In Silico Study on Molecular Sequences for Identification of *Paphiopedilum* Species

Huyen-Trang Vu[1,2], Phuong Huynh[1], Hoang-Dung Tran[1] and Ly Le[2]

[1]Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.
[2]International University – Vietnam National University, Ho Chi Minh City, Vietnam.

Evolutionary Bioinformatics
Volume 14: 1–9

DOI: 10.1177/1176934318774542

ABSTRACT: Our study searched all available sequences of *Paphiopedilum* from NCBI (National Center for Biotechnology Information) and tested for their species resolution capability in single as well as in combination forms. A total of 28 loci were applied for analyses in the study. From the nuclear genome, the highest resolution was of *LFY*, followed by *ACO, DEF*4, and *RAD*51. These 4 loci were found to be even better than the popular region ITS for *Paphiopedilum* identification. Among the chloroplast regions, the intergenic spacer *atpB-rbc*L gave the highest species resolution (76.7%), followed by *mat*K, *trn*L, *rpo*C2, and *ycf*1. The divergence of *CHS, XDH*, 18S, *Nad*1, *ccs*A, *rbc*L, and *ycf*2 was very low and should not be used as identifying markers for *Paphiopedilum*. In addition, 2-locus combinations could improve significantly the resolving capability for the genus, in which 14/36 data sets could be resolved completely (100%) with interspecies relationships. The indel information was also effective supporting data for molecular discrimination of species.

KEYWORDS: *Paphiopedilum*, molecular identification, barcoding, ITS, *mat*K

RECEIVED: October 27, 2017. **ACCEPTED:** April 9, 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 project was funded by the Asian Office of Aerospace Research and Development (AOARD) under grant FA2386-17-1-4032.

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 AUTHOR: Ly Le, International University – Vietnam National University, Ho Chi Minh City, Vietnam. Email: ly.le@hcmiu.edu.vn

Introduction

Paphiopedilum is one of the largest genus of the Slipper orchid group Cypripedioideae.[1,2] Their flowers have the special slipper shapes and attractive colors which are the important characteristic to discriminate among individuals. The overexploitation and illegal trade, especially at the vegetative stage, make this plant come into endangered states.[2,3] Therefore, a demand for exact identification of these species at any stage of their life without the necessity of the fertile part is critical for conservation.

The single region 5′ end of cytochrome c oxidase 1 (*CO*1) from the mitochondrial genome has been proved to effectively identify animal individuals.[3,4] Universal sequences for identification of plants are also being searched from different genomes, mainly from nucleus and chloroplast. However, until now, the effective molecular identification of plants remains challenging because some loci have just been optimal for some particular groups of taxa. Thus, to respond to the demand of plant discrimination, separate sequences for identifying specific taxonomic groups have been done. There were a number of studies screening for potential molecular sequences to authenticate Slipper orchids. The 5 loci *rpo*B, *rpo*C1, *rbc*L, and *mat*K from the chloroplast genome and ITS from the nuclear genome have been used to test for species resolution of Indian *Paphiopedilum* species by Parveen et al[5] in 2012. The 2 plastid loci, *rpo*C2 and the intergenic spacer region between *atpF* and *atpH* genes were used to generate molecular identification system for some species of *Cypripedium* genus.[6] The phylogenetic tree of 25 species of subfamily Cypripedioideae in the research of Guo et al[7] has shown well-resolved monophyletic branches for all studied species based on sequence analyses of 6 maternally inherited chloroplast genes (*mat*K, *rbc*L, *rpo*C1, *rpo*C2, *ycf*1, and *ycf*2) and 2 low-copy nuclear genes (*LFY* and *ACO*). ITS was also used for identification and evaluation of variety relationship of 16 species and 2 varieties of genus *Paphiopedilum* in Vietnam.[8] Many other molecular regions have already been screened for discriminating a wide range of plants. However, the resolution results were not similar in different researches due to the differences in selected taxa, the differences in the number of studied species as well as the sequence length used, and the differences in methods.

In this study, we aim to evaluate all available submitted sequences relating to *Paphiopedilum* species from NCBI (National Center for Biotechnology Information) for species resolution of this genus, using a combination of measurements of tree-based method, indel fragments, variable sites, and even sequence length to search for potential loci which can be used in identification of individuals of the genus. This study may contribute to the use of proper molecular sequences as effective barcoding markers for identification of specific *Paphiopedilum* orchids, serving for conservation and diversity research of the plant.

Materials and Methods

Collected sequences of Paphiopedilum from GenBank

All sequences belonged to *Paphiopedilum* species from GenBank were selected and downloaded into separate loci. From that we obtained 2743 accession numbers of 37 loci

(Additional file 1). The sequence names were then shortened with only scientific species name and their accession number for easy analyses.

The number of sequences and the number of species between loci were different from each other. Among of them, 7 loci which have contained only 1 to 2 sequences, ie, *GLO, DEF, LOAE, IFR, DFR, psa*B, and *psb*A were not used in further tests. The 2 loci, psaA-ycf3 and ndhJ, which have had descriptions with the UNVERIFIED name were also removed. The *trn*K sequences were determined in the same locus with other *mat*K sequences (as *mat*K gene locates inside the region of *trn*K gene) and then we integrated them into one data set of *mat*K, with a total of 266 accessions. Due to the high variation in length of *mat*K sequences, the *mat*K set was divided into 2 subsets (more discussion in the section "Results and Discussion"). The 5S region showed the extreme diversity of conspecific distances after primary aligning that could lead to the paraphyletic relationships and give low species resolution, so we excluded it from our testing. As to some recommendations that the ITS2 was a good barcode in comparison with the whole-length locus ITS,[9,10] in this study, we isolated the data set of ITS2 from ITS to separately examine this short region. Finally, 28 loci were applied for further study: ITS, ITS2, *LFY, ACO, DEF*4, *RAD*51, *CHS, XDH*, and 18S in the nuclear genome; *Nad*1 in the mitochondrial genome; and *acc*D, *ccs*A, *mat*K(1), *mat*K(2), *rbc*L, *rpo*B, *rpo*C1, *rpo*C2, *trn*L, *ycf*1, *ycf*2, the intergenic spacers *atp*B-*rbc*L, *atp*F-*atp*H, *atp*H-*atp*F, *trn*H-*psb*A, *trn*L-*trn*F, *trn*Q-*rps*16, and *trn*S-*trn*FM in the chloroplast genome. In a total number of 107 species, the number and the composition of species were different among different loci. The species coverages of each locus are shown in Table 1. The summary of selected loci of *Paphiopedilum* from NCBI is shown in Additional file 1.

Multiple sequence alignment

Sequences at each locus were aligned using SeaView software version 4.4.0.[11] Alignments were then manually optimized, especially with noncoding regions in chloroplast genome and genes in nuclear genome which are highly divergent and contain many indel fragments. We manually noted all these information as they are also helpful in species resolution. The parsimony and singleton calculations for each data set were performed by Mega 7 program.[12] In the process of alignment, sequences that were too short and too divergent were removed from data sets. The output of alignment analyses is shown in Table 1.

Evaluation of species resolution

Phylogenetic tree for each locus were generated by Neighbor-Joining method using Mega 7 software with Kimura-2-parameter (K2P) model.[12] Evaluation of species resolution was performed using tree-based method and a combination of different measurements. First, species with just only one accession were considered to be distinguishable if the sequence was unique from others. In the phylogenetic tree, this accession would be shown as monophyletic branch.[13] Second, when multiple accessions were collected per species, these all accessions would be grouped into one monophyletic branch in the phylogenetic tree. Third, in converse, if conspecific individuals were not grouped together but separated in paraphyletic branches, then the species was considered as identification failure. A further description of insertions, deletions, and repeats should be included as if there was any difference between them that could help to authenticate them from the others.[14,15] Finally, in case of undiscrimination by K2P distance, different species were grouped in the same branch in the phylogenetic tree. This means that sequences of these accessions were identical. These heterospecies sequences in the same branch would be also more observed with indel information.

In barcoding studies as well as phylogenetic researches, the identification of orthologs and paralogs give much influence on the exact species resolution. Although orthology is one of the most important criteria to evaluate the relationship between individuals, paralogous sequences which are the results of duplicate events and divergence are not correlated with speciation. The presence of paralogs in the same sample may lead to an overestimation of the number of unique species under identification.[16] In practice, this problem can be eliminated by identifying and avoiding paralog contamination before sequence analysis. First, if paralogs really exist in different size, the polymerase chain reaction products using universal primers will be expressed as ghost bands on gel in the electrophoresis step. Second, even the paralogs have the same length as the real gene, and we cannot recognize them under electrophoresis but the sequencing would be shown with ambiguities, double peaks, or noise. These are 2 simple ways to remove paralogs from our analysis. In this in silico study, all the accessions were selected from the available source of GenBank. In case of interspecies comparative, it was considered that the sequences were orthologs. In case of intraspecies relationship analysis, the accessions might be orthologs or paralogs, especially between different clones of the same samples. Instead of deleting the clone sequences, we decided to keep all of them and use as one of the factors affecting the species resolution. The locus with high variation between either intraspecies orthologs or homologous paralogs would result in low species resolution and could not be used as candidate molecular identification markers.

Results and Discussions

Multiple sequence alignment and divergence of the loci

The nuclear loci ITS, ITS2, *LFY, ACO, DEF*4, *RAD*51, *CHS* (except *XDH* and 18S regions) had the parsimony rates (10.8%-49.1%) and singleton rates (9.7%-20.9%) significantly higher than all loci in plastid genome (0.3%-12% in parsimony

Table 1. Comparison parameters of sequences in analysis data sets.

NO.	LOCUS	ALIGNMENT LENGTH, BP	PARSIMONY RATE, %	SINGLETON RATE, %	TOTAL INDELS	NO. OF SPECIES	SPECIES COVERAGE, %	NO. OF SELECTED ACCESSIONS	IDENTIFIED SPECIES			SPECIES RESOLUTION, %
									USING GENETIC DISTANCE	USING INDEL	TOTAL	
1	ITS	634	36.3	16.4	17	91	85.0	339	24	0	24	26.4
2	ITS2	175	49.1	17.7	7	91	85.0	339	19	0	19	20.9
3	LFY	1280	24.8	16.2	31	62	57.9	116	50	0	50	80.6
4	ACO	1333	24.4	16.2	33	71	66.4	133	51	1	52	73.2
5	DEF4	1253	19.9	11.0	59	70	65.4	113	45	0	45	64.3
6	RAD51	914	20.4	17.3	42	67	62.6	124	35	2	37	55.2
7	CHS	1327	10.8	20.9	11	12	11.2	13	10	2	12	100.0
8	XDH	836	5.3	4.2	0	18	16.8	20	18	0	18	100.0
9	18S	1665	0.3	2.0	0	4	3.7	4	4	0	4	100.0
10	Nad 1	1480	0.3	0.1	11	11	10.3	12	3	4	7	63.6
11	accD	715	5.2	1.4	5	77	72.0	107	18	0	18	23.4
12	ccsA	651	0.8	2.0	0	4	3.7	4	4	0	4	100.0
13	matK(1)	1140	6.1	2.6	1	40	37.4	70	21	0	21	52.5
14	matK(2)	598	12.0	1.7	1	87	81.3	253	18	0	18	20.7
15	rbcL	484	3.3	1.2	0	79	73.8	151	5	0	5	6.3
16	rpoB	305	0.3	0.0	0	11	10.3	41	0	1	1	9.1
17	rpoC1	2751	5.5	2.9	4	75	70.1	104	45	0	45	60.0
18	rpoC2	1743	5.7	2.0	15	79	73.8	158	30	2	32	40.5
19	ycf1	1235	6.1	2.2	9	79	73.8	158	30	0	30	38.0
20	ycf2	1512	0.5	1.1	6	7	6.5	7	7	0	7	100.0
21	trnL	620	4.7	5.2	22	77	72.0	84	26	12	38	49.4
22	atpB-rbcL	798	2.5	6.3	52	60	56.1	63	32	14	46	76.7
23	atpI-atpH	725	6.1	2.5	32	76	71.0	107	15	2	17	22.4
24	atpH-atpF	462	4.5	3.5	11	55	51.4	78	15	4	19	34.5
25	trnH-psbA	1030	2.6	2.5	20	20	18.7	20	9	7	16	80.0
26	trnL-trnF	403	3.5	4.0	12	77	72.0	84	10	3	13	16.9
27	trnQ-rps16	956	2.8	2.9	21	17	15.9	17	12	1	13	76.5
28	trnS-trnfM	884	3.6	1.8	16	75	70.1	117	16	5	21	28.0

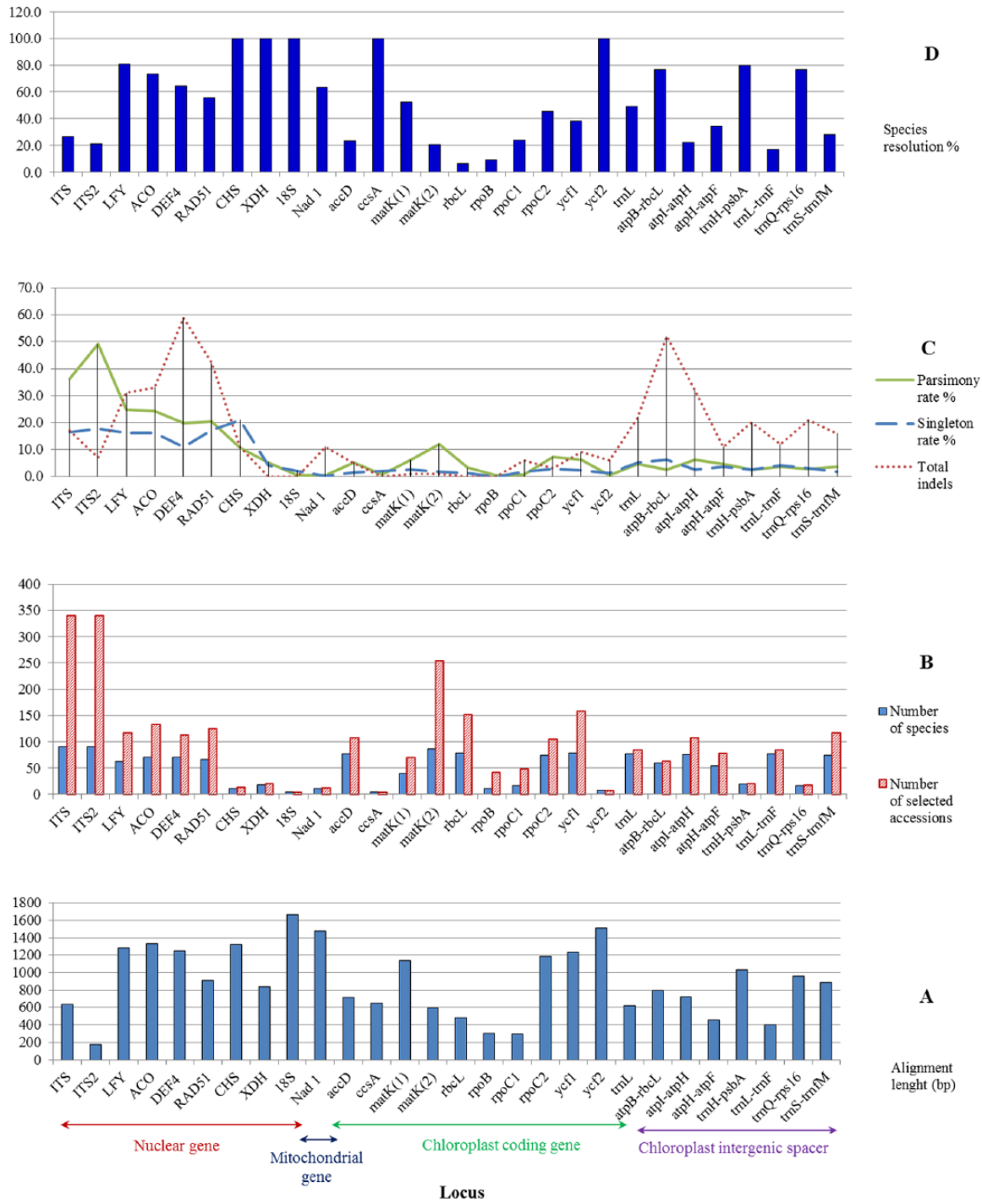

Figure 1. Species resolution of single loci of *Paphiopedilum* sequences and other analyzed information about parsimony rate, singleton rate, indel fragment, number of species, number of sequence, and alignment length per locus.

rate, 0%-6.5% in singleton rate; Table 1). The total indels in chloroplast protein-coding regions, in general, were very low in comparison with most other chloroplast regions and the nuclear genes (except *XDH* and 18S genes; Figure 1C), which explained the different alignment capabilities of these regions as mentioned below.

To have a good discrimination effect, first, the selected sequences should have a proper length that is not too long for simple amplification and not too short for containing enough divergence information. Second, these sequences should be easy to be amplified, sequenced, and aligned. Third, the divergence of the sequence should be high enough for distinguishment at the species level but not too variable at underspecies level.[3] Finally, the high species resolution is a critical criterion. In our study, as the sequences were selected from GenBank, we did not analyze the amplification and sequencing rates. However, the sequence length, the alignment capability, the divergence range, and the resolution rate were all taken into account. The *rpo*C2

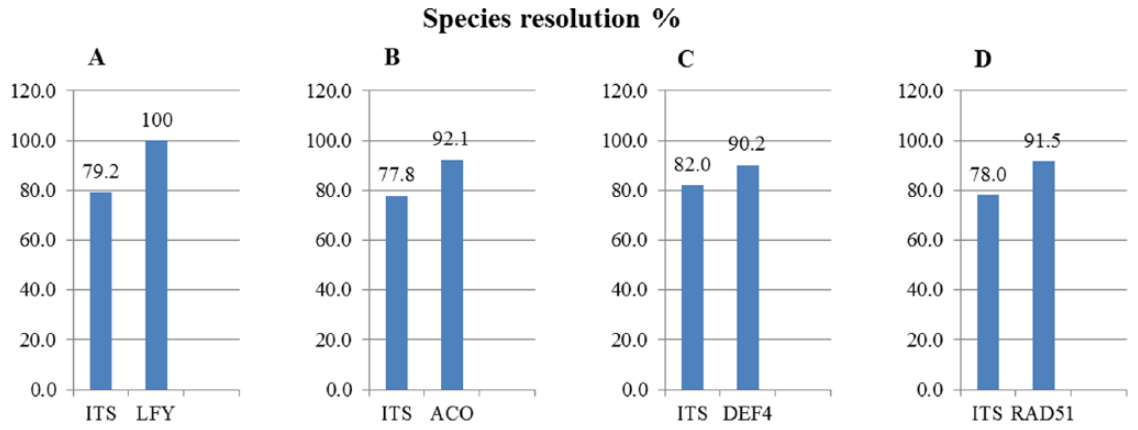

Figure 2. Comparison of resolution between ITS region and each of *LFY, ACO*, DEF4, *RAD*51 regions. The comparisons with the same species number showed that other nuclear genes were better than ITS.

and *ycf*1 regions had rather long sequences (2751 and 1743 bp [base pairs], respectively) which were not suitable to be molecular markers. Therefore, we had to cut the 2 ends of these loci to keep the most variable fragments with suitable length (1188 bp of *rpo*C2 and 1235 bp of *ycf*1) for our study.

The coding-gene regions in the chloroplast genome were easy to be aligned. The noncoding intergenic spacers in chloroplast and the nuclear loci took more time for alignment due to their great indels and polynucleotide repeats inside their sequences. However, they were all successfully aligned. Exceptionally, the *LFY* region contains quite long sequences (from 904 to 3404 bp) in which we could just align the half top fragment (about 1280 bp) for analysis. The other half end fragment which was too complicated to be aligned was removed.

The capability of species identification with single regions

Some of the tested regions achieved 100% species resolution, ie, *CHS, XDH*, 18S in the nuclear genome and *ccs*A, *ycf*2 in the chloroplast genome. However, these loci included extremely low number of species (12, 18, 4, 4, and 7, respectively; Table 1). In addition, their divergent levels were too low (Figure 1C). For these reasons, the results obtained from these loci were not confident enough to evaluate the species discrimination. The similar situations were also obtained from *Nad*1 region (with 11 species), *trn*H-*psb*A region (with 20 species), and *trn*Q-*rps*16 region (with 17 species; Table 1). Then, these 8 regions were removed from our further analyses.

Although the number species of *CHS, XDH*, 18S, *Nad*1, *ccs*A, *rbc*L, *ycf*2, *trn*H-*psb*A, and *trn*Q-*rps*16 loci were very low and could not used for analysis, we still included these regions into this step of the study because we wanted to examine their variability for future studies. The results showed that the divergence of *CHS, XDH*, and 18S of the nuclear genome; the *Nad*1 of the mitochondrial genome; and the *ccs*A, *rbc*L, and *ycf*2 of the chloroplast genome were very low and should not be used as identifying markers. However, *trn*H-*psb*A and *trn*Q-*rps*16 had a rather high total indel fragments and should be examined more with a larger number of studied sequences in future.

In the nuclear genome. The ITS2 locus has been highly noticed as an alternative barcode instead of the full-length ITS region due to ITS high divergence and short length for easy to amplify, to sequence, and to align.[17,18] In this study, indeed ITS2 with a really short sequence (175-282 bp) could achieve just a little lower resolution than ITS (634 bp; Table 1). However, because the identification capability of ITS was still higher than ITS2, and the length of ITS (about 634 bp) was still suitable for amplification and sequencing, and moreover we did not meet any problem in alignment of this region, we still favored the whole ITS rather than the ITS2 in obtaining better resolution.

Among the analyzed regions in the nucleus, ITS and ITS2 loci gave the lowest resolution rate (26.4% and 20.9%). The highest ability was of *LFY* region which could resolve 50/62 species (80.6%; Figure 1D), followed by *ACO, DEF*4, and *RAD*51 (73.2%, 64.3%, and 55.2%, respectively). However, on contrast, the informative parsimony rates of both ITS and ITS2 sequences were higher than the remaining nuclear genes in the study. The reason might be that the insertions and deletions of *LFY, ACO, DEF*4, and *RAD*51 regions were all much higher than of ITS regions (Figure 1C). Besides, the number of sequences per species of ITS regions was very high (Figure 1B) which related to the variable sequences within species. The difference in species number and sequence number between the loci limited the comparison of all the loci simultaneously. For these reasons, to more exactly compare the resolution, we performed further pair tests between ITS locus with each of *LFY, ACO, DEF*4, and *RAD*51 regions in the condition that 2 tested loci contained the same species and the same sequence number (Additional file 2). The results now revealed that all 4 regions *LFY, ACO, DEF*4, and *RAD*51 still showed better resolution than ITS (Figure 2). Although ITS has been widely used in many barcoding as well as phylogenetic studies,[18,19] in this research, we found that other nuclear genes were even better than ITS.

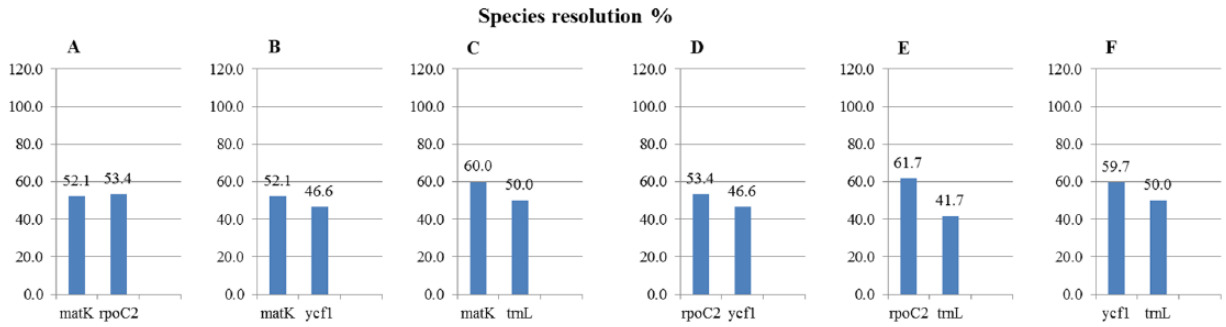

Figure 3. Comparison of resolution with the same species between pairs of regions *mat*K, *rpo*C2, *trn*L, and *ycf*1.

In the chloroplast genome. To the *mat*K region, the lengths between sequences were so different from 589 to 1275 bp. For this reason, after primary aligning, we decided to separate this data set into 2 smaller data sets: the *mat*K(1) including 140 sequences with 1140 bp in length and the *mat*K(2) including all selected 253 sequences from GenBank with 589 bp in length (Table 1). The species resolution of *mat*K(1) was 52.5% which was much higher than *mat*K(2) (20.9%; Figure 1D). The reasons might be either the difference of variation information caused by different length or the difference of species numbers between 2 subsets. Therefore, we performed a further test to compare the resolution of 2 matK sets in the condition of the same number of sequences and the same number of species as well. The results showed that the longer one which was *mat*K(1) gave the higher success rate in species identification (data not shown). *Mat*K sequences were easy to be aligned and the length of 1140 bp was suitable for amplification as well as for enhancing the species distinguishment. We recommend using the longer *mat*K locus in molecular identification of *Paphiopedilum* species.

Among the chloroplast regions, the intergenic spacer *atp*B-*rbc*L gave the highest species resolution (76.7%), followed by *mat*K(1), *trn*L, *rpo*C2, and *ycf*1 (52.5%, 49.4%, 45.9%, and 38%, respectively). The coding regions *rbc*L, *rpo*B, *acc*D, and *rpo*C1 were highly conservative along their sequences and resulted in low resolution percentages (6.3%, 9.1%, 20.8%, and 23.5%, respectively). Except *atp*B-*rbc*L, other noncoding regions had low discrimination rates (*atpH-atpF* 36.4%, *trn*S-*trn*FM 28%, *atp*I-*atp*H 22.4%, and *trn*L-F 16.9%; Table 1). We made a series of further tests of the same species number between *mat*K and other coding loci to more compare their resolutions (Additional file 2). To keep the most number of the same species for testing, *mat*K was analyzed with the length 716 bp. *Mat*K has been popularly used in identifying studies of Orchidaceae,[5,20,21] in our tests, although the length of *mat*K was just 716 bp, this region was really a potential chloroplast marker which had the better resolution than both *trn*L and *ycf*1 regions, and just a little lower than *rpo*C2 due to the decrease in ITS length (Figure 3). *Ycf*1 has been recently considered as "more variable than *mat*K"[22] and highly evaluated as the most variable and parsimony-informative among the 5 chloroplast genes (*mat*K, *rbc*L, *rpo*C1, *rpo*C2, and *ycf*2) on plants.[7,23] However, in our study on *Paphiopedilum*, species distinguishment of this region was lower than *mat*K and *rpo*C2. Besides *mat*K, the locus *rpo*C2 could be a potential marker for identification of *Paphiopedilum*.

In the study of Parveen et al,[5] *mat*K received 100% and ITS got 50% species resolution among the 3 other loci *rpo*B, *rpo*C1, and *rbc*L. However, because the study just resolved only 8 species of *Paphiopedilum*, the result seemed to need more analysis. When Guo et al[18] focused on 9 regions including 8 from the chloroplast and one is the ITS, in this study, we looked inside a larger range of molecular markers of about 28 loci for identification capability *Paphiopedilum* species. The results supported potential region *mat*K and developed some new potential identification markers as *LFY, ACO* which were just used for the aim of phylogenetic relationship analysis in Guo et al.[7,24]

The use of indel information in species identification

In the chloroplast genome, the noncoding regions seemed to have more complexity than the coding regions based on high number of insertions and deletions. These characteristics contributed to the more difficulty in sequence alignment. However, because we could pass all the alignments in previous step, indel fragments on contrast became useful information for discrimination at species level. Specifically, in our study, *atpB-rbc*L locus could identify a total of 46 distinguished species, in which 13 over 46 species have been obviously separated using indel information (Table 1). Besides, *trn*H-*psb*A marker could separate 16/20 species and one-third of these 16 species were discriminated using insertions and deletions. In addition, *trn*L locus could identify 12 out of a total of 26 distinguished species using indel fragments. Moreover, insertions and deletions were also used in such other molecular regions (*ACO, RAD*51, *CHS, Nad*1, *rpo*B, *atp*I-*atp*H, *atp*H-*atp*F, *trn*H-*psb*A, *trn*L-F, *trn*Q-*rps*16, *trn*S-*trn*FM) of our study (Table 1). Although the indel information was not used by bioinformatics tools and by some researchers, it has been proved as effective supporting data for species authentication in a few previous studies.[6,7,14] We totally agree with the use of indel information as the fifth character

Table 2. Comparison of species resolution of single-locus sequences and 2-locus combinations.

NO.	SINGLE-LOCUS SEQUENCE	TOTAL SPECIES NUMBER	IDENTIFIED SPECIES	RESOLUTION, %
1	LFY	62	50	80.6
2	ACO	71	52	73.2
3	DEF4	70	45	64.3
4	RAD51	67	37	55.2
5	matK	40	21	52.5
6	rpoC2	75	34	45.3
7	ycf1	79	30	38.0
8	trnL	77	38.0	49.4
9	atpB-rbcL	60	46	76.7
2-LOCUS COMBINATION				
1	LFY_ACO	60	60	100.0
2	LFY_DEF4	59	59	100.0
3	LFY_Rad51	59	57	96.6
4	LFY_matK	29	29	100.0
5	LFY_ycf1	62	59	95.2
6	LFY_rpoC2	62	60	96.8
7	LFY_trnL	49	49	100.0
8	LFY_atpB-rbcL	43	43	100.0
9	ACO_DEF4	67	65	97.0
10	ACO_Rad51	66	63	95.5
11	ACO_matK	33	33	100.0
12	ACO_ycf1	70	67	95.7
13	ACO_rpoC2	70	67	95.7
14	ACO_trnL	56	56	100.0
15	ACO_atpB-rbcL	45	45	100.0
16	DEF4_Rad51	65	65	100.0
17	DEF4_matK	32	32	100.0
18	DEF4_ycf1	70	65	92.9
19	DEF4_rpoC2	70	62	88.6
20	DEF4_trnL	56	54	96.4
21	DEF4_atpB-rbcL	45	45	100.0
22	RAD51_matK	31	31	100.0
23	RAD51_ycf1	67	58	86.6
24	RAD51_rpoC2	67	52	77.6
25	RAD51_trnL	53	51	96.2
26	RAD51_atpB-rbcL	45	45	100.0
27	matK_ycf1	35	29	82.9

(Continued)

Table 2. (Continued)

2-LOCUS COMBINATION				
28	matK_rpoC2	34	29	85.3
29	matK-trnL	32	22	68.8
30	matK_atpB-rbcL	26	26	100.0
31	ycf1_rpoC2	75	52	69.3
32	ycf1_trnL	63	49	77.8
33	ycf1_atpB-rbcL	50	42	84.0
34	rpoC2_trnL	60	49	81.7
35	rpoC2_atpB-rbcL	48	38	79.2
36	trnL_atpB-rbcL	60	45	75.0

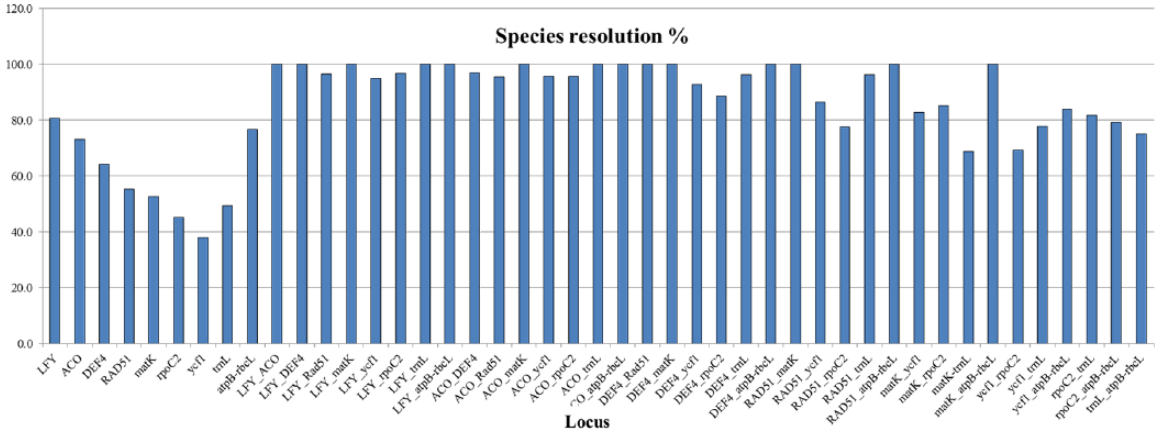

Figure 4. Species resolution of single loci and 2-locus combinations. Percent species resolutions showed more potential in a combination of sequences for identification study.

state in molecular identification of species, besides the 4 traditional characters—A, T, G, and C.

Capability of species identification with multiple region combinations

As no single locus could discriminate all species of *Paphiopedilum* in the study, the 9 recommended potential loci derived from single-locus tests, ie, *LFY, ACO*, DEF4, *RAD*51 from the nuclear genome and *mat*K, *rpo*C2, *trn*L, *ycf*1, and *atp*B-*rbc*L from the chloroplast genome were chosen for further testing with the combination of 2 loci as a unique sequence for identification (Table 2). For species with more than one congeneric sequence, we chose only one representative sequence to calculate the interdistance between species.

The resolutions of combination sequences were all higher than that of the single loci. In 2-locus combination, 14/36 data sets could be resolved completely (100%) with interspecies relationships (Figure 4). *LFY, mat*K, and *atp*B-*rbc*L gave the best results in which 5/8 combinations obtained 100% resolution, followed by *ACO*, DEF4, *RAD*51, and *trn*L (4/8, 4/8, 3/8, and 2/8, respectively). The results revealed high potential for discrimination ability of these combination barcodes.

Conclusions

Our study revealed that the 4 loci, *LFY, ACO, DEF*4, and *RAD*51, in the nuclear genome were better than ITS in identification of *Paphiopedilum* species, and the intergenic spacer *atp*B-*rbc*L had the highest resolution from the chloroplast genome. Along with *mat*K, *atp*B-*rbc*L, and *rpo*C2 in the chloroplast genome had high variability and should be considered as high-potential loci of species resolution for *Paphiopedilum* genus. In addition, the combination of potential loci could improve significantly the resolving capability. Besides, the indel information could be used as effective supporting data for molecular discrimination of species.

Because of the limitation of available sequence information for *Paphiopedilum* and the differences of sequence number among different loci, our evaluation still limited when comparing directly among loci. Therefore, we suggest a further study on both sequencing and in silico analysis for more potential loci from other locations of this plant genome. Furthermore,

continuous efforts based on these potential loci should be done on a higher number of samples to enhance identification ability for many useful applications in conservation and biodiversity foundation for this valuable orchid *Paphiopedilum*.

Acknowledgements

The computer resources were provided by Computational Biology Center of International University, Vietnam National University.

Author Contributions

H-TV analyzed and interpreted the results and was a major contributor in writing the manuscript. PH performed the bioinformatics calculation. H-DT gave advice on the manuscript. LL gave advice and gave final approval of the version to be published. All authors read and approved the final manuscript.

Availability of Data and Materials

All the sequences used in the study were downloaded from GenBank. Details of accession numbers are given in Additional file 3.

REFERENCES

1. Averyanov L, Cribb P, Phan KL, Nguyen TH. *Slipper Orchids of Vietnam With an Introduction to the Flora of Vietnam*. 1st ed. Portland, OR: Timber Press; 2003.
2. Chase MW, Cowan RS, Hollingsworth PM, et al. A proposal for a standardised protocol to barcode all land plants. *Taxon*. 2007;56:295–299.
3. Hebert PD, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. *Proc Biol Sci*. 2003;270:313–321.
4. Shneer VS. DNA barcoding is a new approach in comparative genomics of plants. *Genetika*. 2009;45:1436–1448.
5. Parveen I, Singh HK, Raghuvanshi S, Pradhan UC, Babbar SB. DNA barcoding of endangered Indian Paphiopedilum species. *Mol Ecol Resour*. 2012;12:82–90.
6. Kim JS, Kim HT, Son S-W, Kim J-H. Molecular identification of endangered Korean lady's slipper orchids (Cypripedium, Orchidaceae) and related taxa. *Botany*. 2015;93:603–610.
7. Guo YY, Luo YB, Liu ZJ, Wang XQ. Evolution and biogeography of the slipper orchids: Eocene vicariance of the conduplicate genera in the Old and New World Tropics. *PLoS ONE*. 2012;7:e38788.
8. Khuất HT, Trần ĐK, Lê HH, Trần DD, Khoa T. Molecular phylogeny of the endangered Vietnamese *Paphiopedilum* species based on the internal transcribed spacer of the nuclear ribosomal DNA. *Adv Stud Biol*. 2013;5:337–346.
9. Han J, Zhu Y, Chen X, et al. The short ITS2 sequence serves as an efficient taxonomic sequence tag in comparison with the full-length ITS. *Biomed Res Int*. 2013;2013:741476.
10. Yao H, Song J, Liu C, et al. Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE*. 2010;5:e13102.
11. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 2010;27:221–224.
12. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–1874.
13. Hollingsworth ML, Andra Clark A, Forrest LL, et al. Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol Ecol Resour*. 2009;9:439–457.
14. Shaw J, Lickey EB, Schilling EE, Small RL. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am J Bot*. 2007;94:275–288.
15. Wu C-T, Hsieh C-C, Lin W-C, et al. Internal transcribed spacer sequence-based identification and phylogenic relationship of I-Tiao-Gung originating from Flemingia and Glycine (Leguminosae) in Taiwan. *J Food Drug Anal*. 2013;21:356–362.
16. CBOL Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci U S A*. 2009;106:12794–12797.
17. Feng S, Jiang Y, Wang S, et al. Molecular identification of *Dendrobium* species (Orchidaceae) based on the DNA barcode ITS2 region and its application for phylogenetic study. *Int J Mol Sci*. 2015;16:21975–21988.
18. Guo YY, Huang LQ, Liu ZJ, Wang XQ. Promise and challenge of DNA barcoding in Venus slipper (*Paphiopedilum*). *PLoS ONE*. 2016;11:e0146880.
19. Singh HK, Parveen I, Raghuvanshi S, Babbar SB. The loci recommended as universal barcodes for plants on the basis of floristic studies may not work with congeneric species as exemplified by DNA barcoding of *Dendrobium* species. *BMC Res Notes*. 2012;5:42.
20. Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M. Identification of medicinal ***Dendrobium*** species by phylogenetic analyses using matK and rbcL sequences. *J Nat Med*. 2010;64:133–138.
21. Xu S, Li D, Li J, et al. Evaluation of the DNA barcodes in *Dendrobium* (Orchidaceae) from Mainland Asia. *PLoS ONE*. 2015;10:e0115168.
22. Dong W, Xu C, Li C, et al. ycf1, the most promising plastid DNA barcode of land plants. *Sci Rep*. 2015;5:8348.
23. Neubig KM, Whitten WM, Carlsward BS, et al. Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matK. *Plant Syst Evol*. 2009;277:75–84.
24. Guo YY, Luo YB, Liu ZJ, Wang XQ. Reticulate evolution and sea-level fluctuations together drove species diversification of slipper orchids (*Paphiopedilum*) in South-East Asia. *Mol Ecol*. 2015;24:2838–2855.