Decision Tree Algorithm–Generated Single-Nucleotide Polymorphism Barcodes of *rbcL* Genes for 38 Brassicaceae Species Tagging

Evolutionary Bioinformatics
Volume 14: 1–9

DOI: 10.1177/1176934318760856

Cheng-Hong Yang[1,2], Kuo-Chuan Wu[1,3], Li-Yeh Chuang[4] and Hsueh-Wei Chang[5,6,7]

[1]Department of Electronic Engineering, National Kaohsiung University of Science and Technology, 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 Science and Technology, Kaohsiung, Taiwan. [4]Department of Chemical Engineering, Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan. [5]Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung, Taiwan. [6]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan. [7]Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung, Taiwan.

ABSTRACT: DNA barcode sequences are accumulating in large data sets. A barcode is generally a sequence larger than 1000 base pairs and generates a computational burden. Although the DNA barcode was originally envisioned as straightforward species tags, the identification usage of barcode sequences is rarely emphasized currently. Single-nucleotide polymorphism (SNP) association studies provide us an idea that the SNPs may be the ideal target of feature selection to discriminate between different species. We hypothesize that SNP-based barcodes may be more effective than the full length of DNA barcode sequences for species discrimination. To address this issue, we tested a *r*ibulose diphosphate carboxylase (*rbcL*) *S*NP *b*arcoding (RSB) strategy using a decision tree algorithm. After alignment and trimming, 31 SNPs were discovered in the *rbcL* sequences from 38 Brassicaceae plant species. In the decision tree construction, these SNPs were computed to set up the decision rule to assign the sequences into 2 groups level by level. After algorithm processing, 37 nodes and 31 loci were required for discriminating 38 species. Finally, the sequence tags consisting of 31 *rbcL* SNP barcodes were identified for discriminating 38 Brassicaceae species based on the decision tree–selected SNP pattern using RSB method. Taken together, this study provides the rational that the SNP aspect of DNA barcode for *rbcL* gene is a useful and effective sequence for tagging 38 Brassicaceae species.

KEYWORDS: Decision tree, *rbcL* gene, species tag, SNP, barcoding

RECEIVED: October 21, 2017. **ACCEPTED:** January 24, 2018.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by funds of the Ministry of Science and Technology, Taiwan (MOST 105-2221-E-151-053-MY2 and MOST 104-2320-B-037-013-MY3) and the National Sun Yat-sen University-KMU Joint Research Project (#NSYSU-KMU 107-p001).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Li-Yeh Chuang, Department of Chemical Engineering, Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung 840, Taiwan. Email: chuang@mail.isu.edu.tw;

Hsueh-Wei Chang, Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung 807, Taiwan. Email: changhw@kmu.edu.tw

Introduction

DNA barcoding techniques became popular because the primer sets for the DNA barcode sequencing are convenient to reach.[1] Current works of barcoding are based on the data accumulation for DNA barcode sequences. However, the original idea of "DNA barcoding" was to use an as short as possible DNA sequence to identify a species as reliable as possible. Such molecular species-specific marker was proposed first by Paul Hebert in 2003 for animals.[2]

As yet, the intention of a short identification character is not realized. This is mainly due to the barcode sequence taking an intact gene with all its same and uninformative sequence parts. This is commonly long and the species number too large to make computational discriminations straightforward. This also holds for the chloroplast-encoded ribulose diphosphate carboxylase (*rbcL*) gene.[3,4] This is commonly used for DNA barcoding in plants.[5–8] Its DNA sequence ranges from 1086 to 1410 bp (base pairs). Hence, it is necessary to reduce the computation time for *rbcL* barcoding for tagging species. Detailed inspection revealed that most of the nucleotides of different species were the same in the *rbcL* gene. This evidence inspired us to focus just on the nucleotides which were variable on single-nucleotide polymorphisms (SNPs).[9–14] Accordingly, we proposed a hypothesis that the number of SNPs is smaller than the nucleotide numbers for full-length barcoding sequences. The complexity of data was reduced when SNPs within the full-length barcode sequences were used.

Recently, the concept of SNP barcoding was widely reported with respect to different diseases[15,16] and species identifications.[17–19] For example, the nucleotide signatures consisting of 27 and 37 SNPs of the internal transcribed spacer 2 (ITS2) sequences were used to identify medicinal herbs such as *Panax quinquefolius* (American ginseng) products[20] and *Angelicae sinensis* radix (Danggui).[17] To add another example, 42 SNPs from different chromosomes of the malaria causing protist *Plasmodium vivax* genome were reported to accurately identify *P vivax* infections.[18] A total of 62 SNPs were identified to discriminate between

known strains of the tuberculosis-causing prokaryote *Mycobacterium tuberculosis*.[19] Although SNP barcoding is helpful for species identification in these studies,[17–20] these SNPs were collected together without further shortening the SNP number for essential species identification by computation.

The Brassicaceae family (syn. Cruciferae, mustard or crucifer plants) contains the most common species of the order Brassicales (about 3700 species).[21] Brassicaceae have 4 petals (cross-shaped) and 6 stamina which are distinguished morphologically from other families. The Brassicaceae provide the prominent angiosperm model species, especially *Arabidopsis thaliana* for plant molecular studies.[21] The Brassicaceae were also used for genetic studies in several fields of plant research and the number of species continuously increased.[21,22] Accordingly, it is suitable to choose several Brassicaceae species for species tagging study.

In this study, we proposed an *rbcL SNP barcoding* (RSB) strategy using a decision tree algorithm[8,23–26] for species identification. We used the example of 38 Brassicaceae species for this particular study. We generated an *rbcL* barcode with 31 SNPs that allowed us to discriminate between 38 species using the RSB method for the first time.

Materials and Methods

Data resources

In total, 38 *rbcL* sequences of Brassicaceae plants were downloaded from GenBank (Table 1).[27]

RSB approach

In this article, we proposed an RSB method to apply the decision tree model to classify 38 Brassicaceae species using *rbcL* barcodes that were retrieved from GenBank. The flowchart of the RSB approach is described as follows (Figure 1): Step 1—data processing, step 2—decision tree construction, and step 3—barcode creation. The detailed procedures are explained in the following sections.

Step 1—Data processing. In total, 38 *rbcL* sequences from different Brassicaceae plants were retrieved from GenBank and aligned with ClustalW built in MEGA 7.[28] Subsequently, the 5′ and 3′ protruding sequences were trimmed to the same length for further analysis.

Step 2—Decision tree construction. A similar strategy of nucleotide selection and rule making was outlined before.[29,30] Support data $\boldsymbol{X}$ consisted of the aligned sequences of N species with the same trimmed length for M nucleotides: These data are represented as follows (formula (1)):

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & x_{N,3} & \cdots & x_{N,M} \end{bmatrix} \quad (1)$$

When nucleotides (A, C, G, and T) are included in matrix $\boldsymbol{X}$, the distribution $\boldsymbol{D}$ of nucleotides in each position $p \in [1, M]$ of $\boldsymbol{X}$ is represented as follows (formula (2)):

$$\boldsymbol{D} = \begin{bmatrix} f_{A1} & f_{A2} & f_{A3} & \cdots & f_{AM} \\ f_{C1} & f_{C2} & f_{C3} & \cdots & f_{CM} \\ f_{G1} & f_{G2} & f_{G3} & \cdots & f_{GM} \\ f_{T1} & f_{T2} & f_{T3} & \cdots & f_{TM} \end{bmatrix} \quad (2)$$

where the frequency f of nucleotides in each position is described as follows (formula (3)):

$$f_{ip, i \in \{A,C,G,T\}} = \sum_{k=1}^{N} \left(x_{k,p} \mid i \right) \quad (3)$$

In the decision tree, rules are decided to subgroup species into 2 sides (LL [left leaf] and RL [right leaf] in Figure 1) depending on the score $\boldsymbol{S}$ in each position of sequences, named $score_p$, $\boldsymbol{S}$ and $score_p$. They are represented as follows (formulas (4) and (5)):

$$\boldsymbol{S} = \begin{bmatrix} score_1 & score_2 & score_3 & \cdots & score_M \end{bmatrix} \quad (4)$$

$$score_p = \frac{mid_p - diff_p}{mid_p} + weight_p \quad (5)$$

where mid_p represented half of the number of the data set in the each node; $diff_p$ is the minimum value of $|mid_p - f_{ip}|$; and $weight_p$ is set as 0, 0.33, 0.66, and 1 if the number of appeared nucleotide type is 1, 3, 4, and 2, respectively. When the number of nucleotide types is 2, we set the weight as the highest value, ie, 1, to separate these data into 2 sides for even distribution. Accordingly, species can subgroup at 2 sides based on each $score_p$. Theoretically, any loci with the same score may be able to perform decision tree construction. However, we choose the first appearing nucleotide starring from the beginning in aligned sequences for convenience. The reason for scoring nucleotide in each locus is to divide multiple species sequences into 2 groups correctly for decision tree construction. For different levels, the nodes are subgrouped in the same way and finally the decision tree is constructed. To further explain the calculation processing of mid_p, $diff_p$, $weight_p$, $score_p$, and decision tree construction, an example is provided in Supplementary file S1.

Step 3—Barcode creation. The barcode-generator Web site tool (http://www.barcode-generator.org/) was chosen to create barcode images according to the SNP barcode sequence of the *rbcL* gene. In general, species-specific SNP barcodes forming step 2 were copied and pasted to the window of the barcode-generator Web site with the setting of Code 128 (standard).

Table 1. A total of 38 *rbcL* sequences of the plant family Brassicaceae from GenBank.

SPECIES NAME	LENGTH, BP	ACCESSION NO.	POSITION[a]
Aethionema grandiflora	1347	AY167983.1	114–1167
Anchonium elichrysifolium	1332	FN594834.1	109–1162
Arabidopsis thaliana	1292	AY174633.1	82–1135
Arabis glabra	1154	DQ310542.1	30–1083
Berteroa incana	1348	KM360667.1	120–1173
Biscutella laevigata	1366	KF602144.1	138–1191
Boechera divaricarpa	1381	JX848436.1	140–1193
Bunias orientalis	1408	KM360682.1	120–1173
Cakile maritima	1347	AY167981.1	114–1167
Calepina irregularis	1300	HE616642.1	99–1152
Capsella bursa-pastoris	1384	KM360691.1	96–1149
Chorispora tenella	1380	FN594833.1	128–1181
Cochlearia acaulis	1366	FN594827.1	113–1166
Conringia planisiliqua	1152	JN847840.1	29–1082
Crucihimalaya mollissima	1324	FN594843.1	91–1144
Descurainia sophia	1380	JX848439.1	139–1192
Dontostemon integrifolius	1202	HE616652.1	1–1054[b]
Draba incana	1345	KM360756.1	120–1173
Erysimum capitatum	1347	AY167980.1	114–1167
Halimolobos diffusus	1374	FN594846.1	114–1167
Heliophila pubescens	1399	AM234933.1	111–1164
Hesperis matronalis	1401	KM360815.1	113–1166
Iberis sempervirens	1408	KM360830.1	120–1173
Isatis pachycarpa	1392	FN594830.1	138–1191
Lepidium banksii	1324	KT626727.1	114–1167
Lignariella serpens	1341	JQ933388.1	111–1164
Lobularia maritima	1345	KM360861.1	120–1173
Megacarpaea polyandra	1383	JQ933404.1	111–1164
Noccaea caerulescens	1380	FN594826.1	128–1181
Notothlaspi australe	1318	KT626750.1	114–1167
Olimarabidopsis pumila	1084	DQ310543.1	30–1083
Pachycladon novaezelandiae	1340	FN594852.1	111–1164
Pegaeophyton nepalense	1336	JQ933435.1	111–1164
Physaria arenosa	1379	JX848443.1	138–1191
Rorippa divaricata	1324	KT626842.1	114–1167
Sisymbrium irio	1347	AY167982.1	114–1167
Smelowskia tibetica	1336	JQ933355.1	111–1164
Thlaspi arvense	1345	KM361012.1	120–1173

Abbreviation: bp, base pair.
[a]The position is listed in the reference of its own accession number.
[b]The reference sequence for single-nucleotide polymorphism barcoding and position.

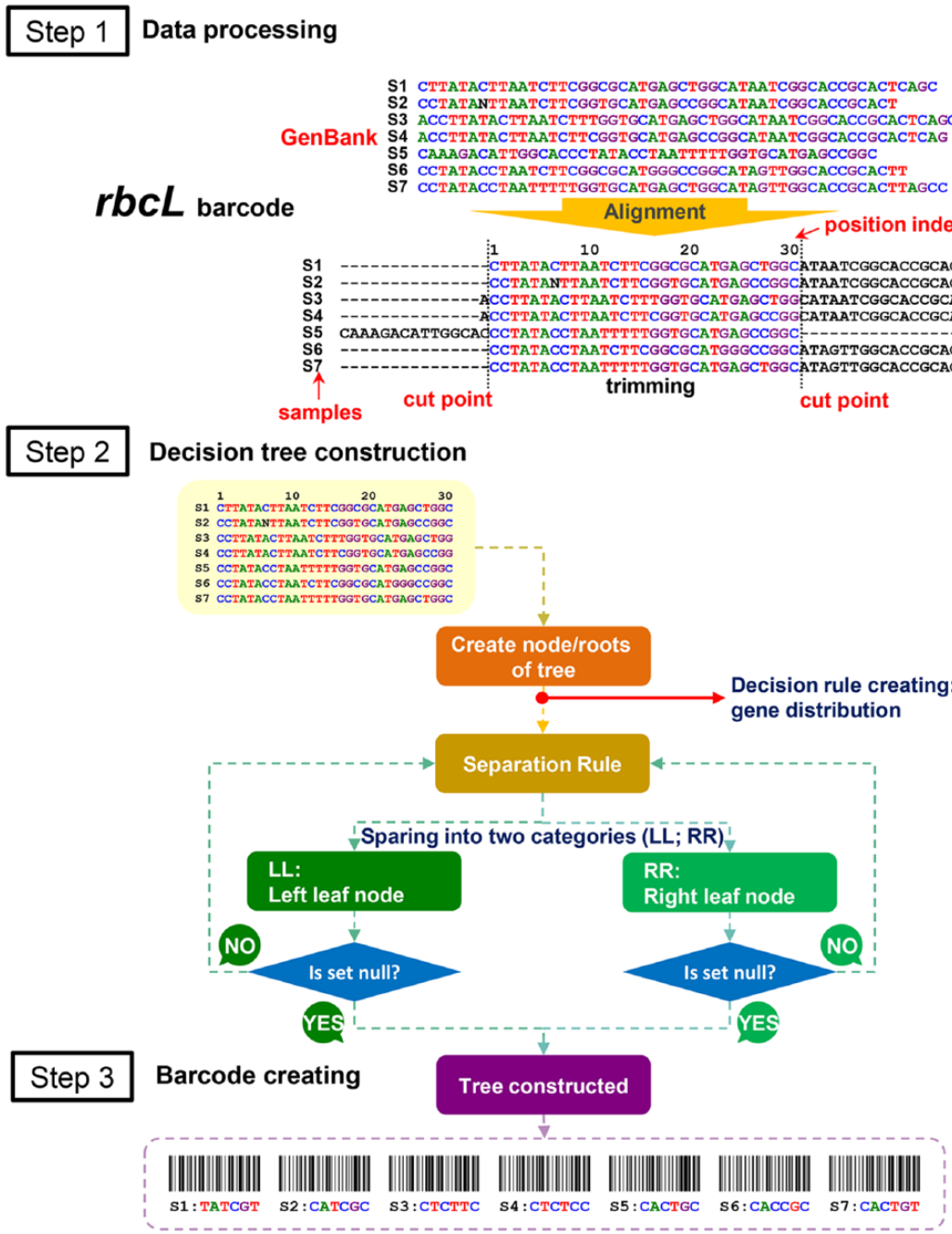

Figure 1. Flowchart of *rbcL* SNP barcoding. Three steps are processed to perform the DTSB method: Step 1—data processing, step 2—decision tree construction, and step 3—barcode creation. For step 1, 7 *rbcL* sequences from different species (S1-S7) were retrieved from GenBank. After alignment, the protruding sequences were trimmed to generate the same length for pretested *rbcL* sequences. For step 2, the aligned sequences were fed for decision tree processing, such as creation of a node root of tree, decision rule creating, and separation rule for sparing into 2 categories (left leaf [LL] node and right leaf [RL] node). For step 3, the tree was constructed and the SNP barcodes of *rbcL* gene for species tagging were generated. Here, hypothesized sequences and SNP barcodes provide an example for species tags. SNP indicates single-nucleotide polymorphism.

After clicking "create barcode," the species-specific SNP barcodes were visualized as strip barcodes.

Results

Step 1—Data processing

Step 1.1. Retrieval of sample *rbcL* sequences: 38 *rbcL* sequences of different Brassicaceae plants (Table 1) are downloaded and available at http://140.127.112.213/Brassicaceae.zip.

Step 1.2. Sequence alignment. Using MEGA 7,[28] 38 *rbcL* sequences were aligned and mostly matched with 5′ and 3′ protruding sequences because they were not of the same length. The alignment status for these aligned sequences is available at http://140.127.112.213/Brassicaceae.zip.

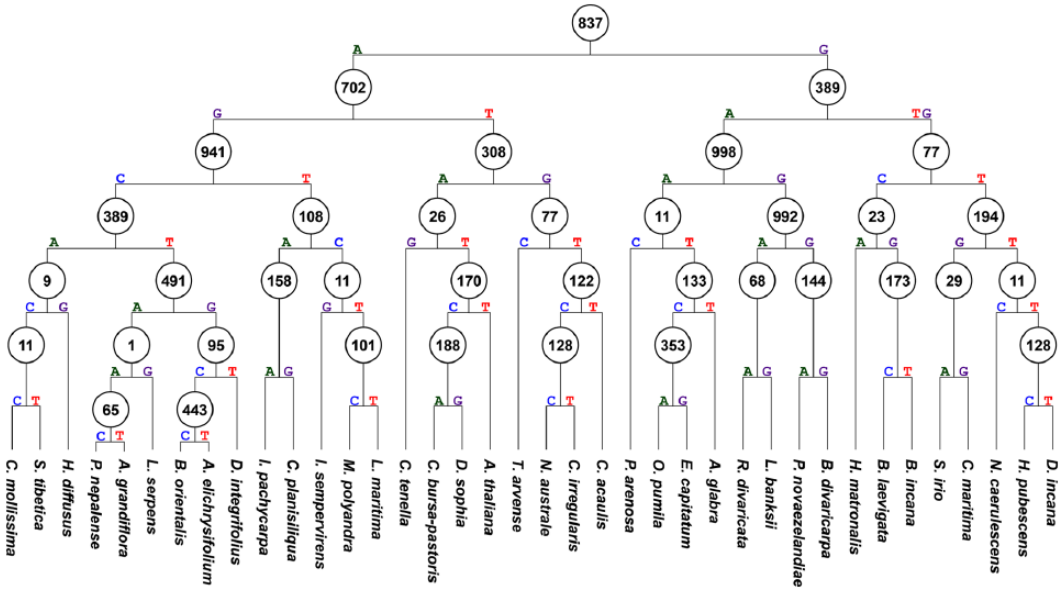

Figure 2. Decision tree making of 38 aligned *rbcL* sequences. The number within each circle is the nucleotide position of the trimmed alignment sequence of *rbcL*. The decision making starts from top to down sequentially level by level. In each level, the letters (nucleotides) in the left and right sides indicate the left leaf (LL) and right leaf (RL) nodes as shown in Figure 1. After collecting the nucleotides from top to down levels, the SNP barcode of *rbcL* can present the tag sequence for different species as shown in Figures 3 and 4. SNP indicates single-nucleotide polymorphism.

Step 1.3. Sequence trimming. To get the same length of *rbcL* sequences for further computational processing, the 5′ and 3′ terminals with protruding sequences were trimmed. The resulting trimmed *rbcL* sequences are available at http://140.127.112.213/Brassicaceae.zip. This trimmed *rbcL* sequence was the same as the sequence from accession no. HE616652.1 (Table 1) and it was regarded as the reference sequence for SNP positioning for the next step. The ranges of this trimmed *rbcL* sequence were listed on the right side of Table 1 for each species.

Step 2—Decision tree construction

Based on our proposed decision rules, the nucleotide distribution (D) and $score_p$ were calculated. Subsequently, the nucleotide with a maximum $score_p$ was identified for subgroup. Among 38 Brassicaceae species, the decision tree was constructed as shown in Figure 2.

For example, *Castanea mollissima* (the left species in Figure 2) required only 6 nodes (837, 702, 941, 389, 9, and 11; yellow background in Figure 3) for the species-specific SNP tag sequences. The others follow the same rule. As different species may need different nodes, we collected all loci and assigned them for the SNP barcode for species discrimination. Finally, 37 nodes and 31 loci were required for discriminating 38 species because some of these loci appeared repeatedly (Figure 3). Finally, the sequence tags for the 38 example species were identified based on the decision tree–selected SNP pattern.

Step 3—Barcode creation

Species-specific *rbcL* SNPs generated from the decision tree algorithm originally appeared from top to down. After sorting, the nodes were listed from small to big numbers as shown in Figure 3. Subsequently, 38 species-specific *rbcL* SNP barcodes were demonstrated in a 1-dimensional barcode pattern (Figure 4).

Discussion

The original idea of "barcode" is designed to make a tag to identify species. However, most barcoding approaches commonly use whole gene sequences of nucleotides with long stretches of uninformative sequence portions but not more targeted characteristics. Here, we developed an RSB method to use the variable parts of otherwise much longer and uninformative *rbcL* barcode for discriminating 38 species belonging to the plant family Brassicaceae. Based on *rbcL* sequences, we developed the RSB to generate the shortest possible DNA barcode with sufficient discriminative power to identify species.

After sequence alignment, the SNP barcode sequences were easily obtained from the decision tree algorithm. In a decision tree model, each sample (species) was separated from root to node or node to node.[31,32] Each root and node needs 1 SNP. If the selected SNPs are not repetitive, the theoretical maximum number of SNPs required to distinguish between each other is N − 1. When the number of chosen SNPs is higher than the number of species, then sufficient species-specific SNP barcodes are available in our proposed RSB method. In the other words, there are 4 possible DNA nucleotides (ie, A, T, C, and G) in each

position index / species	2	9	11	23	26	29	65	68	77	95	101	108	122	128	133	144	158	170	173	188	194	308	353	389	443	491	702	837	941	992	998
A. grandiflora	A	C	T	G	T	G	T	G	C	T	C	C	C	C	C	A	T	C	C	A	T	G	G	T	T	A	G	A	C	G	G
A. elichrysifolium	A	C	T	G	T	G	T	G	T	C	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	G	G	A	C	G	G
A. thaliana	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	T	C	A	T	A	G	A	T	A	T	A	C	G	A
A. glabra	A	C	T	G	T	G	T	G	T	T	C	C	C	C	T	A	G	C	C	A	T	A	G	A	T	A	G	G	C	G	A
B. incana	A	C	T	G	T	G	T	G	C	T	C	C	C	C	C	A	G	C	T	A	T	G	G	T	T	A	G	G	C	G	G
B. laevigata	A	C	T	G	T	G	T	G	C	T	C	C	C	C	C	A	G	C	C	G	T	G	G	T	T	A	G	G	C	G	G
B. divaricarpa	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	G	G	C	C	A	T	G	G	A	T	A	G	G	C	G	G
B. orientalis	A	C	T	G	T	G	T	G	T	C	C	C	C	C	C	A	G	C	C	A	T	G	G	T	C	G	G	A	C	G	A
C. maritima	A	C	T	G	T	G	T	G	T	C	C	C	C	C	C	A	G	C	C	A	G	G	G	T	T	A	G	G	T	G	G
C. irregularis	A	C	T	G	T	G	T	G	T	T	C	C	C	T	C	A	G	C	C	A	T	G	G	T	T	A	T	A	C	G	G
C. bursa-pastoris	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	A	T	A	T	A	C	G	A
C. tenella	A	C	T	G	G	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	T	T	G	T	A	C	G	A
C. acaulis	A	C	T	G	T	G	T	G	T	T	C	C	T	C	C	A	G	C	C	A	T	G	G	T	T	A	T	A	C	G	G
C. planisiliqua	A	C	T	G	T	G	T	G	T	T	C	A	C	C	C	A	G	C	C	A	G	G	G	T	T	A	G	A	T	G	C
C. mollissima	A	C	C	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	A	T	A	G	A	C	G	G
D. sophia	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	G	T	A	G	A	T	A	T	A	C	G	G
D. integrifolius	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	G	G	A	C	G	G
D. incana	A	C	T	G	T	G	T	G	T	T	C	C	C	T	C	A	G	C	C	A	T	G	G	G	T	A	G	G	C	G	G
E. capitatum	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	A	T	A	G	G	C	G	A
H. diffusus	A	G	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	A	T	A	G	A	C	G	G
H. pubescens	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	G	G	C	G	G
H. matronalis	A	C	T	A	T	G	T	G	C	C	C	C	C	C	C	A	G	C	A	A	T	G	G	T	C	G	G	G	C	G	G
I. sempervirens	A	C	G	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	G	C	A	G	A	T	G	G
I. pachycarpa	A	C	T	G	T	G	T	G	T	T	C	A	C	C	C	A	A	C	C	A	G	G	G	T	T	A	G	A	T	G	C
L. banksii	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	A	A	T	G	G	A	T	A	G	G	T	A	G
L. serpens	G	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	G	A	C	G	C
L. maritima	A	C	T	G	T	G	T	G	T	T	T	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	G	A	T	G	G
M. polyandra	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	G	A	T	A	T
N. caerulescens	A	C	C	A	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	G	G	C	G	G
N. australe	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	T	A	C	G	G
O. pumila	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	A	A	T	A	T	G	C	G	A
P. novaezelandiae	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	A	T	A	G	G	C	G	G
P. nepalense	A	C	T	G	T	G	C	G	C	T	C	C	T	C	C	A	G	C	C	A	T	G	G	T	T	A	G	A	C	G	G
P. arenosa	A	C	C	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	A	T	A	G	A	T	A	G	G	C	G	A
R. divaricata	A	C	T	G	T	G	T	A	C	T	C	C	C	C	C	A	G	C	C	A	T	G	G	A	T	A	G	G	C	A	G
S. irio	A	C	T	G	T	A	T	A	T	T	C	C	C	C	C	A	G	C	C	A	G	G	G	T	T	A	G	G	T	G	G
S. tibetica	A	C	T	G	T	G	T	G	T	T	C	C	C	C	C	A	G	C	C	G	T	G	G	A	T	A	G	A	C	G	G
T. arvense	A	C	T	G	T	G	T	G	C	T	C	C	C	C	C	A	G	C	C	A	T	G	G	T	T	A	T	A	C	G	G

Figure 3. Sorting of SNP barcode of *rbcL* sequences from 38 Brassicaceae species. The nucleotides chosen for a decision tree construction depend on the rule of decision tree rather than the order of nucleotide position. Only after sorting, the nucleotides appear in the order of position. Yellow background indicates the minimal SNPs for generating species-specific SNP patterns. The position index is based on the trimmed *rbcL* sequence which is the same as the sequence from accession no. HE616652.1 (Table 1). SNP indicates single-nucleotide polymorphism.

position, it means a permutation of M base pair of DNA sequence had number of $4M$ combinations which is minimum number of SNP difference in the barcodes. In our data, the combination can calculate as $4M > N$, ie, $43 = 64 > 38$. Consequently, it just needs 3-bp sequence in this case. Accordingly, our proposed RSB is tolerant for many species discrimination theoretically.

In our study, the lengths of these sequences ranged from 1086 to 1410 bp from 38 species. After alignment and trimming to the same length for 1054 bp, 155 nucleotides (SNPs) are found. Over a billion of permutation possibilities for SNP combinational patterns can be calculated following the mathematics of combination formula for 155 SNPs. In total, 31 SNPs in 38 species samples were chosen in this study (Figure 4). However, these SNPs may not always conserve. Alternatively, we can use the remaining SNPs, ie, $155 - 31 = 124$, which is allowed to process another computation for the 38 species. Once the SNP frequency changes with variation, the score for each nucleotide is changed accordingly. Finally, other suitable SNPs are able to collect after new computation and generate new SNP barcodes for species tagging. Therefore, our proposed method is tolerated to spontaneous mutations for test species sequences. Among these 38 Brassicaceae species, some are genetically similar and some are distinct in terms of the result of phylogenetic trees constructed by neighbor joining method (Supplementary file S2).[33] Therefore, our proposed RSB approach will be able to discriminate distinct species efficiently by an *rbcL* SNP barcode.

A limitation of RSB is that it can only discriminate species with known sequence data but cannot identify species without sequence background. The universality of SNP barcodes developed by RSB performs well to different species with known *rbcL* sequences. As far as Brassicaceae species are concerned, the decision tree algorithm–generated SNP barcodes of *rbcL* genes are species specific. However, it needs a reanalysis to generate new SNP barcodes when more species are added. This is a supermarket-like idea that our proposed method only aimed to provide the specific SNP tags (strip barcode) for different

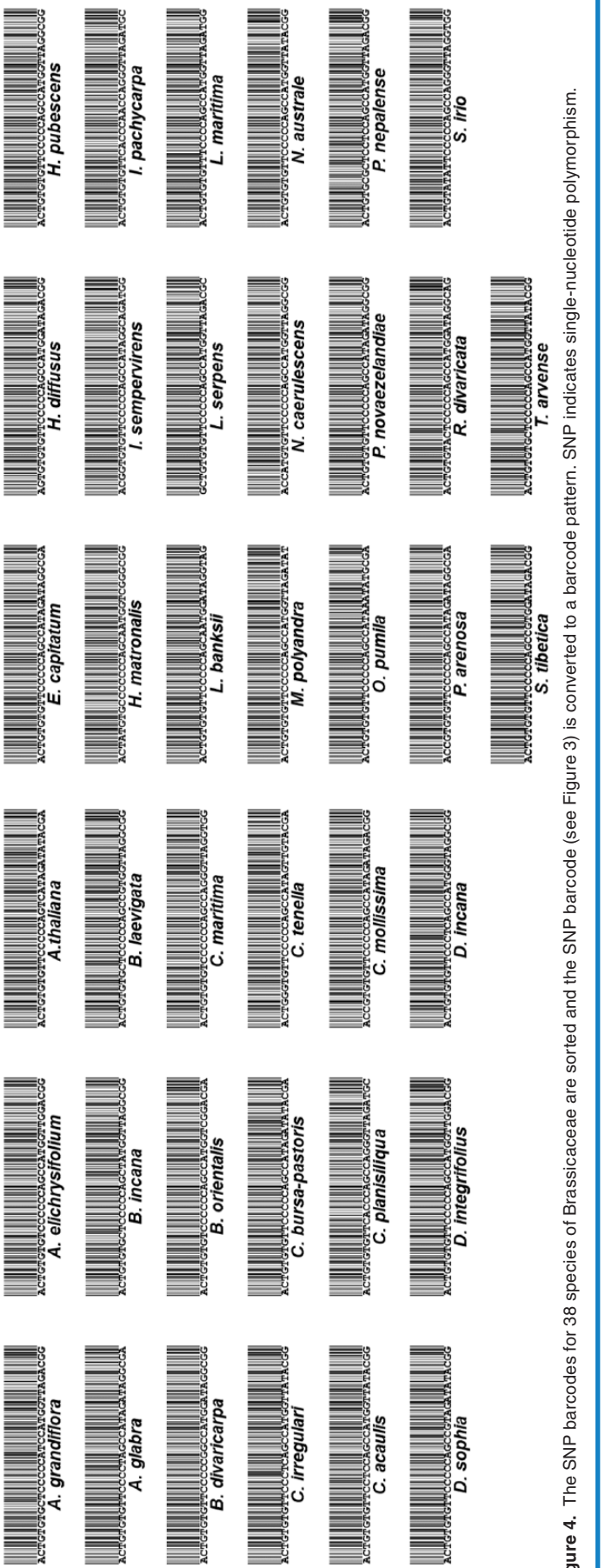

Figure 4. The SNP barcodes for 38 species of Brassicaceae are sorted and the SNP barcode (see Figure 3) is converted to a barcode pattern. SNP indicates single-nucleotide polymorphism.

species (goods) among the known species (all products in a supermarket). Once new species are added, more SNP numbers are required and the species-specific SNP barcodes need to be regenerated by computation. It was warranted to construct the database for certain interesting species groups in the future. The present contribution suggests that longer traditional barcode sequences could be narrowed down to smaller SNP barcodes for tagging species and reducing the overall nucleotide number for computation. However, the reference sequence used as the standard for trimming sequence ends rather than the whole sequence data, limiting its availability of evolutionary significance using our proposed method. It warrants further investigation for developing algorithm-based SNP patterns for long whole sequence data in evolutionary studies.

In general, the classification model must be inferred from a larger set of sequences. Many machine learning–based approaches were developed for DNA barcoding classification, such as DNA-Bar,[34] BLOG,[35] VIP Barcoding,[36] LAF,[29] and logic mining technique.[37] For example, DNA-Bar provides selecting DNA distinguishers (specific nucleotides) from genomic sequences to identify different microorganisms.[34] It finds a near-minimum number of distinguishers and gives specific pattern for each microorganism by a greedy distinguisher selection strategy. Although the DNA distinguishers represent a list of 10 nucleotide patterns, these nucleotides were extracted from genomic sequences rather than from the same gene region as shown in our study with the *rbcL* gene. Because the combinational nucleotides (DNA barcodes) identified in our study is derived from the same gene, these DNA barcodes belong to the SNP barcodes. BLOG 2.0 provides the character-based species classification for DNA barcoding by data mining.[35] VIP Barcoding provides a composition vector–based method for fast species classification.[36] LAF combined alignment-free techniques and rule-based classification algorithms for bacterial genome classification.[29] In general, most of these classification model studies solve the problem of allocating an unidentified species to a known species based on its DNA barcode.[26]

Although the *rbcL* gene represents a suitable nucleotide sequence for distinguishing different plant species, the longer traditional barcode sequences should be narrowed down to small number of SNP barcode for the ease of tagging species and reducing the amount of information for computation. Only one set of SNP barcodes is sufficient for the reliable tagging of a species. Therefore, the model establishment is not necessary in our study. For this purpose, we use the decision tree idea to identify the specific pattern of SNP barcodes for tagging a species. In detail, we use the concept of decision tree that creates decision rules (IF-THEN-ELSE) to find the best DNA sequence to dichotomize in each node. Our proposed method uses nucleotides to distinguish all species through tree traversal. The result of node traversal is summarized to generate SNP barcode and finally produces the strip barcode for tagging a species. It is different from a general decision tree in machine learning that needs to separate existing data into training sets and testing sets.[26,37,38] Therefore, it is not necessary in our study to use a training set to build the classification model and then use a testing set to verify the performance.

An important perspective of this work is further development of software that could automate our proposed algorithms for tagging species. A brief version of the "SNP barcodes for species tagging" software is available at the Web site http://203.64.88.159/barcoding/. It provides 6 items for demonstration, such as introduction, download, usage (user manual), data set (3 family), demo (YouTube video for operating 3 sample data sets), and author information. The developed species-specific SNP barcodes are readable by the barcode reader (scanner) or smartphone apps (such as QuickMark QR Code); you can download iOS (https://itunes.apple.com/us/app/qr-code-reader-quickmark-barcode/id384883554?mt=8) or Android (https://play.google.com/store/apps/details?id=tw.com.quickmark) using cell phone. In this way, the species-specific SNP barcodes are readable and distinguishable from each other. In the future, more species can be included to provide unique SNP barcodes for Brassicaceae plants and other organisms in principle.

Conclusions

The full length of the *rbcL* gene is commonly used for taxonomic identification of spermatophyte plants. Because most of the nucleotides do not vary and are not informative for species identification and authentication, we developed an RSB method, which combines the sequence alignment with a decision tree algorithm to generate the shortest barcode with 31 *rbcL* SNPs for discriminating of 38 Brassicaceae species as an example. As the *rbcL* gene barcode sequence is shortened from the full length (~1086 to 1410 bp) to 31 SNPs, the computational time is dramatically reduced. Therefore, the RSB-aided *rbcL* SNP barcode proposed here provides an effective way of species identification for plants.

Acknowledgements

The authors thank Dr Hans-Uwe Dahms for the English editing.

Author Contributions

L-YC and H-WC conceived and designed the research and wrote the paper. C-HY instructed K-CW for algorithm processing. K-CW also contributed to sequence retrieval. C-HY and H-WC revised the paper. All authors read and approved the final manuscript.

Supplementary Materials

Supplementary file S1: Example of strategy of decision tree construction in *rbcL* aligned sequences.Supplementary Material
Supplementary file S2: Phylogenetic trees of 38 Brassicaceae species.

REFERENCES

1. Ratnasingham S, Hebert PD. Bold: the barcode of life data system. *Mol Ecol Notes*. 2007;7:355–364. http://www.barcodinglife.org.
2. Hebert PD, Cywinska A, Ball SL, DeWaard JR. Biological identifications through DNA barcodes. *Proc Royal Soc B Biol Sci*. 2003;270:313–321.
3. Bafeel SO, Arif IA, Bakir MA, Al Homaidan AA, Al Farhan AH, Khan HA. DNA barcoding of arid wild plants using *rbcL* gene sequences. *Genet Mol Res*. 2012;11:1934–1941.
4. Bafeel SO, Alaklabi AA, Arif IA, et al. Ribulose-1, 5-biphosphate carboxylase (*rbcL*) gene sequence and random amplification of polymorphic DNA (RAPD) profile of regionally endangered tree species *Coptosperma graveolens* subsp. *arabicum* (*S. moore*) Degreef. *Plant Omics J*. 2012;5:285–290.
5. Barabaschi D, Tondelli A, Desiderio F, et al. Next generation breeding. *Plant Sci*. 2016;242:3–13.
6. Garcia-Robledo C, Erickson DL, Staines CL, Erwin TL, Kress WJ. Tropical plant-herbivore networks: reconstructing species interactions using DNA barcodes. *PLoS ONE*. 2013;8:e52967.
7. CBOL Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci U S A*. 2009;106:12794–12797.
8. Pei N, Erickson DL, Chen B, et al. Closely-related taxa influence woody species discrimination via DNA barcoding: evidence from global forest dynamics plots. *Sci Rep*. 2015;5:15127.
9. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. SNP markers and their impact on plant breeding. *Int J Plant Genomics*. 2012;2012:728398.
10. Appleby N, Edwards D, Batley J. New technologies for ultra-high throughput genotyping in plants. *Methods Mol Biol*. 2009;513:19–39.
11. Arif IA, Bakir MA, Khan HA, et al. A brief review of molecular techniques to assess plant diversity. *Int J Mol Sci*. 2010;11:2079–2096.
12. Chuang LY, Yang CS, Ho CH, Yang CH. Tag SNP selection using particle swarm optimization. *Biotechnol Prog*. 2010;26:580–588.
13. Shikha M, Kanika A, Rao AR, Mallikarjuna MG, Gupta HS, Nepolean T. Genomic selection for drought tolerance using genome-wide SNPs in maize. *Front Plant Sci*. 2017;8:550.
14. Tsai LC, Lee JC, Liao SP, Weng LH, Linacre A, Hsieh HM. Establishing the mitochondrial DNA D-loop structure of *Columba livia*. *Electrophoresis*. 2009;30:3058–3062.
15. Chen JB, Chuang LY, Lin YD, et al. Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. *Mitochondrial DNA*. 2014;25:231–237.
16. Yang CH, Chuang LY, Cheng YH, et al. Single nucleotide polymorphism barcoding to evaluate oral cancer risk using odds ratio-based genetic algorithms. *Kaohsiung J Med Sci*. 2012;28:362–368.
17. Wang X, Liu Y, Wang L, Han J, Chen S. A nucleotide signature for the identification of *Angelicae sinensis* radix (Danggui) and its products. *Sci Rep*. 2016;6:34940.
18. Baniecki ML, Faust AL, Schaffner SF, et al. Development of a single nucleotide polymorphism barcode to genotype *Plasmodium vivax* infections. *PLoS Negl Trop Dis*. 2015;9:e0003539.
19. Coll F, McNerney R, Guerra-Assuncao JA, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun*. 2014;5:4812.
20. Liu Y, Wang X, Wang L, Chen X, Pang X, Han J. A nucleotide signature for the identification of American ginseng and its products. *Front Plant Sci*. 2016;7:319.
21. Franzke A, Koch MA, Mummenhoff K. Turnip time travels: age estimates in Brassicaceae. *Trends Plant Sci*. 2016;21:554–561.
22. Schmidt R, Acarkan A, Boivin K. Comparative structural genomics in the Brassicaceae family. *Plant Physiol Biochem*. 2001;39:253–262.
23. Younsi R, MacLean D. Using 2k + 2 bubble searches to find single nucleotide polymorphisms in k-mer graphs. *Bioinformatics*. 2015;31:642–646.
24. Zhang Q_, Abel H, Wells A, et al. Selection of models for the analysis of risk-factor trees: leveraging biological knowledge to mine large sets of risk factors with application to microbiome data. *Bioinformatics*. 2015;31:1607–1613.
25. Loh WY. Classification and regression trees. *Wiley Interdis Rev*. 2011;1:14–23.
26. Weitschek E, Fiscon G, Felici G. Supervised DNA barcodes species classification: analysis, comparisons and results. *BioData Min*. 2014;7:4.
27. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2011;39:D32–D37.
28. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–1874.
29. Weitschek E, Cunial F, Felici G. LAF: logic alignment free and its application to bacterial genomes classification. *BioData Min*. 2015;8:39.
30. Polychronopoulos D, Weitschek E, Dimitrieva S, Bucher P, Felici G, Almirantis Y. Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers. *Genomics*. 2014;104:79–86.
31. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27:130–135.
32. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. 2008;26:1011–1013.
33. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–425.
34. DasGupta B, Konwar KM, Mandoiu II, Shvartsman AA. DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*. 2005;21:3424–3426.
35. Weitschek E, Van Velzen R, Felici G, Bertolazzi P. 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*. 2013;13:1043–1046.
36. Fan L, Hui JH, Yu ZG, Chu KH. VIP barcoding: composition vector-based software for rapid species identification based on DNA barcoding. *Mol Ecol Resour*. 2014;14:871–881.
37. Bertolazzi P, Felici G, Weitschek E. Learning to classify species with barcodes. *BMC Bioinformatics*. 2009;10:S7.
38. van Velzen R, Weitschek E, Felici G, Bakker FT. DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE*. 2012;7:e30490.