Received: 27 October 2016 | Accepted: 16 April 2017

DOI: 10.1002/ece3.3045

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

Single nucleotide polymorphism barcoding of cytochrome c oxidase I sequences for discriminating 17 species of Columbidae by decision tree algorithm

Cheng-Hong Yang[1,2] | Kuo-Chuan Wu[1,3] | Hans-Uwe Dahms[4] | Li-Yeh Chuang[5] | Hsueh-Wei Chang[4,6,7,8]

[1]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

[2]Graduate Institute of Clinical Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[3]Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

[4]Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung, Taiwan

[5]Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan

[6]Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung, Taiwan

[7]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan

[8]Research Center for Natural Products and Drug Development, Kaohsiung Medical University, Kaohsiung, Taiwan

Correspondence
Li-Yeh Chuang, Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan.
Email: chuang@isu.edu.tw
and
Hsueh-Wei Chang, Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung, Taiwan.
Email: changhw@kmu.edu.tw

Funding information
Ministry of Science and Technology, Taiwan, Grant/Award Number: MOST 102-2221-E-151-024-MY3 and MOST 104-2320-B-037-013-MY3; National Sun Yat-sen University-KMU Joint Research Project, Grant/Award Number: #NSYSU-KMU 106-p001

Abstract

DNA barcodes are widely used in taxonomy, systematics, species identification, food safety, and forensic science. Most of the conventional DNA barcode sequences contain the whole information of a given barcoding gene. Most of the sequence information does not vary and is uninformative for a given group of taxa within a monophylum. We suggest here a method that reduces the amount of noninformative nucleotides in a given barcoding sequence of a major taxon, like the prokaryotes, or eukaryotic animals, plants, or fungi. The actual differences in genetic sequences, called single nucleotide polymorphism (SNP) genotyping, provide a tool for developing a rapid, reliable, and high-throughput assay for the discrimination between known species. Here, we investigated SNPs as robust markers of genetic variation for identifying different pigeon species based on available cytochrome c oxidase I (COI) data. We propose here a decision tree-based SNP barcoding (DTSB) algorithm where SNP patterns are selected from the DNA barcoding sequence of several evolutionarily related species in order to identify a single species with pigeons as an example. This approach can make use of any established barcoding system. We here firstly used as an example the mitochondrial gene COI information of 17 pigeon species (Columbidae, Aves) using DTSB after sequence trimming and alignment. SNPs were chosen which followed the rule of decision tree and species-specific SNP barcodes. The shortest barcode of about 11 bp was then generated for discriminating 17 pigeon species using the DTSB method. This method provides a sequence alignment and tree decision approach to parsimoniously assign a unique and shortest SNP barcode for any known species of a chosen monophyletic taxon where a barcoding sequence is available.

KEYWORDS
COI, Decision tree, SNP barcode, species discrimination, species tag

1 | INTRODUCTION

The original idea of DNA barcoding was to use a short DNA sequence as a species-specific marker for species identification and authentication (Hebert, Cywinska, Ball, & deWaard, 2003). It differs from molecular phylogeny approaches as the main purpose is not to analyze evolutionary relationships but to identify an unknown species within a known phylogenetic classification system (a monophylum) using DNA sequences (Kress, Wurdack, Zimmer, Weigt, & Janzen, 2005).

The DNA barcoding technique utilizes a short DNA sequence of the genome that provides enough variation at the species level to unequivocally define a taxon at the species level (http://www.barcodeoflife.org/content/about/what-dna-barcoding). A suitable barcoding gene is commonly of unique reproducibility, sequence versatility, and comparability among different species belonging to the same major kingdom, like animals, plants, or fungi (Yan et al., 2013). For animals and some other eukaryotes, the most successful results were provided by the mitochondrial gene cytochrome c oxidase I (COI), which is the standard gene region in the range of 650 base pairs (bp) (Hebert et al., 2003). Within vertebrate animals, COI was proposed as a potential barcode for the identification of 260 North American bird species (Hebert, Stoeckle, Zemlak, & Francis, 2004). Such information was applied in food authentication and safety (Vandamme et al., 2016) as well as for forensic purposes (Bell, Burgess, Okamoto, Aranda, & Brosi, 2016; Desmyter & Gosselin, 2009; Dubey, Meganathan, & Haque, 2011). Although the COI sequence is conventionally used as an unarbitrary barcode for the discrimination between eukaryotic and animal species, its major shortcoming is that it takes substantial memory and processing time for computational comparisons, particularly when dealing with large data. Such large data are increasingly available with metagenomic approaches to species diversity, even if only using a single promising barcoding gene, like COI (Gao, Jia, & Kong, 2016). Therefore, reducing the amount of noninformative data for computational analysis in species identification remains a challenge with promising applications.

Several genetic markers were developed in the past for the purpose of species and population characterization (Grover & Sharma, 2016). Restriction fragment length polymorphisms (RFLPs) were among the first genomic markers. RFLPs have the disadvantage of being complex, costly, and showing a comparatively low rate of polymorphism. Often considered as a second generation of genomic markers, SSRs (simple sequence repeats) are easy to obtain at lower cost showing a higher polymorphism rate (Gao et al., 2016). Single nucleotide polymorphisms (SNPs) are considered as the third-generation of markers. With the development of next-generation sequencing (NGS) technology and low-cost genome sequencing, a large number of SNPs have recently been identified with the microarray technology based on a standardized protocol (Unterseer et al., 2014). SNP arrays comprise loci with unique positions along chromosomes or genomes, thereby largely avoiding the confusion associated with multiple sequence variants but still at comparatively high costs, for example, specific patterns of SNP. A bioinformatics-based approach in reducing the computational database to an amount of informative gene sequences would be helpful to deal with this problem. Such an approach would effectively reduce the complexity of a given barcoding sequence information in terms of SNP. An economic and easy to apply molecular barcode with a high-throughput possibility is required for determining species for above applications. Here, we show the development of a minimal set of SNP markers that is robust enough to fingerprint a diverse collection of species. In our previously developed software Seq-SNPing (Chang et al., 2009), the SNPs were easy to identify after sequence alignment.

In this article, we propose a decision tree-based SNP barcoding (DTSB) algorithm that automatically generates barcodes for species identification through a decision tree approach. This will facilitate to discriminate biota at species level based on a machine learning technique to analyze given COI sequences from 17 pigeon species. We hypothesize that SNPs from aligned COI sequences of different know species can be used as a new of straightforward way to strip barcoding sequence information from nonvariable and noninformative information to gain shortest variable bp information allowing speedy computational comparisons for the purpose of species discrimination.

2 | MATERIAL AND METHODS

2.1 | Data sources

Seventeen COI sequences of the bird family Columbidae containing four genera were used in this study. These data were obtained from GenBank; the details are provided in Table 1.

2.2 | Decision trees

An introduction of tree-like structures (graph or model) of decision tree algorithm is shown in Figure 1. This simplified algorithm recursively implements from top (root, see circle A of Figure 1) to bottom (leaf nodes, see rectangle R_1 to R_5 of Figure 1). The procedure of generating trees begins with the root, each node in the tree is according to the rule, and it is determined which path from the decision node to another decision node or left node is taken. This procedure continues until arriving at a left node. Two common measures are used: Entropy and Gini index. In general, the decision tree criterion is used for the decision rule that is splitting the decision node to branch into a leaf that is called information gain measurement, which is expressed as (formula 1):

$$\mathrm{Gain}(A,S)=\mathrm{Info}(S)-E(S) \tag{1}$$

where Gain(A, S) is an estimate of the amount of set of uncertainty (S) and set of attributes (A). The entropy $E(S)$ is defined as follows (formulas 2 and 3):

$$E(S)=\sum_{i=1}^{k}\left(\frac{(p_i+n_i)}{(p+n)}\right)\cdot \mathrm{Info}(p_i,n_i) \tag{2}$$

where k is the number of attribute, $S = p + n$, p is the number of samples with positive target, and n is the number of samples with negative target.

$$\mathrm{Info}(p,n)=-D_p\log_2(D_p)-D_n\log_2(D_n) \tag{3}$$

TABLE 1 17 COI sequences of bird species belonging to the pigeon family (Columbidae) from GenBank

Family	Genera	Species name	Length (bp)	Accession no.
Columbidae	*Columbina*	*C. talpacoti*	694	FJ027432.1
		C. picui	694	FJ027428.1
		C. passerina	681	DQ433537.1
		C. inca	674	DQ433529.1
	Columba	*C. oenas*	722	GU571344.1
		C. rupestris	694	GQ481615.1
		C. palumbus	694	GQ481607.1
		C. livia	694	GQ481606.1
	Zenaida	*Z. auriculata*	694	FJ028598.1
		Z. macroura	697	DQ434834.1
		Z. asiatica	652	DQ433271.1
	Patagioenas	*P. picazuro*	694	FJ027979.1
		P. maculosa	694	FJ027973.1
		P. cayennensis	694	FJ027970.1
		P. araucana	694	FJ027968.1
		P. flavirostris	682	DQ433887.1
		P. fasciata	680	DQ433886.1

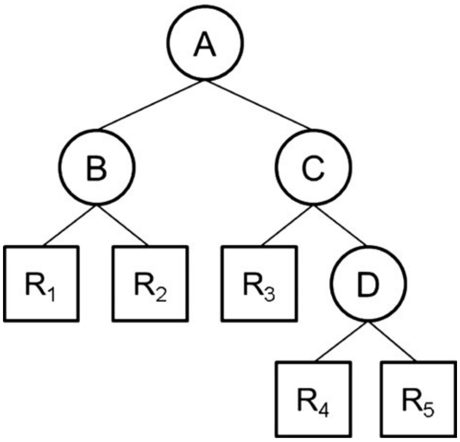


FIGURE 1 An illustration of a simple decision tree. The circle represents decision nodes (parent nodes), which is a decision-maker to make a choice according to decision rules. The rectangle represents leaf nodes (child nodes)

where,

$$D_p = \frac{p}{(p+n)},\ D_n = \frac{n}{(p+n)}$$

The Gini index Gini(*e*) is defined as (formula 4):

$$\text{Gini}(e) = 1 - \sum_i P(i|e)^2 \tag{4}$$

where *P*(*i*|*e*) is conditional probability of category *i* at node *e* of the tree and defined as follows (formulas 5, 6, and 7):

$$P(i|e) = \frac{P(i,e)}{P(e)} \tag{5}$$

$$P(i,e) = \frac{P(\pi_i \cdot N_i(e))}{N_i} \tag{6}$$

$$P(e) = \sum_i P(i,e) \tag{7}$$

where π_i is the prior probability value for class *i*, $N_i(e)$ is the number of records in class *i* of node *e*, and N_i is the number of records of class *i* in the root.

In Gini index, when the value of Gini(*e*) is bigger, the distribution of class of samples is average. Otherwise the distribution of class of samples is unaverage. In a decision tree, three main parameters in the decision tree include (1) the tree constraints, (2) the splitting criterion, and 3) the tree pruning method (De Mántaras, 1991; Gelly, Chiche, & Gracy, 2005; Quinlan, 1986).

2.3 | TSB approach

In this study, we propose a novel approach to implement the species identification process, called decision tree-based SNP barcoding (DTSB) method. The DTSB method is generating a shorter barcode through decision tree construction using COI DNA sequences. The flowchart of the DTSB approach is given as follows (Figure 2): Step 1) data processing, Step 2) decision tree making, and Step 3) barcode sequence creation.

2.3.1 | TSB step 1) data processing

Several COI data were collected from GenBank. Because the variation of the length of COI in different accession numbers, they were aligned with ClustalW program (Tamura et al., 2011), and the resulting

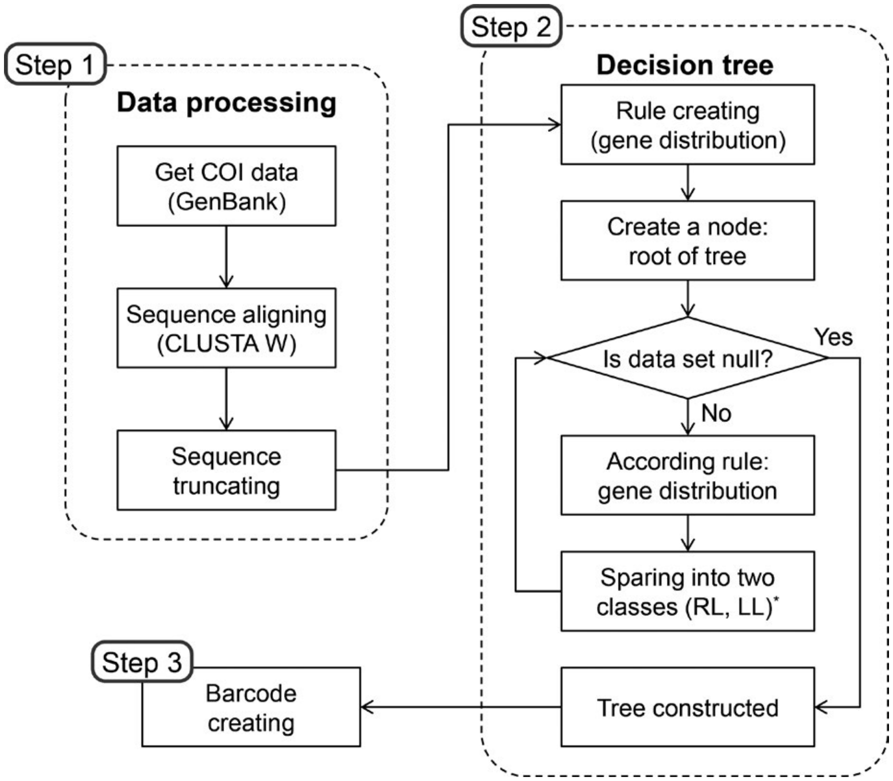

Step 1
Data processing
Get COI data (GenBank)
Sequence aligning (CLUSTA W)
Sequence truncating
Step 2
Decision tree
Rule creating (gene distribution)
Create a node: root of tree
Is data set null?
Yes
No
According rule: gene distribution
Sparing into two classes (RL, LL)*
Tree constructed
Step 3
Barcode creating

FIGURE 2 The flowchart of the DTSB approach. *RL: Right leaf node, LL: Left leaf node

sequences were truncated to keep the same sequence length for further processing.

2.3.2 | TSB step 2) decision tree making for barcoding

Support the data ***X*** are composed of the alignment with *N* sequences (from each species) of the same length of *M* nucleotides, and it can be written as follows (formula 8):

$$X=\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & x_{N,3} & \cdots & x_{N,M} \end{bmatrix} \tag{8}$$

When the nucleotides A, C, G, and T are used in the matrix ***X***, we obtain the distribution ***D*** of nucleotides in each position $p \in [1, M]$ of ***X***, and the distribution ***D*** is represented by (formulas 9 and 10):

$$D=\begin{bmatrix} f_{A1} & f_{A2} & f_{A3} \cdots f_{AM} \\ f_{C1} & f_{C2} & f_{C3} \cdots f_{CM} \\ f_{G1} & f_{G2} & f_{G3} \cdots f_{GM} \\ f_{T1} & f_{T2} & f_{T3} \cdots f_{TM} \end{bmatrix} \tag{9}$$

where

$$f_{ip,\ i\in\{A,\,C,G,\,T\}}=\sum_{k}^{N}(x_{k,p}|i) \tag{10}$$

In the decision tree, the rules are designed to distinguish species and subgroups into two sides (right and left leaves) based on the score ***S*** in each position of sequences. The ***S*** is represented as follows (formula 11):

$$S=\left[\ \mathrm{score}_1 \quad \mathrm{score}_2 \quad \mathrm{score}_3 \quad \cdots \quad \mathrm{score}_M\ \right] \tag{11}$$

where the score at the position (*p*), namely score_p, is calculated as (formulas 12–15):

$$\mathrm{score}_p=\frac{\mathrm{mid}_p-\mathrm{diff}_p}{\mathrm{mid}_p}+\mathrm{weight}_p \tag{12}$$

where

$$\mathrm{mid}_p=\frac{\text{number of data set in node}}{2} \tag{13}$$

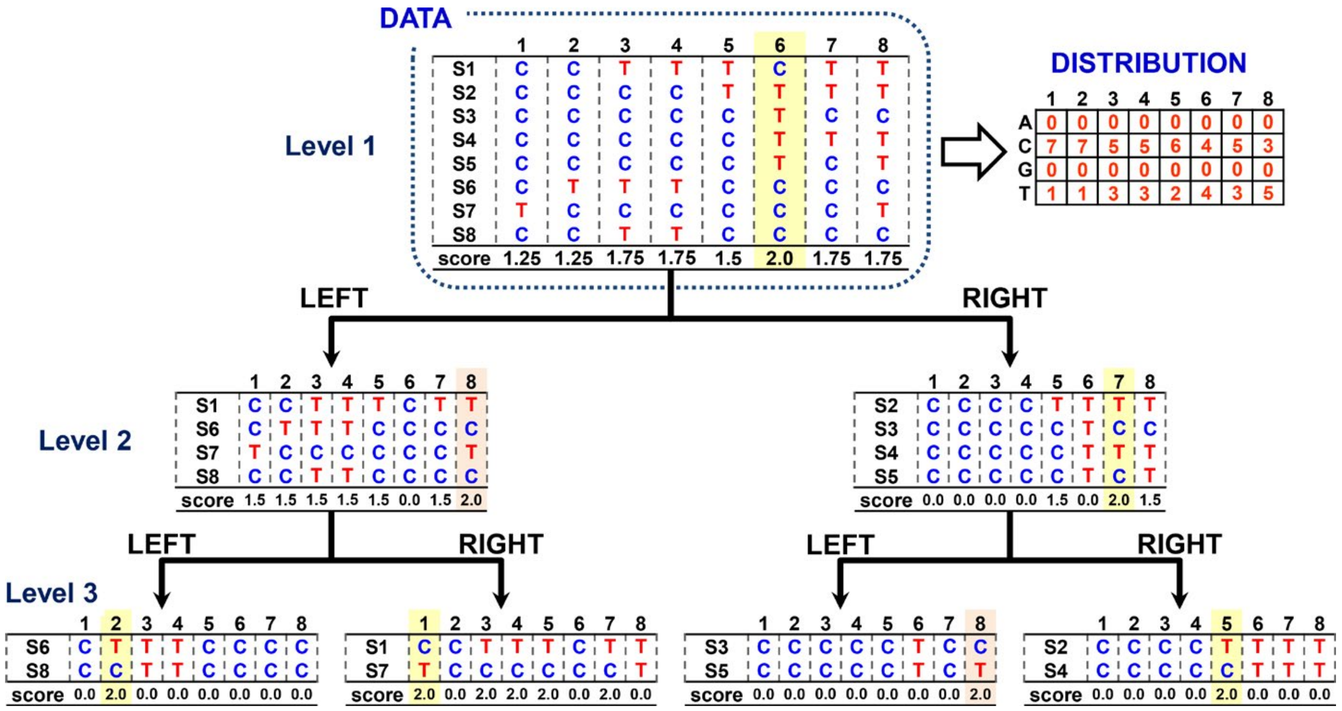
and

$$\mathrm{diff}_p=\min_{i\in\{A,\,C,G,\,T\}}\left\{\left|\mathrm{mid}_p-f_{ip}\right|\right\} \tag{14}$$

and

$$\mathrm{weight}_p=\begin{cases} 0, & \text{if the number of appeared nucleotide type is 1} \\ 1, & \text{if the number of appeared nucleotide type is 2} \\ 0.66, & \text{if the number of appeared nucleotide type is 3} \\ 0.33, & \text{if the number of appeared nucleotide type is 4} \end{cases} \tag{15}$$

Consequently, the species can be divided into two sides according to the score calculation for each score_p. The rest nodes at different levels are performed in the same manner, and the tree is finally constructed.



DATA (Level 1)

	1	2	3	4	5	6	7	8
S1	C	C	T	T	T	C	T	T
S2	C	C	C	C	T	T	T	T
S3	C	C	C	C	C	T	C	C
S4	C	C	C	C	C	T	T	T
S5	C	C	C	C	C	T	C	T
S6	C	T	T	T	C	C	C	C
S7	T	C	C	C	C	C	C	T
S8	C	C	T	T	C	C	C	C
score	1.25	1.25	1.75	1.75	1.5	2.0	1.75	1.75

DISTRIBUTION

	1	2	3	4	5	6	7	8
A	0	0	0	0	0	0	0	0
C	7	7	5	5	6	4	5	3
G	0	0	0	0	0	0	0	0
T	1	1	3	3	2	4	3	5

Level 2 LEFT

	1	2	3	4	5	6	7	8
S1	C	C	T	T	T	C	T	T
S6	C	T	T	T	C	C	C	C
S7	T	C	C	C	C	C	C	T
S8	C	C	T	T	C	C	C	C
score	1.5	1.5	1.5	1.5	1.5	0.0	1.5	2.0

Level 2 RIGHT

	1	2	3	4	5	6	7	8
S2	C	C	C	C	T	T	T	T
S3	C	C	C	C	C	T	C	C
S4	C	C	C	C	C	T	T	T
S5	C	C	C	C	C	T	C	T
score	0.0	0.0	0.0	0.0	1.5	0.0	2.0	1.5

Level 3

	1	2	3	4	5	6	7	8
S6	C	T	T	T	C	C	C	C
S8	C	C	T	T	C	C	C	C
score	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0

	1	2	3	4	5	6	7	8
S1	C	C	T	T	T	C	T	T
S7	T	C	C	C	C	C	C	T
score	2.0	0.0	2.0	2.0	2.0	0.0	2.0	0.0

	1	2	3	4	5	6	7	8
S3	C	C	C	C	C	T	C	C
S5	C	C	C	C	C	T	C	T
score	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0

	1	2	3	4	5	6	7	8
S2	C	C	C	C	T	T	T	T
S4	C	C	C	C	C	T	T	T
score	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0

FIGURE 3 Example of decision tree making in DTSB. On each level, decision trees are made, and left and right sides are generated. S1–S8 indicate the eight sequences with the lengths of eight nucleotides collected from eight species. The number at the top of each table indicates the order of nucleotides. The number at the bottom of each table indicates the scores for each position, for example, score_p

Assuming that we get the "data" of 8 sequences (species) of a length for eight nucleotides (Figure 3), then the "distribution" is counted from "data" and the scores S (score_p) are calculated using formula 12. For example, the positions p_1 and p_6 in level one has eight sequences (species); therefore, the mid_1 and mid_6 are $\frac{8}{2}=4$ (formula 13) and the diff_1 and diff_6 are calculated as follows (formula 14):

$$\text{diff}_1=\min\begin{cases} f_{A1}=|4-0|=4 \\ f_{C1}=\ |4-7|=3 \\ f_{G1}=\ |4-0|=4 \\ f_{T1}=\ |4-1|=3 \end{cases}=3$$

and

$$\text{diff}_6=\min\begin{cases} f_{A6}=|4-0|=4 \\ f_{C6}=\ |4-4|=0 \\ f_{G6}=\ |4-0|=4 \\ f_{T6}=\ |4-4|=0 \end{cases}=0$$

where there are two types in p_1 and p_6 (C and T); hence, weight 1 and weight 6 are 2 (formula 15). The scores are calculated as follows (formula 12):

$$\text{score}_1=\frac{4-3}{4}+1=1.25$$

and

$$\text{score}_6=\frac{4-0}{4}+1=2$$

Therefore, we can get all scores of positions $p_1 \sim p_8$ (see Figure 3) and the maximum score in position p_6 is obtained at level one. All sequences are subgrouped into "left" and "right" sides as branches according to nucleotides (here it is C and T). Subsequently, the subtrees follow the same procedure as described above until the end of tree. This way, the positions p_1, p_2, p_5, p_6, p_7, and p_8 are found (Figure 3). The position p_8 is chosen twice, that is, (1) the left side of level 2 and (2) the right side of level 2 and its left side of level 3. Therefore, short barcode sequences are sometimes available using DTSB.

2.3.3 | TSB step 3) generating barcode sequences

The code 128 (standard) of one dimension barcode is commonly used for alpha-numerical or numerical applications only. We use the website tool (http://www.barcode-generator.org/) to create barcode images.

3 | RESULTS

For the DTSB approach, the results of each step are obtained as follows (Figure 4).

3.1 | Step 1) data processing

Step 1.1. COI data collection: Seventeen COI sequences of bird family Columbidae were collected and aligned as shown in Fig. S1 (Supporting information).

Step 1.2. *Aligning multiple sequences:* After performing MEGA 6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013), the alignment

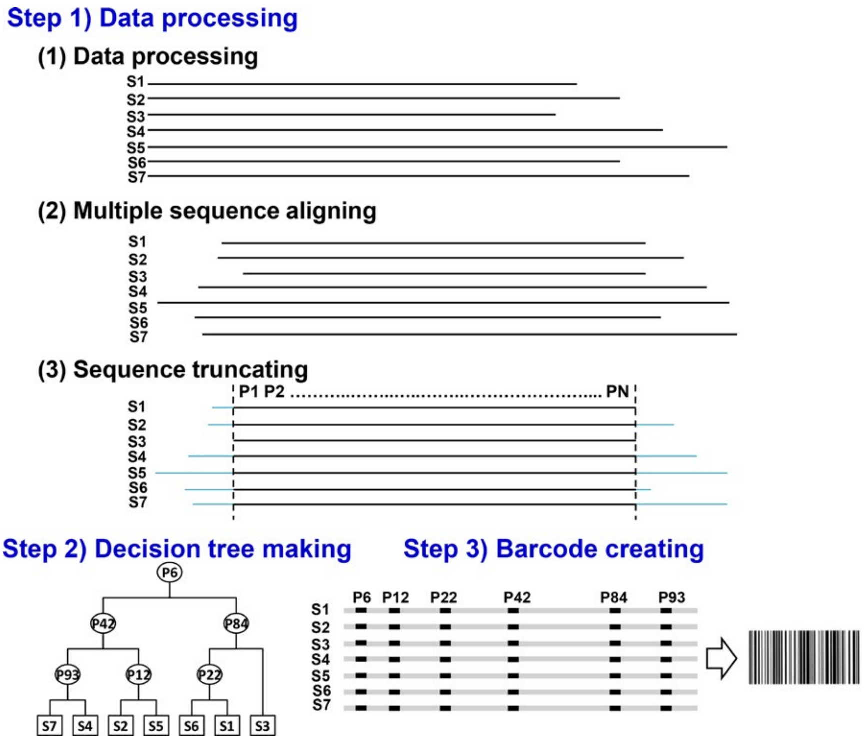


FIGURE 4 Illustration of DTSB approach. Three steps are processed to perform the DTSB method. S1–S7 indicate the example sequences from different species. In step 1, data processing is performed. Light blue lines in step 1 (3) indicate the protruding sequence after alignment. P1–PN indicate the position numbers for the sequence of common COI regions (Fig. S3) after trimming. In step 2, the decision tree making is performed. Subsequently, the species-specific grouping is generated. The P numbers indicate the position numbers, and the S number indicates the example sequences. In step 3, barcode creating is performed. The SNPs identified from the decision tree are visualized and used to generate species-specific SNP barcoding patterns

of these seventeen COI sequences is shown in Fig. S2 (Supporting information).

Step 1.3. *Sequence truncation:* Because the length of COI for different species commonly differs, the sequence of the 5′ and 3′ ends of some COI sequences may be protruding and needs further trimming to generate equal bp lengths with blunt ends. Finally, a common region of the COI sequences was identified as shown in Fig. S3 (Supporting information). This common sequence was then used as the reference sequence for nucleotide position numbering.

3.2 | Step 2) decision tree making

The decision rule was used to construct a decision tree according to the COI gene distribution sparing into two classes as shown in Figure 5. Based on the alignment-generated SNP pattern, the tested 17 species were discriminated from each other using our proposed decision tree algorithm.

3.3 | Step 3) generating barcode sequences

Finally, species-specific COI SNPs generated from the decision tree algorithm were visualized into SNP barcode as shown in Figure 6a. However, the SNP pattern generated from decision tree was listed in the order of decision levels. The selected SNP from the top level was designed to appear first. The level next to the top level appeared next and so on. Accordingly, the number order of selected SNP appeared randomly. After sorting, species-specific SNP barcodes were listed in the order of the position number (Figure 6b). Subsequently, species-specific SNP barcode patterns were generated (Figure 7).

4 | DISCUSSION

Currently, the DNA barcoding is widely used in systematics and species identification in evolutionary, ecological, and conservation research (Austerlitz et al., 2009; DasGupta, Konwar, Mandoiu, & Shvartsman, 2005; Hebert et al., 2003; Kress, Garcia-Robledo, Uriarte, & Erickson, 2015; Meier, Shiyang, Vaidya, & Ng, 2006). The original intention of "barcode" is designed to make a straightforward, reliable, and cheap tag to identify species. Such tag should be similar to tag the goods in the supermarket which can easily obtain the product information through barcodes (Hebert et al., 2003).

Based on given barcodes, our intention is to reduce the nonvariable and noninformative sequence information. For this purpose, we implement an algorithm to process known sequences and obtain the shortest SNP barcode for species identification.

In the present study, we are using COI sequences (~652 to 722 bps) to generate reliably much shorter genetic species tags that are based on the actual nucleotide differences (SNPs) of a given group of species. This way we do not have to consider the bulk of uninformative nucleotide sequences many species share in a pool of phylogenetically related taxa. The DTSB method applied here to a group of 17 pigeon species generates the shortest possible DNA barcode for species identification. Such SNP barcode sequences are easily obtained after sequence alignment; for example, the number of *M* bp for SNP barcode sequence may be identified from *N* species with COI sequences. Each SNP barcode sequence can reliably identify each species of a given taxon, and the SNP barcodes are generated from decision tree algorithm that searches for *P* nodes, where $P = [p_{\min}, p_{\max}]$, $p_{\min} = N/4$, $p_{\max} = N - 1$, and P nodes can be repeatedly selected if needed. For example, we assume that 10 species are

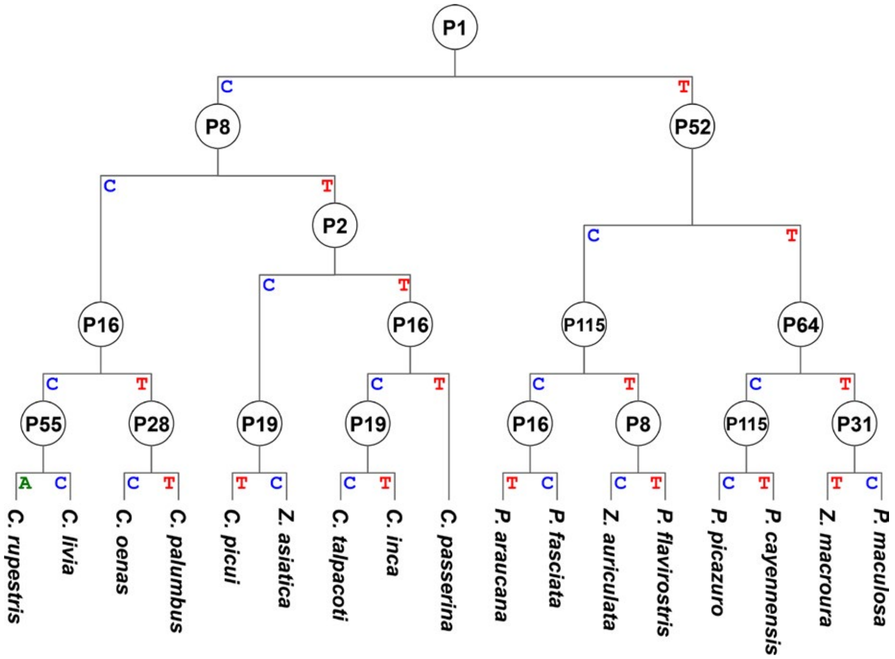
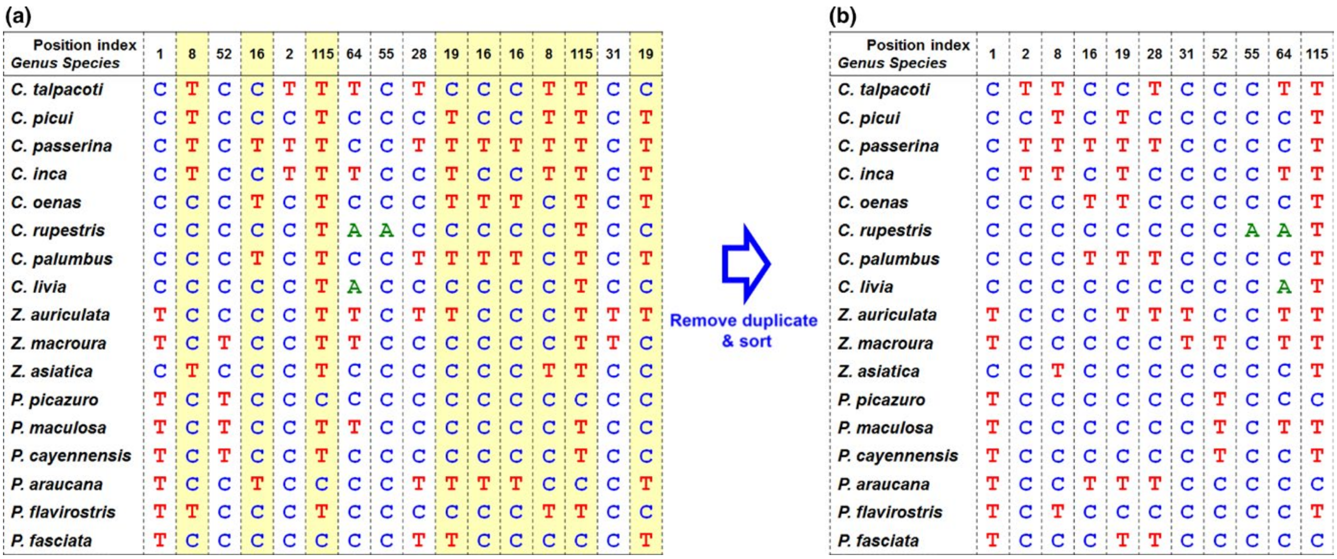


FIGURE 5 Decision tree outcome of the DTSB approach. P1–PN in each circle indicate the position numbers for the sequence of common COI regions (Fig. S3) after trimming. The letter in each turn indicates the nucleotide for grouping. Collection of the different levels becomes the SNP barcode for each test species as shown in Figure 6a


(a)

Position index Genus Species	1	8	52	16	2	115	64	55	28	19	16	16	8	115	31	19
C. talpacoti	C	T	C	C	T	T	T	C	T	C	C	C	T	T	C	C
C. picui	C	T	C	C	C	T	C	C	C	T	C	C	T	T	C	T
C. passerina	C	T	C	T	T	T	C	C	T	T	T	T	T	T	C	T
C. inca	C	T	C	C	T	T	T	C	C	T	C	C	T	T	C	T
C. oenas	C	C	C	T	C	T	C	C	C	T	T	T	C	T	C	T
C. rupestris	C	C	C	C	C	T	A	A	C	C	C	C	C	T	C	C
C. palumbus	C	C	C	T	C	T	C	C	T	T	T	T	C	T	C	T
C. livia	C	C	C	C	C	T	A	C	C	C	C	C	C	T	C	C
Z. auriculata	T	C	C	C	C	T	T	C	T	T	C	C	C	T	T	T
Z. macroura	T	C	T	C	C	T	T	C	C	C	C	C	C	T	T	C
Z. asiatica	C	T	C	C	C	T	C	C	C	C	C	C	T	T	C	C
P. picazuro	T	C	T	C	C	C	C	C	C	C	C	C	C	C	C	C
P. maculosa	T	C	T	C	C	T	T	C	C	C	C	C	C	T	C	C
P. cayennensis	T	C	T	C	C	T	C	C	C	C	C	C	C	T	C	C
P. araucana	T	C	C	T	C	C	C	C	T	T	T	T	C	C	C	T
P. flavirostris	T	T	C	C	C	T	C	C	C	C	C	C	T	T	C	C
P. fasciata	T	C	C	C	C	C	C	C	T	T	C	C	C	C	C	T

(b)

Position index Genus Species	1	2	8	16	19	28	31	52	55	64	115
C. talpacoti	C	T	T	C	C	T	C	C	C	T	T
C. picui	C	C	T	C	T	C	C	C	C	C	T
C. passerina	C	T	T	T	T	T	C	C	C	C	T
C. inca	C	T	T	C	T	C	C	C	C	T	T
C. oenas	C	C	C	T	T	C	C	C	C	C	T
C. rupestris	C	C	C	C	C	C	C	C	A	A	T
C. palumbus	C	C	C	T	T	T	C	C	C	C	T
C. livia	C	C	C	C	C	C	C	C	C	A	T
Z. auriculata	T	C	C	C	T	T	T	C	C	T	T
Z. macroura	T	C	C	C	C	C	T	T	C	T	T
Z. asiatica	C	C	T	C	C	C	C	C	C	C	T
P. picazuro	T	C	C	C	C	C	C	T	C	C	C
P. maculosa	T	C	C	C	C	C	C	T	C	T	T
P. cayennensis	T	C	C	C	C	C	C	T	C	C	T
P. araucana	T	C	C	T	T	T	C	C	C	C	C
P. flavirostris	T	C	T	C	C	C	C	C	C	C	T
P. fasciata	T	C	C	C	T	T	C	C	C	C	C

FIGURE 6 SNP barcode creation using the DTSB approach. (a) The SNP barcode from the decision tree outcome of the DTSB approach. The color background indicates that the nucleotide is repeatedly chosen. (b) The sorted SNP barcode. The position numbers were sorted from small to high, and the corresponding nucleotides were moved together. Repeated nucleotides are processed to keep only one nucleotide for the same position. The sorted SNP barcode is transformed into barcoding patterns as shown in Figure 7

included for classifying by SNP barcoding. $P = [3, 9]$ is calculated in the formula "$P = [p_{\min}, p_{\max}]$." It means that in the most parsimonious case it just needs 3 bp to identify 10 species. In the least informative case, this would need 9 bp.

A permutation of M bp of DNA sequence had a 4^M combination of nucleotides, where "4" represents the four available DNA nucleotides "A, T, C, and G." It is a permutation with repetition problem described as $H_{N-1}^{M} = C_{N-1}^{N-1+M-1}$ when P is $p_{\max}$. In the example of the current study, the sequence data with 652~722 bp from 17 species ($N = 17$) are obtained through sequence alignment and they are trimmed into the same length for 652 bp. After removing the same nucleotides, nucleotides representing 185 bp ($M = 185$) are discovered as SNPs. These provide the possibility of combinational pattern permutation through the following: $H_{N-1}^{M} = C_{N-1}^{N-1+M-1} = C_{17-1}^{17-1+185-1} = C_{16}^{200} > 16\text{E}+21$. This calculation indicates that 17 species with 185 SNPs from the same aligned nucleotides can provide huge combinations, which is larger than the requirement for combination in the making rule of decision tree for correct classification in present study. Therefore, our proposed method can shorten the needs for SNP barcode encoding in species discrimination.

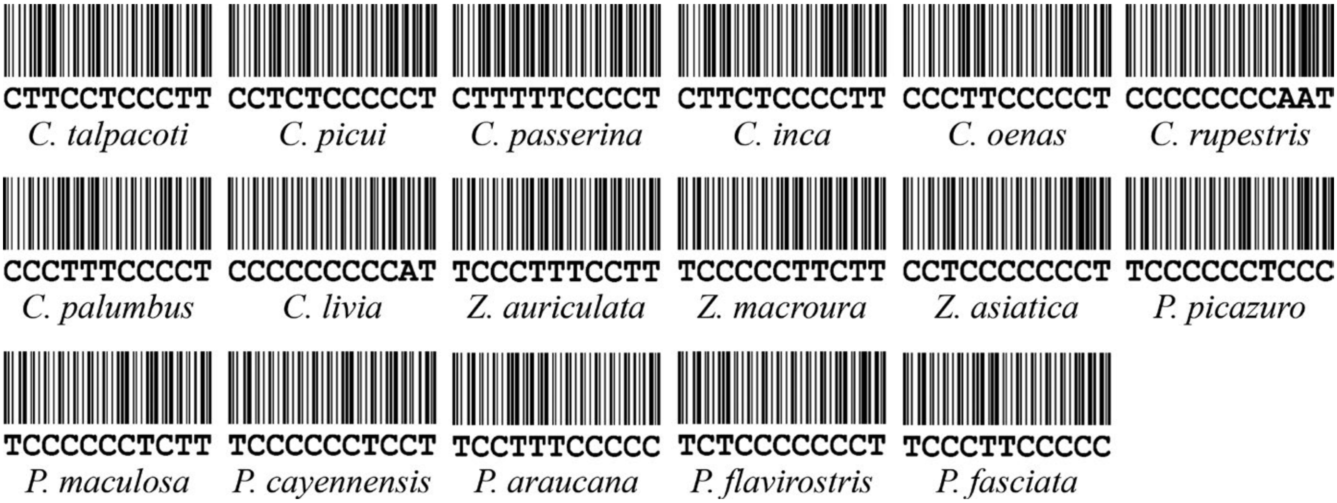


FIGURE 7 SNP barcodes for 17 Columbidae species

Moreover, the DTSB is functional for a monophylum, that is, a phylogenetically coherent group of species. Within 17 Columbidae species, the number of SNP for the phylogenetically distantly related species is commonly higher than for closer related species. Therefore, a DTSB had a high potential to classify species reliably, efficiently, at low cost, and with high-throughput potential with a short species-specific SNP barcode.

Although SNPs can easily be identified, some tasks for computation are still necessary for species identification. The decision tree algorithm (Pei et al., 2015; Quinlan, 1986) is characterized by its high-sensitive property to variations in the training data (Weitschek, Fiscon, & Felici, 2014). Consequently, the decision tree algorithm may be suitable to classify the SNPs generated in the COI barcoding computation but warrants further validation.

The DTSB is limited to discriminate species within the aligned sequences of known species. It is not suitable to extend identified SNP barcodes for the purpose of unknown species identification. In the current study, DTSB is only performed using 17 known COI sequences of the pigeon family, Columbidae. However, this method allows for and is particularly useful to apply for much larger sample sizes. It may also apply to other COI sequences from the BOLD: The Barcode of Life Data System (Ratnasingham & Hebert, 2007) and other non-COI barcoding systems. This holds for the kingdom plants where a universal DNA barcode is still undetermined. However, the CBOL Plant Working Group suggested the so-called core barcode (rbcL+matK) from chloroplast DNA for land plants in general (CBOL Plant Working Group, 2009). Some more alternative barcodes (trnH-psbA, ITS) and the RuBisCo gene of plants were used in plants before (Dong et al., 2014). The sequence of ITS (internal transcribed spacer) is commonly applied for fungi (Schoch et al., 2012; Seifert, 2009).

5 | CONCLUSION

The full length of cytochrome oxidase 1 (COI) sequence is suitable for species identification and phylogenetic inference. However, the full-length "tag" makes it unfriendly for species identification and authentication. To function as the supermarket tagging, we propose a DTSB method to generate the shortest SNP barcode of a given COI sequence to discriminate 17 Columbidae species through this tree decision algorithm. The computational loading for full length (~650 bp) has been reduced to a SNP barcode of about 11 bp. In the future, these species-specific short SNP barcode may provide a reliable, faster species identification. The DTSB method is also flexible to apply to non-COI sequences, like ITS, RuBisCo for tagging biota from other kingdoms than animals.

ACKNOWLEDGMENTS

This work was supported by funds of the Ministry of Science and Technology, Taiwan (MOST 102-2221-E-151-024-MY3 and MOST 104-2320-B-037-013-MY3), the National Sun Yat-sen University-KMU Joint Research Project (#NSYSU-KMU 106-p001).

COMPETING INTERESTS

The authors declare there are no competing interests.

REFERENCES

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., … Laredo, C. (2009). DNA barcode analysis: A comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, *10*(Suppl 14), S10.

Bell, K. L., Burgess, K. S., Okamoto, K. C., Aranda, R., & Brosi, B. J. (2016). Review and future prospects for DNA barcoding methods in forensic palynology. *Forensic Science International: Genetics*, *21*, 110–116.

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 12794–12797.

Chang, H. W., Chuang, L. Y., Cheng, Y. H., et al. (2009). Seq-SNPing: Multiple-alignment tool for SNP discovery, SNP ID identification, and RFLP genotyping. *OMICS: A Journal of Integrative Biology*, *13*, 253–260.

DasGupta, B., Konwar, K. M., Mandoiu, I. I., & Shvartsman, A. A. (2005). DNA-BAR: Distinguisher selection for DNA barcoding. *Bioinformatics*, *21*, 3424–3426.

De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning*, *6*, 81–92.

Desmyter, S., & Gosselin, M. (2009). COI sequence variability between Chrysomyinae of forensic interest. *Forensic Science International. Genetics*, *3*, 89–95.

Dong, W., Cheng, T., Li, C., Xu, C., Long, P., Chen, C., & Zhou, S. (2014). Discriminating plants using the DNA barcode rbcLb: An appraisal based on a large data set. *Molecular Ecology Resources*, *14*, 336–343.

Dubey, B., Meganathan, P. R., & Haque, I. (2011). DNA mini-barcoding: An approach for forensic identification of some endangered Indian snake species. *Forensic Science International. Genetics*, *5*, 181–184.

Gao, L., Jia, J., & Kong, X. (2016). A SNP-based molecular barcode for characterization of common wheat. *PLoS ONE*, *11*, e0150947.

Gelly, J. C., Chiche, L., & Gracy, J. (2005). EvDTree: Structure-dependent substitution profiles based on decision tree classification of 3D environments. *BMC Bioinformatics*, *6*, 4.

Grover, A., & Sharma, P. C. (2016). Development and use of molecular markers: Past and present. *Critical Reviews in Biotechnology*, *36*, 290–302.

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological sciences*, *270*, 313–321.

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, *2*, e312.

Kress, W. J., Garcia-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, *30*, 25–35.

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 8369–8374.

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

Pei, N., Erickson, D. L., Chen, B., Ge, X., Mi, X., Swenson, N. G., … Kress, W. J. (2015). Closely-related taxa influence woody species discrimination via DNA barcoding: Evidence from global forest dynamics plots. *Scientific Reports*, *5*, 15127.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Ratnasingham, S., & Hebert, P. D. (2007). bold: The Barcode of Life Data System. (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*, 355–364.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Author, L. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 6241–6246.

Seifert, K. A. (2009). Progress towards DNA barcoding of fungi. *Molecular Ecology Resources*, *9*(Suppl s1), 83–89.

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

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., … Schon, C. C. (2014). A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, *15*, 823.

Vandamme, S. G., Griffiths, A. M., Taylor, S. A., Di Muri, C., Hankard, E. A., Towne, J. A., … Mariani, S. (2016). Sushi barcoding in the UK: Another kettle of fish. *PeerJ*, *4*, e1891.

Weitschek, E., Fiscon, G., & Felici, G. (2014). Supervised DNA Barcodes species classification: Analysis, comparisons and results. *BioData Mining*, *7*, 4.

Yan, D., Luo, J. Y., Han, Y. M., Peng, C., Dong, X. P., Chen, S. L., … Xiao, X. H. (2013). Forensic DNA barcoding and bio-response studies of animal horn products used in traditional medicine. *PLoS ONE*, *8*, e55854.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Yang C-H, Wu K-C, Dahms H-U, Chuang L-Y, Chang H-W. Single nucleotide polymorphism barcoding of cytochrome c oxidase I sequences for discriminating 17 species of Columbidae by decision tree algorithm. *Ecol Evol*. 2017;7: 4717–4725. https://doi.org/10.1002/ece3.3045