + *trnH–psbA* and nuclear alone: ITS).
Figure S7. Species and genus discrimination rates of plastid specific, standard matrices and nuclear ITS (all *P* values were determined by one-way ANOVA with Tukey's multiple comparison test).
Note S1. The barcode matrices and datasets.
Table S1. Specimens analysed in the present study for barcodes specific to Lauraceae including details of amplification and sequencing success, with GenBank accession numbers.
Table S2. Information for the barcodes analysed in the present study (specific plastid matrix: *ycf1* + *ndhH–rps15* + *trnL–ycf2*; standard plastid matrix: *rbcL* + *matK* + *trnH–psbA*; and nuclear matrix: ITS).
Table S3. Species discrimination rates of three DNA barcode matrices for different individuals based on ML and NJ methods (specific plastid matrix: *ycf1* + *ndhH–rps15* + *trnL–ycf2*; standard plastid matrix: *rbcL* + *matK* + *trnH–psbA* and nuclear alone: ITS).