VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY

**INTERNATIONAL UNIVERSITY**

PHD

VŨ THỊ HUYỀN TRANG

# MOLECULAR IDENTIFICATION OF VIETNAM *PAPHIOPEDILUM* ORCHIDS USING DNA SEQUENCES

Doctor of Philosophy in Biotechnology
Code: 9420201

THESIS SUMMARY

HO CHI MINH CITY – 2020

# ABSTRACT

*Paphiopedilum* is a valuable genus in Vietnam. The dissertation constructed DNA sequences for genetic characterization and identification of *Paphiopedilum* species in Vietnam and to determine certain markers for detecting correct orchid origin and genetic diversity. The sequence analyses mainly based on the tree-based method. Species are identified when all sequences of the same species are clustered in the same branch on phylogenetic trees. Nucleotide polymorphism characteristics including variation, parsimony, singleton, mono-indel, insertion, deletion parameters were also recorded for sequence analyses.

First, for screening potential sequences, 5 loci ITS, *mat*K, *rpo*B, *rpo*C1 and *trn*H-*psb*A were sequenced and analyzed on 8 *Paphiopedilum* species. Simultaneously, 37 loci of *Paphiopedilum* from GenBank were also examined for identification capability. Besides, proper primers for those 5 potential loci, i.e. *ACO*, *LFY*, *atp*B-*rbc*L, *trn*L and *mat*K were developed and tested on fresh samples. Finally, the ITS in the nuclear genome and the two loci *mat*K, *trn*L in the chloroplast genome were selected for large-scale applications.

Second, the large-scale identification of Vietnam *Paphiopedilum* population was conducted on a total of 95 samples belonging to 22 species including 20 original species and 2 natural hybrids using 3 short sequences ITS, *mat*K and *trn*L. The result was that 17 out of 22 species of Vietnam *Paphiopedilum* species were well-identified. Furthermore, the combination of preliminary vegetative classification with later molecular recognition step resulted in the successful identification of up to 20 *Paphiopedilum* species in the study.

Third, for further comparison of species identification capability between whole chloroplast genome and short sequences, the complete chloroplast genome of the species *Paphiopedilum delenatii* was sequenced, assembled, and analyzed with that of other Orchidaceae species. The results showed that the complete genome could also be used as potential identification sequence, especially for closely related species.

All data from the study was upload to GenBank library for extensive research. The endemic species *Paphiopedilum x dalatense* of Vietnam was first reported. The results were also given back to scientific Institutes which contribute to the conservation and control of the illegal trade of *Paphiopedilum* species in Vietnam.

# CHAPTER 1. GENERAL INTRODUCTION

## 1.1 Problem statement

*Paphiopedilum* is a distinct genus of the Orchidaceae family which has a different flower structure with deep pocket-shaped lip that looks like a colorful lovely slipper. Vietnam is the country which has a high diversity of about 20 original species with different varieties and many other natural hybrids of *Paphiopedilum* orchids. Unfortunately, most Vietnam *Paphiopedilum* taxa are threatened with decrease and extinction due to uncontrolled sampling and deforestation.

The conservation of populations in nature is overly complicated. In addition to establishing protected areas and setting out regulations to prohibit illegal trade, customs and quarantine inspectors should have a basic understanding of plants, to be able to differentiate between rare and common species. Furthermore, most of the trading samples are at the vegetative developmental stage, with high morphological similarities among species, leading to misidentification, causing difficulty in illegal trade control and conservation work. Hence it is necessary to develop more effective identification methods for *Paphiopedilum* species in Vietnam.

There are some different goals in identification researches. Some researchers want to distinguish taxa among a known group of samples. Other researchers want to find out whether the sample is a specific taxon that contain a gene that has been well-known before. And some studies aim to determine the origin of an unknown sample or even want report a new taxon in science. With a low cost, PCR-based techniques can effectively solve the two first targets, using amplification reactions and specific characters of electrophoresis bands. However, an unknown taxon cannot be determined using these PCR-band techniques.

Since each site of the DNA sequence is considered as a character in bioinformatic analysis, the sequence-based method gives more variable information for identification and can overcome the problems of morphological and PCR-based methods. This approach allows the read of every nucleotide among the samples. Hence short DNA fragments represented for the organism can be used as an identifying sequence like the human fingerprint, with just a small amount of sample, making it efficient for evaluating the genetic relationship of a new unknown sample based on the available sequence library and so allowing to consult the similar origins of the unknown species.

## 1.2 Research objectives

In this study, we aimed to construct DNA sequences for genetic characterization and identification of *Paphiopedilum* species in Vietnam and to determine certain markers for detecting correct orchid germplasm and genetic diversity.

## 1.3 Significance of the study

Collection of genetic information, for looking up origins of a wide range of organisms linked all over the world, is an advanced idea and essential for the protection of species, phylogenetic inference, management, and development of genetic diversity, especially at biodiversity hotspots.

Understanding of genetic features of *Paphiopedilum* significantly contributes to the genetic conservation and diversity management of these valuable species. Data of Vietnam *Paphiopedilum* species was partial of the global GenBank library for extensive research. Our results were also given back to scientific Institutes which will contribute to the control of the illegal trade, conservation task and development of *Paphiopedilum* species in Vietnam.

### 1.4 Methodology

The sequence analyses mainly based on tree-based method. Species are identified when all sequences of the same species are clustered in the same branch on phylogenetic trees. Nucleotide polymorphism characteristics including variation, parsimony, singleton, mono-indel, insertion, deletion parameters were also recorded for sequence analyses. Details of methodology were presented in each result chapter.

### 1.5 Thesis organization

The dissertation was structured into 5 chapters. Chapter 1 briefly provided an introduction of the thesis. Chapter 2 presented theoretical basis and research situation related to the topic from which to propose thesis objectives and research methods. Chapter 3 described in details the materials used in the study and problem-solving methodology. Chapter 4 showed and discussed the results of different research. Finally, Chapter 5 provided conclusions about main results, contributions, and implications of the dissertation.

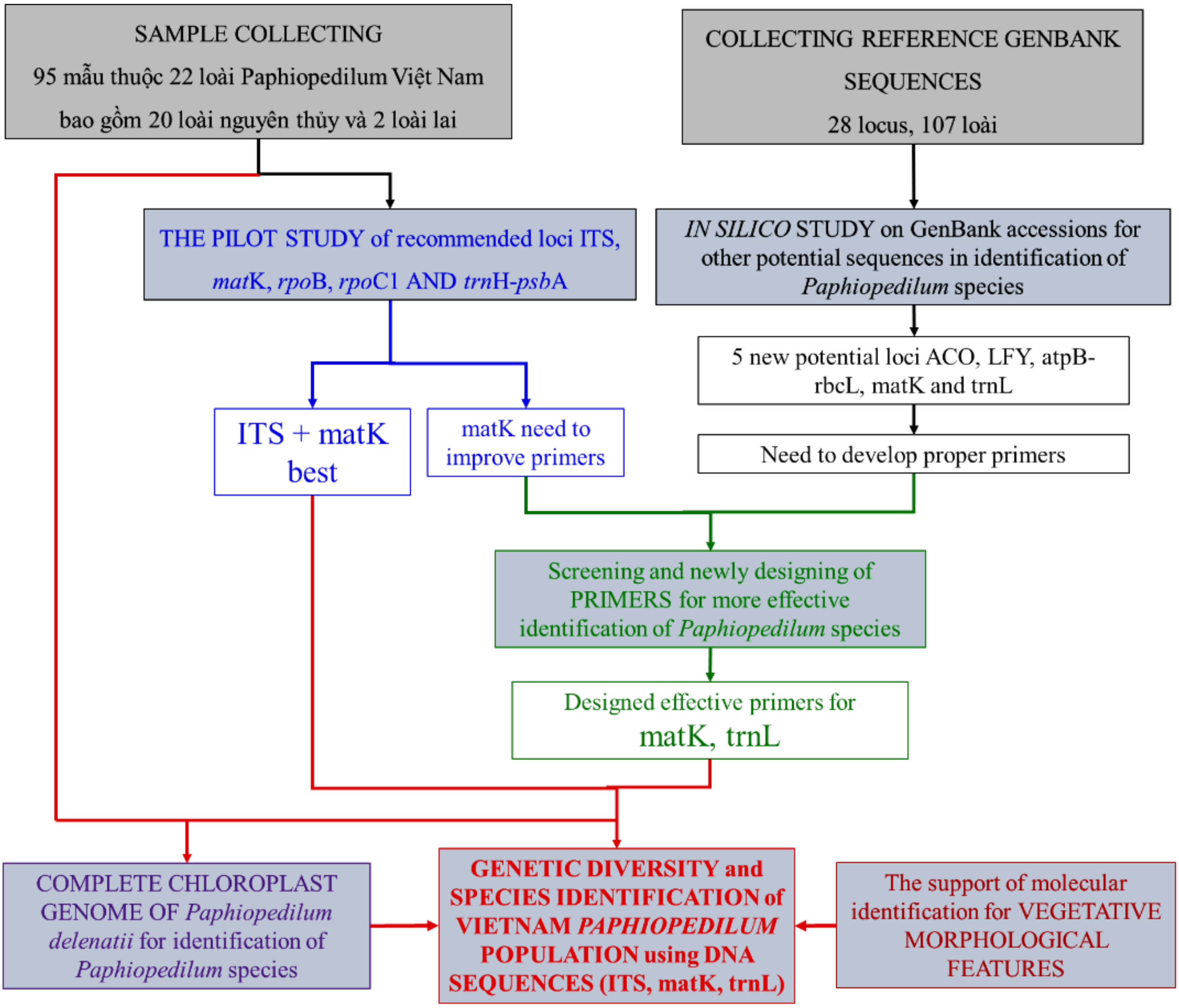
SAMPLE COLLECTING
95 mẫu thuộc 22 loài Paphiopedilum Việt Nam
bao gồm 20 loài nguyên thủy và 2 loài lai
COLLECTING REFERENCE GENBANK SEQUENCES
28 locus, 107 loài
THE PILOT STUDY of recommended loci ITS, *mat*K, *rpo*B, *rpo*C1 AND *trn*H-*psb*A
*IN SILICO* STUDY on GenBank accessions for other potential sequences in identification of *Paphiopedilum* species
5 new potential loci ACO, LFY, atpB-rbcL, matK and trnL
ITS + matK best
matK need to improve primers
Need to develop proper primers
Screening and newly designing of PRIMERS for more effective identification of *Paphiopedilum* species
Designed effective primers for matK, trnL
COMPLETE CHLOROPLAST GENOME OF *Paphiopedilum delenatii* for identification of *Paphiopedilum* species
GENETIC DIVERSITY and SPECIES IDENTIFICATION of VIETNAM *PAPHIOPEDILUM* POPULATION using DNA SEQUENCES (ITS, matK, trnL)
The support of molecular identification for VEGETATIVE MORPHOLOGICAL FEATURES

**Figure 1.1** Scheme of order of studies in the dissertation

# CHAPTER 2. LITERATURE REVIEW

## 2.1 *Paphiopedilum* species

### *2.1.1 Scientifiic taxonomy and morphological description*

*Paphiopedilum* Pfitz.1886 is a distinct genus of the Orchidaceae family which has a different flower structure with deep pocket-shaped lip that looks like a colorful lovely slipper. Hence they are commonly called 'Lady's or Venus's Slipper' orchids.

### *2.1.2 Paphiopedilum population in Vietnam*

Vietnam has a diversity of about 20 original species with different varieties and many natural hybrids. According to IUCN Red List 2020-2, 2 species are in vulnerable state, 7 are in endangered state and up to 11 are in critically endangered.

## 2.2 Identification techniques

### *2.2.1 Common identification techniques*

Until now, the most popular method for identification of different organisms is morphological-based classification method (Averyanov *et al.*, 2004). However, in cases that the samples are not totally grown or are partial damaged, the process of identification is difficult or misleading. PCR-based methods can be used with just a small piece of sample. However, an unknown taxon cannot be determined using these PCR-band techniques.

Since each site of the sequence is considered as a character in bioinformatics analysis, the sequence-based method gives more variable information. This approach allows the read of every nucleotide among the samples. The specific and stable features of monomers are useful in evaluating the genetic relationship of a new unknown sample based on the available sequence library and so allow to consult the similar origins of the unknown species. Hence short DNA fragments represented for the organism can be used as an identifying sequence like the human fingerprint. This quick and sensitive method allows researchers to successfully identify species with just a small amount of sample, making it efficient for species conservation, especially at biodiversity hotspots and for the discrimination of medicinal species from adulterants.

### *2.2.2 Identification techniques using DNA sequences - DNA barcodes*

The ‘DNA barcode’ concept was first introduced by Paul Heber in 2003 for “quickly and accurately identifying taxonomic species, through all of living forms from bacteria, fungi, plants to animal and human; a tool for determination of the origin of organism” (Lahaye *et al.*, 2008). The single region 5’ end of cytochrome c oxidase 1 (*CO*1) from the mitochondrial genome has been proved to effectively identify animal individuals (Hebert *et al.*, 2003, Shneer, 2009). However, searching for DNA barcode in plants is far more challenging than in animals. Universal sequences for identification of plants are also being searched from different genomes, mainly from the nucleus and chloroplast.

### *2.2.3 Common bioinformatic analyses in identification techniques using DNA sequences*

Bioinformatic procedure for species identification using sequences comprises two basic steps: First, the sequence alignment must be conducted to the basis of the comparative step. Second, similarity or variation character comparative are investigated based on the available alignment data to achieve the identifying target. The tree-based method is a simple and visualized approach that is most common in such classification. The two species are separated completely when each species sequences are clustered in different branches, called monophyletic relationship (Meyer and Paulay, 2005). Another method also applied in some circumstances the utilization of nucleotide polymorphism, such as variable sites, sequence length variation, indel information, GC% content as tools of identification technique. This approach is known as character-based method in which each nucleotide is consider as the fifth character beside the four traditional characters A, T, C and G.

## 2.3 Research on sequence identification of Orchidaceae species

For slipper orchids, there were a number of studies screening for potential molecular sequences to authenticate species. In Vietnam, a phylogenetic study of Khuat Huu Trung et al. was conducted on 16 Paphiopedilum species genus using a single ITS region (Trung et al., 2013). As the sample size could significantly affect the resolution results (Meyer and Paulay, 2005, Guo et al., 2016), a larger number of samples, as well as additional discriminatory sequences, was recommended to develop a comprehensive identification method. The study of Khuat Huu Trung et al. (2013) was the first in the construction of a molecular database of *Paphiopedilum* in Vietnam. So far, no other research has been carried out on barcoding the Vietnamese *Paphiopedilum* population.

# CHAPTER 3. MATERIALS AND METHODS

## 3.1 Materials and research scope

Fresh leaves of specimens were mostly obtained from the orchid collections of the Tay Nguyen Institute for Scientific Research and the Agricultural Genetics Institute, Vietnam. Plants from these sources had been collected from different geographic areas and provinces of Vietnam and correctly identified based on flower morphological description (Cribb, 1998, Averyanov *et al.*, 2004, Averyanov *et al.*, 2010) by plant experts and were on culture of orchid collections of the Institutes. In addition, some plants collected from private cultures and trading markets were also studied. The samples being at flowered stage were identified based on flower morphological description following Averyanov et al. (2004) by the author and Professor Tran Hop. The mutated sample DEL-158T with white flower and 8 other samples, i.e., ARM-41, CAL-166, CON-115, COC-150, COC-151, TRA-177, TRA-178, and VIE-129 which were non-flowered and unidentified by morphology, were also included to confirm their scientific name based on this identification technique.

In total, 95 samples of 22 Vietnam *Paphiopedilum* species including 20 original species and 2 hybrids were analyzed for species resolution. Two varieties *P. malipoense* var. *malipoense* and *P. malipoense* var. *jackii* were treated as the same species. At least three samples for each species were analyzed except for two species *P. vietnamense* and *P. herrmannii,* which are rare and almost extinct in nature, only one sample was obtained. For each analysis, the closely related Slipper species, *Phragmipedium longifolium*, was included as an outgroup.

## 3.2 Methods

### *3.2.1 DNA extraction and amplification and sequencing for short sequences*

Total DNA from fresh leaves was extracted using Isolate II Plant DNA kit BIO-52069 (TBR company, Ho Chi Minh City, Vietnam). The DNA was then stored in TE solution at −20 °C and used as the template (100 ng per 50 µL reaction volume) for the amplification process.

**Table 3.1** Primers used for amplification reactions in the study.

| Locus | Annealing Temperature ($^o$ C) | Primer name | Primer sequences (5’–3’) | PCR product length | Source |
|---|---|---|---|---|---|
| ITS | 58 | IT1–F | AGTCGTAACAAGGTTTC | 900 | (Tsai, 2011) |
| | | IT2–R | GTAAGTTTCTTCTCCTCC | | |
| *trn*H-*psb*A | 53 | *psb*A3’f | CGCGCATGGTGGTTCACAATCC | 900 | (Hollingsworth *et al.*, 2009) |
| | | *trn*Hf | GTTATGCATGAACGTAATGCTC | | |
| *rpo*B | 53 | 2F | ATGCAACGTCAAGCAGTTCC | 600 | |
| | | 4R | GATCCCAGCATCACAATTCC | | |
| *rpo*C1 | 53 | 1.1F | GTGGATACACTTCTTGATAATGG | 600 | |
| | | 1.3R | TGAGAAAACATAAGTAAAGGGC | | |
| *mat*K | 55 | 2.1F | CCTATCCATCTGGAAATCTTAG | 800 | |
| | | 5R | GTTCTAGCACAAGAAAGTCG | | |
| *mat*K | 55 | F56–mo | GGCAACAAAACTTCCTATA | 1200 | This study |
| | | R1326–mo | TCTAGCACACGAAAGTCGA | | |
| *trn*L | 62 | *trn*L–F | GGTAGAGCTACGACTTGATT | 600 | |
| | | *trn*L–R | CGGTATTGACATGTAAAATGGGACT | | |

PCR reaction components included 10 µl Taq DNA poly 2X - premix (0.1 U / mL Taq DNA Polymerase; 0.4 mM dATP; 0.4 mM dGTP; 0.4 mM dCTP; 0.4 mM dTTP; 4 mM $MgSO_4$; 20 mM KCl; 16 mM $(NH_4)_2SO_4$; 20

mM Tris-HCl, pH8); 1 µl forward primer; 1 µl of reverse primer; 1 µl 50 ng / µl mold DNA; Add $H_2O$ to a volume of 25 µl / reaction.

The thermal cycle was as follows: one cycle of DNA pre-denaturation at 94 °C for 10 min, followed by 30 cycles of 30 s denaturation at 94 °C, 30 s at annealing temperature (Ta °C), and 40 s extension at 72 °C, with a final elongation of 5 min at 72 °C, using SimpliAmp™ Thermal Cycler A24811 (Thermo Fisher Scientific company, Waltham, MA USA). The Ta °C was different depending on the corresponding primer pairs. Details of Ta °C and primers used for the amplification of the ITS, *mat*K, *trn*L regions are shown in Table 3.1.

The quality of all PCR products was checked using the electrophoresis technique for the presence of a clear, unique band in agarose gel 1%. 40 µL volumes of the unpurified PCR products, which with bright, thick, and single band, were sent to 1stBASE company (Singapore) for Sanger sequencing on both forward and reverse directions. The primers used for sequencing were the same as those in the PCR reactions (Table 3.1).

### *3.2.2 Sequence adjustment and alignment*

Raw sequences were trimmed off ambiguous ends using FinchTV (Geospiza, 2004). Reliability and accuracy of raw data were checked by comparing the forward and reverse sequences before the consensus DNA sequence was created using Seaview 4.0 software (Gouy *et al.*, 2009). All consensus sequences were submitted to the GenBank; their accession numbers were shown in Appendix A1. The alignments were managed automatically using the  Seaview software (Gouy *et al.*, 2009) and then manually optimized, especially with non-coding regions in the chloroplast genome and coding regions in the nuclear genome which are highly divergent and contain many indel fragments.

### *3.2.3 Sequence analysis*

Sequences from the study were checked for accuracy by comparing it with reference sequences from the GenBank database using BLAST (Basic Local Alignment Search Tool).

The genetic characteristics as conserved regions, polymorphism sites and specific genetic distance were calculated using MEGA7 software (Kumar *et al.*, 2008). Nucleotide substitutions including Variation, Parsimony, Singleton. Mono-indel, Insertion, Deletion parameters were recorded manually. Unique variable-site characters, also known as species-specific SNP (Single Nucleotide Polymorphism), could help to distinguish one species from the others.

Tree-based construction was carried out using different phylogenetic methods. The neighbor-joining (NJ) method was conducted using MEGA7 (Kumar *et al.*, 2008), and the Maximum Likelihood (ML) and Maximum Parsimony (MP) methods in the PAUP* 4.0 tool (Swofford, 2003), as well as the Bayesian Inference (BA) method in the MRBAYES program (Huelsenbeck and Ronquist, 2001). Tree rooting was performed using the outgroup method. The nucleotide substitution models set up in each phylogeny running were inferred from the jModeltest program (Posada, 2008). The optimal model for ITS, *mat*K, *trn*L, and *mat*K + ITS was K80 + G, TIM1 + I, TPM1uf + I, and TPM1uf + G, respectively. These proposed models for each DNA locus were applied to the PAUP* and MRBAYES programs. The model Kimura-2-parameters was used for MEGA analysis. Bootstrap 1000 was applied for reliability estimations. The tree-topology obtained from all phylogenetic running was visualized using the Figtree v1.4.3 program (Rambaut, 2009).

### *3.2.4 Evaluation of species resolution*

The species resolution was estimated mostly based on the tree-based method, in combination with the polymorphism character-based method.

Firstly, species with just only one accession was distinguishable if the sequence was unique from others. In the phylogenetic tree, this accession would be shown as a monophyletic branch (Hollingsworth *et al.*, 2009).

Secondly, when multiple accessions were collected per species, these all accessions would be grouped into one monophyletic branch in the phylogenetic tree. Thirdly, 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 (Shaw *et al.*, 2007, Wu *et al.*, 2013). Finally, in case of indiscrimination by K2P distance, different species were grouped in the same branch in the phylogenetic tree. This means that the sequences of these accessions were identical. These hetero-species sequences in the same branch would be also more observed with indel information.

### *3.2.5 New primer designing*

Primers for 5 loci *ACO*, *LEAFY*, *atp*B-*rbc*L, *mat*K, *trn*L specific for Vietnamese *Paphiopedilum* were developed through two main steps. First, primers used to amplify these three loci from the published articles were searched. Available primer sequences were aligned with reference sequences of the examined regions of GenBank *Paphiopedilum* accessions downloaded in the study of Chapter 4. Identical nucleotide sites between primers and alignment sequences were screened using Seaview 4.0 (Gouy *et al.*, 2009). The primers which did not match or gave significantly different from the *Paphiopedilum* sequences would be removed from the study. The primers which contain small difference from 0-2 nucleotides would be kept for further tested using Primer3Plus (Untergasser *et al.*, 2007) to restrict the disparity of the annealing temperature (Ta) between forward and reverse primers, hairpin structure forming, repeat forming and dimer forming.

Second, in case no proper primers for *Paphiopedilum* were found, novel primers were newly designed also using Primer3Plus program. A representative sequence of each region was used as a template for the priming in Primer3Plus. The output primers proposed by the program were then put into Seaview software for the alignment and check of conserved characteristics and the length of the amplified products.

All the selected or designed primers were tested for amplification on 54 samples in the study. The PCR products were tested for expression in gel agarose 0.8 %.

### *3.2.6 Vegetative morphological identification*

Studied samples were taken photographs and recorded measurements serving for intra- and inter-specific morphological analysis. The distinguishing vegetative characters were analyzed focus on leaf morphology based on following criteria: blade/lamina shape (oblong, elliptic), size, leaf tip, midrib, vein, small netted vein, leaf margin, leaf base, color uperside, color underside, surface (rough-glossy/smooth), cilia, leaf thickness, leaf toughness, leaf direction. Each characteristic was observed on studied samples and pictures, referenced by monograph book of *Paphiopedilum* (Tran, 1998, Averyanov *et al.*, 2004) for more reliable. Studied species were not described in separate but in comparison with their other sisters. The order of description started from most to fewer notable characteristics which could stand out the differences among species samples. Observed variables based on qualitative features of leaf shape, color, vein feature, thickness, toughness, and quantitative features of leaf size were saved and organized using Microsoft Excel 2010.

### *3.2.7 Methods for plastomic analyses*

#### *3.2.7.1 DNA extraction for Next generation sequencing (NGS)*

For major of the thesis, short sequences were analyzed for identification target. Amplification and Sanger sequencing used for these mini-barcodes required moderate quality of extracted DNA. However, to test the efficiency of a super barcode in comparison with the short ones, one of our samples, the voucher DEL_2 of the *Paphiopedilum delenatii* species was sequenced for whole genome analysis using Next generation sequencing, which required highly quality of total DNA. The extraction using available kits resulted in insufficient amount of DNA for quality test, library construction and sequencing steps. Therefore, DNA extraction was manually applied as follow.

0.2 g leaf was ground with 5 µl proteinase K, 3ml of a mixture of beta-mer + extract buffer at 65 $^{o}$C then incubated for 30 minutes at 65 $^{o}$C. The sample was added with 600 µl P:C:I and centrifuged for 10 minutes at 10000 rpm. After adding 5ul of RNAse and incubated at 37 $^{o}$C, the sample was added with 600ul C:I. DNA was precipitated by isopropanol and incubated overnight at -20 $^{o}$C. The pellet, obtained by centrifugation, was washed with 70%, 80%, 90% ethanol. DNA has suspended in 25 µl TE and stored at -20 $^{o}$C. The library construction and whole-genome sequencing of *P. delenatii* were performed by GENEWIZ (South Plainfield, NJ, USA). Sequencing was carried out on an Illumina HiSeq using a 2x150 pair-end (PE) configuration.

*3.2.7.2 Read data processing and chloroplast genome assembly*

Demultiplexing was performed by bcl2fastq 2.17. Raw data were filtered as follows: Discard pair-end reads with adapter; (2) Discard pair-end reads when the content of N bases is more than 10% in either read, and (3) Discard pair-end reads when the ration of bases of low quality (Q<20) is more than 0.5 in either read.

The chloroplast genome of *P. delenatii* was reconstructed using NOVOPlasty 2.7.2 (Dierckxsens *et al.*, 2016), with the complete chloroplast genome of *P. armeniacum* (RefSeq: NC_026779.1) as the reference genome and the *rbcL* from the same plastid genome of *P. armeniacum* as the seed sequence. The annotation was done by GeSeq (Tillich *et al.*, 2017) and further manually curated by comparing to the annotations of representative species *P. armeniacum*, *P. dianthum*, and *P. niveum* in GenBank. The genome map was drawn by OGDRAW (Lohse *et al.*, 2007).

*3.2.7.3 Repeat sequence and microsatellite identification*

REPuter (Kurtz *et al.*, 2001) was used to calculate DNA repeats including forward, reverse, complement, and palindromic kinds of repeat sequences. The repeats were identified with a hamming distance of 3 and a minimum repeat size of 30 (Dong *et al.*, 2018). MISA (Beier *et al.*, 2017) was used to identify microsatellite sequences with default parameters.

*3.2.7.4 Examination of IR junctions*

We manually examined the IR junctions of all included orchid species. Annotations of IRs, SSC, LSC, and genes were based on their respective annotations on the RefSeq database. For genomes without IR annotations, we used REPuter to identify their pairs of inverted repeats.

*3.2.7.5 Phylogenetic analysis*

The phylogenetic analysis was based on the complete genome sequences of representative orchid species under the maximum likelihood criterion and the GTR + I + G nucleotide substitution model using R package phangorn *(Schliep, 2010)*. Node was calculated from 1000 bootstrap replicates. Figtree (Rambaut, 2009) was used to visualize the resulting tree. The multiple alignment data of these plastomes was used to calculate variable sites and genetic distance matrices using MEGA *(Kumar et al., 2008)*. MAFFT *(Katoh and Standley, 2013)* was used to pairwise align and construct dot-plot graphs.

*3.2.7.6 Nucleotide variability calculation*

DnaSP v6.1 was used to extract the parsimony variable sites density over the plastid genome alignment of four analyzed *Paphiopedilum* species with a sliding window (window length ≤ 600 and step size = 200). Nucleotide diversity was calculated by ratio of Pi and window length. The diversity threshold was 0.079 calculated by sum of the average and double the standard deviation (Bi *et al.*, 2018). Regions with diversity higher than threshold were recommended as highly variable regions.

# CHAPTER 4. RESULTS AND DISCUSSION

## 4.1 The pilot study of recommended sequences ITS, *mat*K, *rpo*B, *rpo*C1 and *trn*H-*psb*A

Five barcodes ITS, *mat*K, *rpo*B, *rpo*C1 and *trn*H-*psb*A were pilot screened on the identification of 8 endemic *Paphiopedilum* species of Vietnam as potential identification markers for further application on Vietnam population.

### *4.1.1 Amplification and sequencing rates*

Four regions of ITS sequences, *mat*K, *rpo*B, *rpo*C1 showed successful PCR amplification results on all 23 studied samples and the sequencing efficiency also reached 100%. However, for success amplification of *mat*K on all 23 samples, we had to repeat multi-PCR reactions for some samples. As for *trn*H-*psb*A region, the amplification rate was at 82.62% and the sequencing rate was extremely low at 31.58%.

### *4.1.2 Sequence characteristics of studied regions*

As for genetic variation, ITS had the highest ratio of all parsimony (166), singleton (71), insertion and deletion (20), followed by *mat*K. The lowest rates belonged to *rpo*B and *rpo*C1.

In terms of intra-specific and inter-specific genetic distance, samples of the same species showed 100% homologous similarity and intra-specific genetic distance was 0, except for *P. dalatense*.

### *4.1.3 Identification of species using the tree-based method*

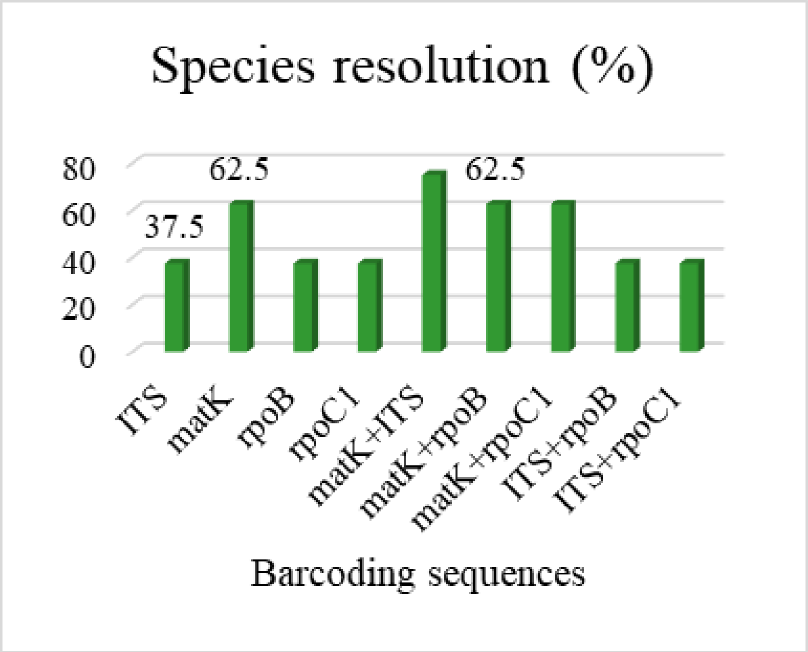

**Figure 4.1** Species resolution of barcoding sequences on 8 endemic *Paphiopedilum* species

As for 4 single-sequence barcodes, *mat*K gave the best resolution (5 over 8 species), which rated 62.5%. The remaining regions ITS, *rpo*B and *rpo*C1 all reached 3 over 8. The combination of two loci allowed identifying capability of 6 over 8 species based on *mat*K + ITS barcode.

### *4.1.4 Discussion about the effective of 5 recommended sequences in identification studied Paphiopedilum species*

The IT1/IT2 primer pair used in many previous studies (Tsai *et al.*, 2003, Gigot *et al.*, 2007), was achieved high amplification efficiency of ITS region on *Paphiopedilum* in this study.

Although the amplification rate of *trn*H-*psb*A was 82.61%, none of the products could be obtained in the first reactions. The sequencing rate was even lower 31,58%. *trn*H-*psb*A is an intergenic spacer region which is considered an unstable region with many repeats and pseudogenes, coupled with a high rate of DNA mutation. The problem was also discussed by previous publications (Gigot *et al.*, 2007, Dong *et al.*, 2014). This locus is not suitable for identification of *Paphiopedilum*.

The combination of two loci allowed identifying capability of 6 over 8 species based on *mat*K + ITS barcode. This result was also consistent with Chochai *et al.* 2012 and other previous studies.

#### *4.1.5 Conclusions*

The two sequences ITS and *mat*K were suggested for further study, especially in combination of both. The primer pair IT1/IT2 for ITS was good but new primers for *mat*K should be considered. *trn*H-*psb*A, *rpo*B and *rpo*C1 were all not suitable for *Paphiopedilum* identification.

### 4.2 *In silico* study on GenBank accessions for searching other potential sequences in identification of *Paphiopedilum* species

From the above pilot results, although the combined marker *mat*K+ITS was reported to be the best, the species resolution till could not reach 100%. Hence other potential sequences should be screened. In this part, for screening other DNA regions, we evaluated 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 individuals of the genus.

#### *4.2.1 Results of searching for reference Paphiopedilum sequences from GenBank*

All sequences belonged to *Paphiopedilum* species from GenBank were selected and downloaded into separate loci. From that, 28 selected loci were applied for further study. In a total number of 107 species, the number and the composition of species were different among different loci.

#### *4.2.2 Multiple sequence alignment and divergence of studied loci*

The nuclear loci including 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%), which were significantly higher than all loci in plastid genome (0.3-12% in parsimony rate, 0–6.5% in singleton rate). 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).

#### *4.2.3 The capability of species identification using single regions*

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%), following by *ACO*, *DEF*4, and *RAD*51 (73.2%, 64.3% and 55.2%, respectively).

Among the chloroplast regions, the intergenic spacer *atp*B-*rbc*L gave the highest species resolution (76.7%), following by *mat*K (1), *trn*L, *rpo*C2 and *ycf*1 (52.5%, 49.4%, 45.9% and 38%, respectively).

#### *4.2.4 The capability of species identification using multiple region combinations*

The resolutions of combination sequences were all higher than of the single loci. In a two-locus combination, 14 out of 36 datasets could be resolved completely (100%) with inter-species relationships. *LFY*, *mat*K and *atp*B-*rbc*L gave the best results in which 5/8 combinations obtained 100% resolution. The followings were *ACO*, *DEF*4, *RAD*51 and *trn*L (4/8, 4/8, 3/8 and 2/8, respectively).

### *4.2.5 Discussion about the use of indel information in species identification*

We totally agreed with the use of indel information as the fifth character state in molecular identification of species, beside four traditional characters A, T, G, C. However, indel information should be stable to be applied on a wide range of species. Otherwise, this information may confuse the sequence comparison which was the same as *trn*H-*psb*A mentioned earlier.

### *4.2.6 Discussion about the identification effect of studied loci from GenBank*

ITS2 was favored than ITS due to higher capability of identification. Other nuclear regions, i.e. *ACO, LFY*, *DEF*4 and *RAD*51 were even better than ITS. However, the total indel number of *DEF*4 and *RAD*51 were significantly higher than *LFY* and *ACO* which makes the sequences difficult to develop good primers for successful amplification.

### *4.2.7 Conclusions*

*ACO* and *LFY* regions from nuclear genome, the non-coding intergenic spacer *atp*B-*rbc*L and the two coding chloroplast sequences *trn*L and *mat*K were recommended for the identification technique. However, for application of these loci, proper primers and amplification capability should be studied. In the next section, we studied specific primers for amplification 5 regions *ACO*, *LFY*, *atp*B-*rbc*L, *mat*K and *trn*L on *Paphiopedilum*.

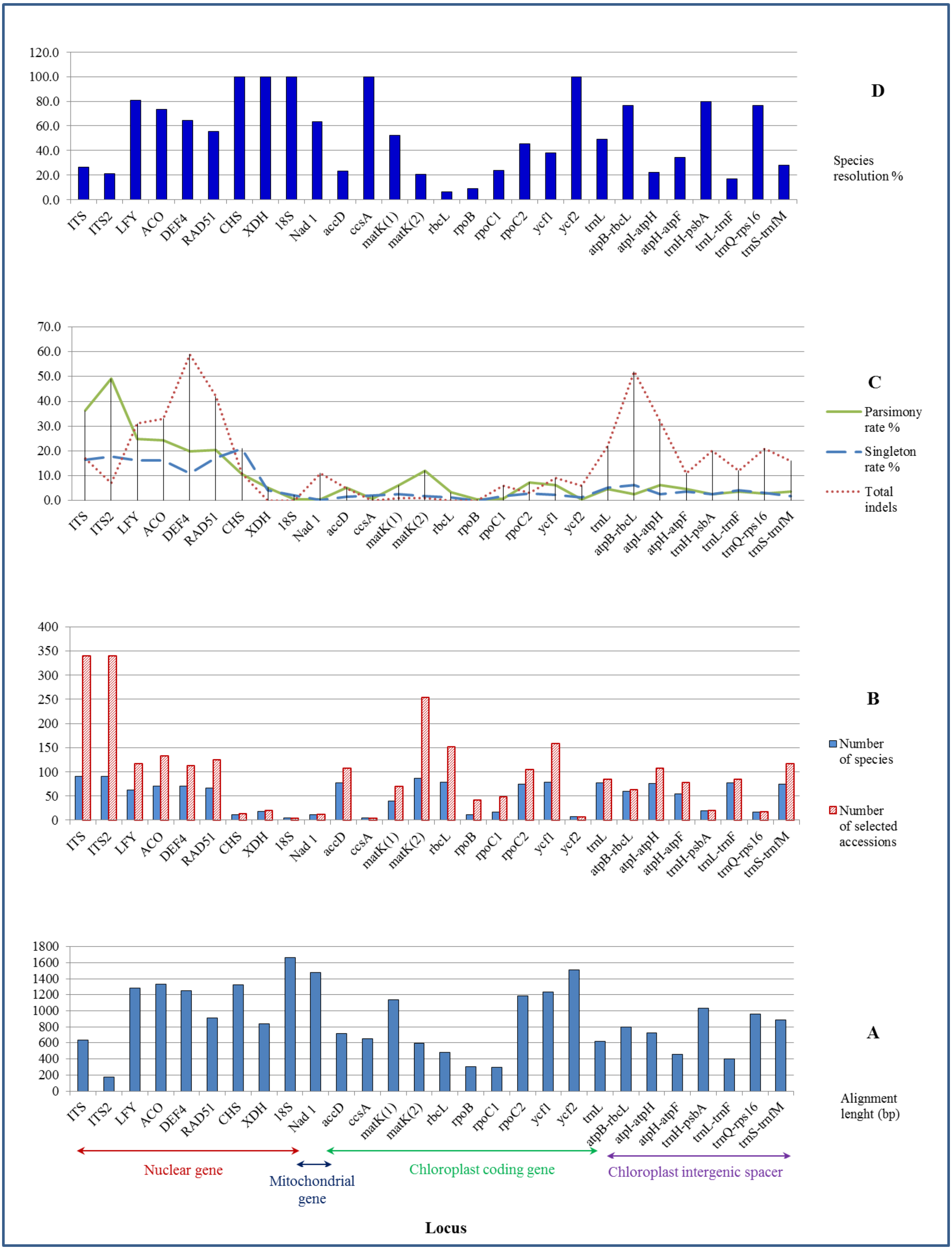

**Figure 4.7** Species resolution of single loci of *Paphiopedilum* sequences and other analyzed information about parsimony rate, singleton rate, indel fragment, number of species, number of sequence and alignment length per locus.

## 4.3 Screening and newly designing primers for more effective identification of *Paphiopedilum* species

Primers for effective amplification and sequencing play an important role in increasing the sensitive of identification performance. As for new suggested regions *ACO*, *LEAFY*, *atp*B-*rbc*L, and *trn*L, the primers used to amplify the sequence of were previously designed to be universal for amplification a broad range of plants and hence were low efficient on specific *Paphiopedilum* or produced unexpected length products. Hence new primers for achieving these sequences needed to be developed for *Paphiopedilum*. In addition, *mat*K showed the best among studied markers from the previous pilot study, but amplification efficiency was still not optimal. Thererfore we also designed new primers for this locus to increase its PCR rate. In total, 5 regions, i.e. *ACO*, *LEAFY*, *atp*B-*rbc*L, *trn*L and *mat*K were included in this part of the study.

### *4.3.1 Primer development for amplification of ACO sequences*

The electrophoresis products of 54 samples gave 24 bright and clear PCR bands with the expected size 1000 bp. However, those primers were inefficient, showing a 44% amplification rate (under 50%) after multiple repetitions and different annealing temperature tests.

### *4.3.2 Primer development for amplification of matK sequences*

The primer pair F56-mo: 5’-GGCAACAAAACTTCCTATA-3’/R1326-mo: 5’-TCTAGCACACGAAAGTCGA-3’ was newly developed. The electrophoresis results by *mat*K showed bright and clear PCR band on 48 out of 54 samples for the first time of amplification with the expected size 1100 bp, and gave 100% success at the second repeats.

### *4.3.3 Primer development for amplification of trnL sequences*

The primer pair trnL-F: 5’- GGTAGACGCTACGGACTTGATT-3’/trnL-R: 5’-CGGTATTGACATGTAAAATGGGACT-3’ was newly designed. Amplification rate of these two loci were all 100%.

### *4.3.4 Primer development for amplification of LFY and atpB-rbcL sequences*

On contrast, amplification rate was 0% for both *LFY* and *atp*B-*rbc*L regions using the new primers. Other primers could not be developed due to excess divergence.

**Table 4.7** Newly designed primers for amplification of *ACO, LEAFY, atp*B-*rbc*L*, trn*L and *mat*K regions on *Paphiopedilum* sequences.

| Locus | Primer name | Sequence | Primer length (bp) | $T_m$ (°C) | Origin | Expected product length (bp) |
|---|---|---|---|---|---|---|
| *ACO* | ACO-F | 5’-TCCCTGTTTCCAACATCTCC-3’ | 20 | 58,4 | This study | 1100 |
| | ACO-R | 5’-GAGGCGATTGACATCCTGTT-3’ | 20 | 58,4 | This study | |
| *LEAFY* | LFY-F | 5'-CGAAGCCCTCCTGTTTCAGT-3' | 20 | 62 | This study | 1090 |
| | LFY-R | 5'-GGACCGTGTGTTGACCATG-3' | 19 | 60 | This study | |
| *trn*L | trnL-F | 5’-GGTAGACGCTACGGACTTGATT-3’ | 22 | 62,5 | This study | 500 |
| | trnL-R | 5’-GAGGCGATTGACATCCTGTT-3’ | 20 | 62,1 | This study | |
| *atp*B-*rbc*L | F-mo | 5'-ATCTAGGATTTACATATAC-3' | 19 | 46 | This study | 800 |
| | R-rbcL | 5'-GTCAATTTGTAATCTTTAAC-3' | 20 | 50 | Tsai et al 2010 | |
| *mat*K | 56F-mo | 5’-GGCAACAAAACTTCCTATA-3’ | 19 | 47 | This study | 1100 |
| | 1326R-mo | 5’-TCTAGCACACGAAAGTCGA-3’ | 19 | 49 | This study | |

### 4.3.5 *Discussion about the effectiveness of the primers*

The new primers for *mat*K result was much better, with 89% after the first amplification reaction, higher than the primers on Indian *Paphiopedilum* (87,5%) in Parveen *et al.* (2012) and Singh *et al.* (2012) (Lee, Chang and Chung, 2011). This rate was even at 100% after one or two repeated reactions.

As for *trn*L, all sequences of *Paphiopedilum* from the GenBank were submitted without accompanying published papers. We could not find the available primers that fixed *Paphiopedilum*. The newly designed primers for this locus was successful applied on all our *Paphiopedilum* samples.

The three regions *LFY*, *ACO* and *atp*B-*rbc*L all met the problem of containing few and short conserved regions inside their nucleotide sequences. None of the new designing primer pairs could satisfy primer criteria. As a result, amplification of these loci was not a success on studied samples although many times of amplification reactions were tried. The issue limited the specific primer building and might be the explanation why *LFY*, *ACO* and *atp*B-*rbc*L were not popularly used in identification research.

For the reasons, the nuclear region ITS and the two chloroplast regions *mat*K and *trn*L, were used for further analyses on the large scale of identification of the Vietnamese *Paphiopedilum* population.

### 4.3.6 *Conclusions*

The two new primer pairs F56-mo: 5’-GGCAACAAAACTTCCTATA-3’/R1326-mo: 5’-TCTAGCACACGAAAGTCGA-3’ for *mat*K and trnL-F: 5’- GGTAGACGCTACGGACTTGATT-3’/trnL-R: 5’-CGGTATTGACATGTAAAATGGGACT-3’ for *trn*L were successfully designed and effectively applied on *Paphiopedilum*. The three loci *ACO*, *LFY* and *atp*B-*rbc*L could not achieved proper primers due to the excess nucleotide divergence and not be recommended for identification of *Paphiopedilum*.

## 4.4 Genetic diversity and species identification of Vietnam *Paphiopedilum* population using DNA sequences

From the pilot study, ITS and *mat*K were proposed as the potential sequences. Besides, new primers for *mat*K and *trn*L really did their job best. Hence in this section we applied sequence identification on large scale of Vietnam *Paphiopedilum* population using 3 loci, i.e ITS, *mat*K and *trn*L, which were both high in species resolution and got proper amplification primers. A total of 95 samples of 22 Vietnam *Paphiopedilum* species including 20 original species and 2 hybrids were analyzed for species resolution (Appendix Table A.2) using tree-based method.

Furthermore, in practical conservation, a quick and accurate tool for identification is most neccesary. There are different bioinformatic algorithms such as Neighbor-joining (NJ), Maximum Likelihood (ML), Maximum Parsimony (MP), Bayesian (BA) or Unweighted Pair Group Method with Arithmetic mean (UPGMA) for tree-based reconstruction using different bioinformatic tools as MEGA, PAUP* and MRBAYES. This section also investigated and discussed the proper tools for the practical use of the barcoding technique in controlling the trade of *Paphiopedilum* genus.

### 4.4.1 *Effects of different bioinformatic tools on the identification of the Vietnam Paphiopedilum population*

A total of 95 samples of 22 Vietnam *Paphiopedilum* species including 20 original species and 2 hybrids were analyzed for species resolution. Single ITS gave the best resolution for 14 out of 21 species, followed by *mat*K (12 species) and *trn*L (6 species).

In terms of species resolution, three methods, ML, NJ, and BA, gave the same identification results, while MP showed a slight lower rate in each of the three loci, ITS, *mat*K, and *trn*L. In terms of time efficiency, the MEGA and MRBAYES programs seemed to meet the criteria.

### *4.4.2 Identification of trading Paphiopedilum samples using DNA sequences*

Standard sequences were applied on 8 samples were collected from trading markets. The approach helped to verify 5 samples and adjusted the correct scientific names of 4 samples.

### *4.4.3 Genetic diversity and the relationship of Vietnam Paphiopedilum species*

The combination of *mat*K+ITS could indeed help to increase the number of identified species to 16. Due to this multilocus combination, *P. dalatense* accessions which were not separated in both single ITS and *mat*K files were now grouped into a monophyletic branch.

### *4.4.4 Discussion about the effective identification using DNA sequences*

Although *mat*K was proposed as the best barcode, with 100% resolution in two previous *Paphiopedilum* studies (Parveen *et al.*, 2012, Rajaram *et al.*, 2019), our study agreed with Parveen *et al.* (2017), who found that denser sampling decreased the resolution of *mat*K (Parveen *et al.*, 2017). However, the combination of *mat*K with another locus was recommended as well.

In our study, it was recommended that *mat*K be directly combined with ITS, resulting in a resolution of 77.27%, higher than that reported by Guo *et al.* (2016). Furthermore, the two Vietnamese endemic species, *P. dalatense*, and *P. herrmannii*, which were not mentioned in Guo *et al.* (2016) were first discussed here. There were some other different results between the two studies. In finally, the application of the results from our study in the identification of Vietnam *Paphiopedilum* species was shown to be effective and practical. All sequences of samples from this study were registered in GenBank.

Our result was also consistent with the observations of Trung *et al.* (2013). However, the study of Trung and his colleagues used the single ITS region and had a small sample size in which only 16 Vietnamese *Paphiopedilum* species, each with only one representative specimen, were included. Hence the intra-specific genetic distance in Trung *et al.* was not examined which could affect the species resolution result.

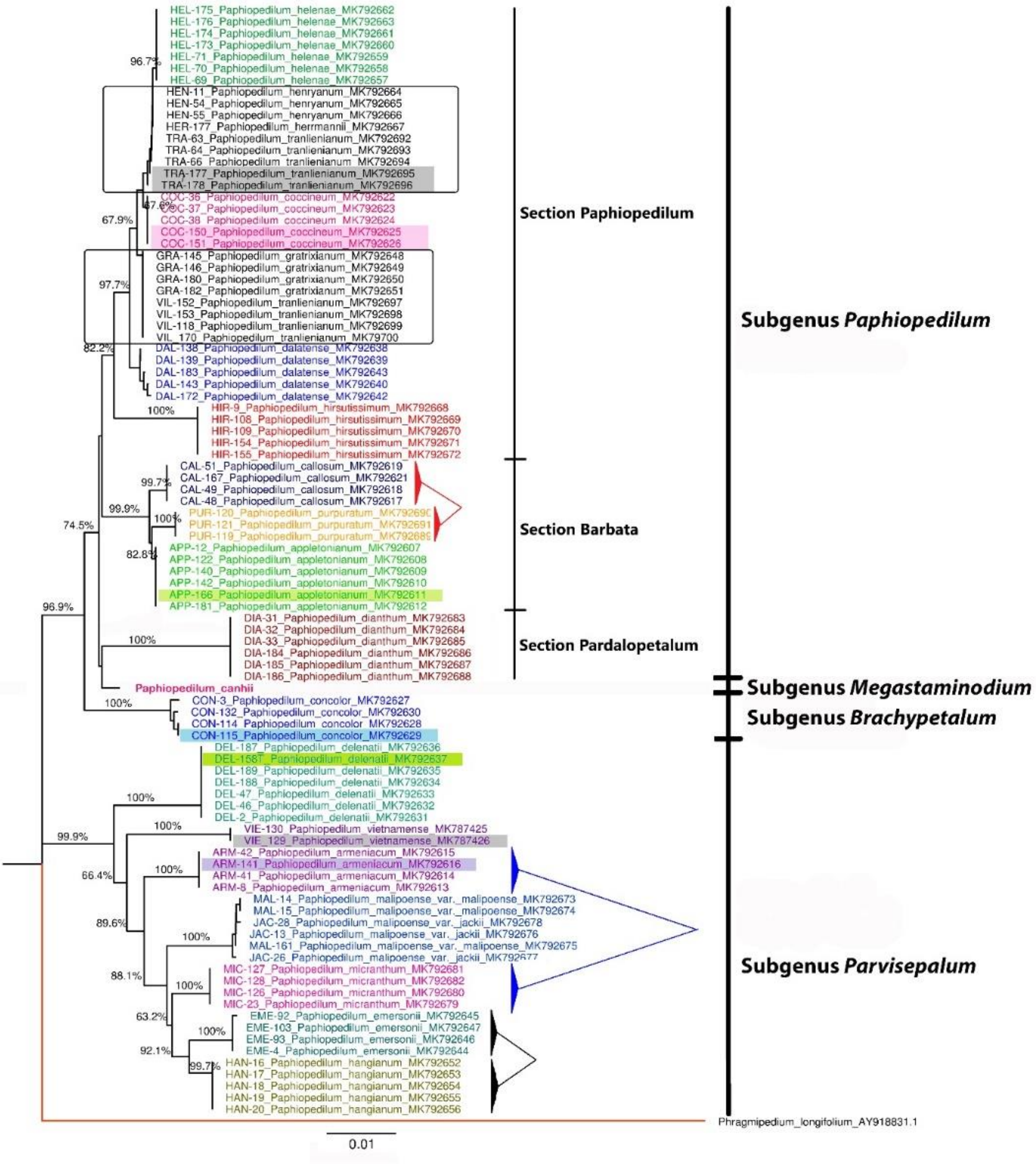
HEL-175_Paphiopedilum_helenae_MK792662
HEL-176_Paphiopedilum_helenae_MK792663
HEL-174_Paphiopedilum_helenae_MK792661
HEL-173_Paphiopedilum_helenae_MK792660
96.7%
HEL-71_Paphiopedilum_helenae_MK792659
HEL-70_Paphiopedilum_helenae_MK792658
HEL-69_Paphiopedilum_helenae_MK792657
HEN-11_Paphiopedilum_henryanum_MK792664
HEN-54_Paphiopedilum_henryanum_MK792665
HEN-55_Paphiopedilum_henryanum_MK792666
HER-177_Paphiopedilum_herrmannii_MK792667
TRA-63_Paphiopedilum_tranlienianum_MK792692
TRA-64_Paphiopedilum_tranlienianum_MK792693
TRA-66_Paphiopedilum_tranlienianum_MK792694
TRA-177_Paphiopedilum_tranlienianum_MK792695
TRA-178_Paphiopedilum_tranlienianum_MK792696
COC-36_Paphiopedilum_coccineum_MK792622
67.6%
COC-37_Paphiopedilum_coccineum_MK792623
67.9%
COC-38_Paphiopedilum_coccineum_MK792624
COC-150_Paphiopedilum_coccineum_MK792625
COC-151_Paphiopedilum_coccineum_MK792626
GRA-145_Paphiopedilum_gratrixianum_MK792648
GRA-146_Paphiopedilum_gratrixianum_MK792649
GRA-180_Paphiopedilum_gratrixianum_MK792650
97.7%
GRA-182_Paphiopedilum_gratrixianum_MK792651
VIL-152_Paphiopedilum_tranlienianum_MK792697
VIL-153_Paphiopedilum_tranlienianum_MK792698
VIL-118_Paphiopedilum_tranlienianum_MK792699
VIL_170_Paphiopedilum_tranlienianum_MK79700
82.2%
DAL-138_Paphiopedilum_dalatense_MK792638
DAL-139_Paphiopedilum_dalatense_MK792639
DAL-183_Paphiopedilum_dalatense_MK792643
DAL-143_Paphiopedilum_dalatense_MK792640
DAL-172_Paphiopedilum_dalatense_MK792642
100%
HIR-9_Paphiopedilum_hirsutissimum_MK792668
HIR-108_Paphiopedilum_hirsutissimum_MK792669
HIR-109_Paphiopedilum_hirsutissimum_MK792670
HIR-154_Paphiopedilum_hirsutissimum_MK792671
HIR-155_Paphiopedilum_hirsutissimum_MK792672
CAL-51_Paphiopedilum_callosum_MK792619
99.7%
CAL-167_Paphiopedilum_callosum_MK792621
CAL-49_Paphiopedilum_callosum_MK792618
99.9%
CAL-48_Paphiopedilum_callosum_MK792617
100%
PUR-120_Paphiopedilum_purpuratum_MK792690
74.5%
PUR-121_Paphiopedilum_purpuratum_MK792691
PUR-119_Paphiopedilum_purpuratum_MK792689
APP-12_Paphiopedilum_appletonianum_MK792607
82.8%
APP-122_Paphiopedilum_appletonianum_MK792608
APP-140_Paphiopedilum_appletonianum_MK792609
APP-142_Paphiopedilum_appletonianum_MK792610
APP-166_Paphiopedilum_appletonianum_MK792611
96.9%
APP-181_Paphiopedilum_appletonianum_MK792612
DIA-31_Paphiopedilum_dianthum_MK792683
DIA-32_Paphiopedilum_dianthum_MK792684
100%
DIA-33_Paphiopedilum_dianthum_MK792685
DIA-184_Paphiopedilum_dianthum_MK792686
DIA-185_Paphiopedilum_dianthum_MK792687
DIA-186_Paphiopedilum_dianthum_MK792688
Paphiopedilum_canhii
CON-3_Paphiopedilum_concolor_MK792627
100%
CON-132_Paphiopedilum_concolor_MK792630
CON-114_Paphiopedilum_concolor_MK792628
CON-115_Paphiopedilum_concolor_MK792629
DEL-187_Paphiopedilum_delenatii_MK792636
DEL-158T_Paphiopedilum_delenatii_MK792637
DEL-189_Paphiopedilum_delenatii_MK792635
DEL-188_Paphiopedilum_delenatii_MK792634
100%
DEL-47_Paphiopedilum_delenatii_MK792633
DEL-46_Paphiopedilum_delenatii_MK792632
DEL-2_Paphiopedilum_delenatii_MK792631
100%
VIE-130_Paphiopedilum_vietnamense_MK787425
99.9%
VIE_129_Paphiopedilum_vietnamense_MK787426
ARM-42_Paphiopedilum_armeniacum_MK792615
ARM-141_Paphiopedilum_armeniacum_MK792616
66.4%
100%
ARM-41_Paphiopedilum_armeniacum_MK792614
ARM-8_Paphiopedilum_armeniacum_MK792613
MAL-14_Paphiopedilum_malipoense_var._malipoense_MK792673
MAL-15_Paphiopedilum_malipoense_var._malipoense_MK792674
89.6%
JAC-28_Paphiopedilum_malipoense_var._jackii_MK792678
100%
JAC-13_Paphiopedilum_malipoense_var._jackii_MK792676
MAL-161_Paphiopedilum_malipoense_var._malipoense_MK792675
JAC-26_Paphiopedilum_malipoense_var._jackii_MK792677
88.1%
MIC-127_Paphiopedilum_micranthum_MK792681
MIC-128_Paphiopedilum_micranthum_MK792682
100%
MIC-126_Paphiopedilum_micranthum_MK792680
MIC-23_Paphiopedilum_micranthum_MK792679
63.2%
EME-92_Paphiopedilum_emersonii_MK792645
100%
EME-103_Paphiopedilum_emersonii_MK792647
EME-93_Paphiopedilum_emersonii_MK792646
92.1%
EME-4_Paphiopedilum_emersonii_MK792644
99.7%
HAN-16_Paphiopedilum_hangianum_MK792652
HAN-17_Paphiopedilum_hangianum_MK792653
HAN-18_Paphiopedilum_hangianum_MK792654
HAN-19_Paphiopedilum_hangianum_MK792655
HAN-20_Paphiopedilum_hangianum_MK792656
Phragmipedium_longifolium_AY918831.1
0.01
Section Paphiopedilum
Section Barbata
Section Pardalopetalum
Subgenus Paphiopedilum
Subgenus Megastaminodium
Subgenus Brachypetalum
Subgenus Parvisepalum

**Figure 4.16** Phylogenetic tree of 22 Vietnamese *Paphiopedilum* species based on *mat*K + ITS barcode using the neighbor-joining method in the MEGA software. The sequences of *Phragmipedium* from the GenBank database were used as outgroups.

### *4.4.5 Discussion about bioinformatics tools and analyses for identification using tree-based method*

For tree-based reconstruction, the probability model should be estimated. Neighbor-joining (NJ), Maximum Likelihood (ML), Maximum Parsimony (MP), Bayesian (BA) or Unweighted Pair Group Method with Arithmetic mean (UPGMA) are common phylogenetic algorithms inferred along with suitable models. The common software and programs for phylogenetic tree estimating were MEGA, PAUP* and MRBAYES. In this study, MRBAYES and MEGA with Neighbor-joining method were recommended for practical use.

### *4.4.6 Conclusions*

Our study was applied in the certain area that is Vietnam. A total of 22 species including 20 original species and 2 hybrids belonging to *Paphiopedilum* population were examined. The results returned in 17 out of 22 species of Vietnam *Paphiopedilum* species were well-identified. We also suggested the use of Neighbor-joining in MEGA software. MRBAYES could also be a good alternate. The application of the results from our study in the identification of Vietnam *Paphiopedilum* species was shown to be effective and practical. All sequences of samples from this study were registered in GenBank.

## 4.5 The support of molecular identification for vegetative morphological features on Vietnam *Paphiopedilum* species

Morphology identification is a crucial technique serving for many applications. Traditional identification of *Paphiopedilum* based on flower morphology is the most popular. However, most of plant life are at the stage of non-flower. This Chapter constructed detailed description data of Vietnam *Paphiopedilum* species based on vegetative morphology. The study aimed to confirm the support of molecular features in quick and effective identification of Vietnam *Paphiopedilum* using the combination of vegetative morphology and DNA sequences.

94 *Paphiopedilum* plants belonging to 20 original species including 2 varieties of genus Paphiopedilum in Vietnam, excluding *P. canhii* and *P. herrmannii* due to sample degradation, were used for this study. Those samples have been collected from different geographic areas and provinces of Vietnam which were preliminary identified by both floral characteristics (Averyanov et al., 2004) and molecular barcodes of ITS and *mat*K sequences from Chapter 6.

All samples were taken photographs and recorded measurements serving for intra- and inter-specific morphological analysis.

### *4.5.1 General characteristics of vegetative organs of Paphiopedilum genus*

*Paphiopedilum* species are herbal plants with extremely short rhizomes and stems except *P. malipoense* and *P. micranthum* with linked multi-rhizomes into underground nets. There are 3 – 7 distichous leaves which are elliptic, oblong, or obovate. Leaf apex is round or acute and often asymmetrically bilobate or trilobate. Leaf base is conduplicate into V-shaped, enclosing around stem. Blade is plain green pattern or tessellated with dark and pale green above. This mosaic pattern is very typical for many *Paphiopedilum* orchids. These mosaic spots are thought to be due to the uneven distribution of chlorophylls on the leaf surface resulting in alternate light and dark spots (Cribb, 1998). Some species spotted with purple spots at base or throughout the lower surface. The characteristics and density of the blue mosaic spots on the upper and lower surfaces of leaves are specific to species when observing in detail.

### *4.5.2 Distinguish Paphiopedilum species using vegetative polymorphism*

14 out of 20 species were clearly distinguished. Six species that were misidentified had morphological similar in pairs, i.e. *P. hangianum* versus *P. emersonii*, *P. callosum* versus *P. purpuratum*, and *P. armeniacum* versus *P. micranthum.* Using molecular sequences, all these species were separated into different monophyletic clades with high support values from 92.1%. to 99.9%. Hence the combination of vegetative and molecular identification techniques could resolve up to 20 species.

### *4.5.3 Artificial key based on leaf morphology to the genus Paphiopedilum*

An artificial key based on leaf morphology to the genus *Paphiopedilum* were described which helped to distinguish 14 out of 20 species.

### *4.5.4 Combination of vegetative morphological and molecular methods in identification of Vietnam Paphiopedilum species*

Using molecular sequences, all these species were separated into different monophyletic clades with high support values from 92.1%. to 99.9%. Hence the combination of vegetative and molecular identification techniques could resolve up to 20 species.

### *4.5.5 Discussion about the combination of vegetative morphological and molecular methods in identification of Vietnam Paphiopedilum species*

Discrimination and identification of species are basis for conservation, development, or breeding of valuable genetic resources. Various identification methods have been developed, of which morphological and molecular techniques are commonly used. However, research based only on vegetative morphology was hardly found so far.

In this study, artificial key for 20 *Paphiopedilum* species of Vietnam were also established which did not show the differences or similarities between taxa but described leaf traits in an order manner so that readers can easily lookup for species identification. Furthermore, the combination of vegetative and molecular identification techniques could resolve up to 20 species. We recommended this new approach for effective identification of *Paphiopedilum* that was significantly contributed to practical conservation of the genus.

### *4.5.6 Conclusions*

According to our observation, some detailed leaf characters could be utilized as effective tools for quickly identification of *Paphiopedilum* taxa even to species that similar in morphology and closed in relationship. In this study, artificial key for 20 *Paphiopedilum* species of Vietnam were also established which did not show the differences or similarities between taxa but described leaf traits in an order manner so that readers can easily lookup for species identification. Furthermore, the combination of vegetative and molecular identification techniques could resolve up to 20 species. We recommended this new approach for effective identification of *Paphiopedilum* that was significantly contributed to practical conservation of the genus.

## 4.6 Complete chloroplast genome of *Paphiopedilum delenatii* for identification of *Paphiopedilum* species

We carried out next generation sequencing of *P. delenatii* and assembled its complete chloroplast genome. Our end goal is providing genome information for further identification studies and evaluating the identification capability of whole genome in comparison with short sequence barcoded.

### *4.6.1 Chloroplast genome of Paphiopedilum delenatii*

The assembled chloroplast genome of *P. delenatii* exhibited a quadripartite structure of 160955 bp and a total of 130 genes.

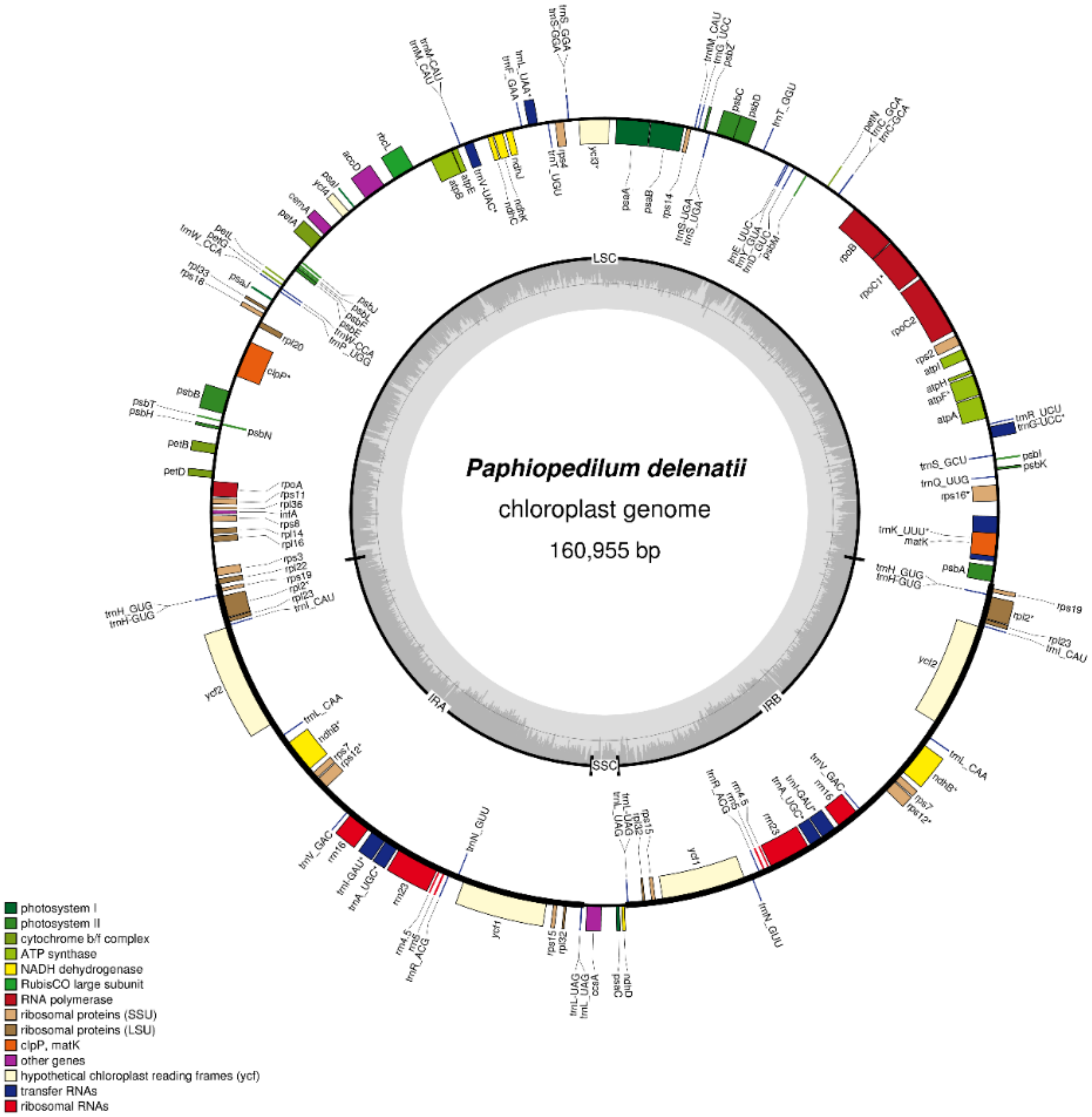

**Figure 4.24** Gene map of *Paphiopedilum delenatii* chloroplast genome.

### *4.6.2 Phylogenetic and species resolution analyses*

Whole-plastome tree and phylogenetic trees based on two popular short barcodes *mat*K and *rbc*L of 16 representative species were constructed. All three phylogenetic trees presented a full separation of all 16 species.

### *4.6.3 Divergence of hotspot regions*

We calculated the nucleotide variability in all chloroplast genomes of four analyzed species *Paphiopedilum delenatii*, *Paphiopedilum armeniacum*, *Paphiopedilum niveum* and *Paphiopedilum dianthum*. Eight highly variable regions were suggested as the potential markers in barcoding and phylogeny study of *Paphiopedilum* genus.

### *4.6.4 Discussion about the role of complete chloroplast genomic sequence as meta-barcode in species identification*

Using short sequences for identification of species is still the universal method up to now due to its simple and time-saving trait. Indeed, the average divergence of *mat*K (0.122) was even higher than of the plastome (0.112) in our study. The reason is that the plastome contains both diverse and conserved regions while *mat*K is a selected-high variable locus. This result proved the species-separation capability of this mini-barcode in comparison with the super-one.

However, as none of can resolve entirely plant species, chloroplast genome was proposed to be used as meta-barcode by some previous studies (Nock *et al.*, 2011a, Singh *et al.*, 2012, Chen *et al.*, 2018, Krawczyk *et al.*, 2018). In this study, the min specific genetic distance of *matK* (0.005) was much lower than that of the plastome (0.012). The result suggested that the entire cp genome could do better than *mat*K in identifying closely related species.

### *4.6.5 Discussion about the use of plastomic information to develop other identification tools*

The sequence data of *P. delenatii* complete chloroplast genome could be used directly in identification of *Paphiopedilum* species based on the differences in GC%, total length, IR template, number of genes, and phylogenetic trees; or for development of other identification markers such as SSR, barcoding or species-specific PCR-based techniques.

### *4.6.6 Conclusions*

Since the cost for whole genome sequencing has significantly decreased, from $2.7 billion in 2003 for the first human genome to $300,000 in 2006 and from that to $1,000 in 2016, a series of studies on barcoding plants was published in the two following years 2017 and 2018. Along with sequencing and assembling techniques, whole-plastome barcode may offer more informative sites and is considered as an accurate and effective single barcode for identification in plants. Our result suggested that the entire cp genome could do better than *mat*K in identifying closely related species. Besides, you can also use this meta-data to develop potential mini-barcodes which are high variability for quick authentication of some certain taxa. Divergent and conserved regions in genome provided useful information to establish DNA-based as well as PCR-based identification markers serving protection and management of species.

# CHAPTER 5. GENERAL CONCLUSIONS AND RECOMMENDATIONS

## 5.1 The main results of the dissertation

For application of DNA identification on *Paphiopedilum* population in Vietnam, a total of 22 species including 20 original species and 2 hybrids were examined. Accordingly, 17 out of 22 species of Vietnam *Paphiopedilum* species were well-identified.

In the barcoding technique using DNA sequences, the chosen loci are causally related to the identification results. The experiments with different loci, i.e. ITS, *LFY*, *ACO* from nuclear genome and *mat*K, *rpo*B, *rpo*C1, *trn*H-*psb*A, *atp*B-*rbc*L, *trn*L from the chloroplast genome recommended the use of combined ITS + *mat*K as the most effective method for the molecular identification of Vietnamese *Paphiopedilum* species. Besides, single ITS was also used in combination with nucleotide polymorphism to increase the species resolution. A potential barcode should be balanced between high divergence for high species resolution and a sufficiently conserved level for the design of universal primers.

The neighbor-joining method using MEGA software was proposed for simple and effective use for barcoding of target plants. In addition, the indel information could be used as effective supporting data for molecular discrimination of species.

The study confirmed the use of available primers IT1/IT2 for ITS amplification and contributed the two new primer pairs F56-mo: 5’-GGCAACAAAACTTCCTATA-3’/R1326-mo: 5’-TCTAGCACACGAAAGTCGA-3’ for more effective amplification of *mat*K and the primer pair trnL-F: 5’-GGTAGACGCTACGGACTTGATT-3’/trnL-R: 5’-CGGTATTGACATGTAAAATGGGACT-3’ for *trn*L applied on *Paphiopedilum*.

For practical identification of *Paphiopedilum* species, DNA sequences proved to be effectively used as the support step following preliminary leaf morphology classification on unidentified samples when resolved up to 20 species. This strategy can help to reduce the time and cost in biodiversity control and national resource conservation assignment.

The complete chloroplast genome of the species *Paphiopedilum delenatii* was sequenced and assembled. This plastome and other available plastomes of *Paphiopedilum* from GenBank were analyzed for barcoding effect in comparison with short DNA sequences. Although the average intra-specific genetic distances were not much different, the min value using whole plastome was much lower, which in turn could better identify closely related species. These super-barcodes could be used directly in identification or for development of other identification markers such as SSR, barcoding or species-specific PCR-based techniques.

## 5.2 The scientific and practical contributions of the dissertation

1. Constructing the sequence datasets of ITS, *mat*K and *trn*L regions for species of *Paphiopedilum* population of Vietnam. All data from the study was upload to GenBank library for extensive research. Our results were also given back to scientific Institutes which will contribute to the conservation and control of the illegal trade of *Paphiopedilum* species in Vietnam.

2. Sequences of the natural hybrid endemic species *Paphiopedilum x dalatense* of Vietnam were first reported in GenBank.

3. Proposing the use of the combination sequence matK + ITS for identification of Paphiopedilum in Vietnam and on the world.

4. Developing new effective primers for amplification of the universal locus *mat*K, and the *trn*L fragment.

5. Providing the detailed vegetative artificial key for quick identification of *Paphiopedilum* in practical as a support tool in the combination with molecular tool.
6. Successful identifying 17 Vietnam Paphiopedilum species using DNA sequence and  20 species using the combination of vegetative and DNA features.
7. Giving the comparison among different bioinformatics tools and different bioinformatics analyses in barcoding technique and suggesting the proper tools for practical use.
8. The complete chloroplast genome of the species *Paphiopedilum delenatii* was first sequenced, assembled, annotated, and registered with GenBank. Character information from this platome was recommended to be effectively used for identification of closely related taxa and for developing many other identification markers.

### 5.3 Recommendation

Up to now, mini barcodes are still the most popular and convenient due to cost-effective price of short sequencing. We proposed to use the Paphiopedilum sequence information published on GenBank for identification species serving for the management and conservation of Paphiopedilum resources in Vietnam.

Although it is still costly and time-consuming than short DNA sequencing, genome sequencing costs have decreased significantly in recent years. Some technologies have evolved in direct sequencing without going through a separate amplification step, e.g. the nanopore technique. We recommended further research and development of whole genome sequencing and the application of this information in easy and efficient species identification in the future.

## PUBLICATIONS

### International articles

1. **Vu H-T**, Nguyen T-D, Vu Q-L, Tran N, Nguyen T-C, Luu P-N, Tran D-D, Nguyen T-K & Le L* (2020). "Genetic Diversity and Identification of Vietnamese *Paphiopedilum* Species Using DNA Sequences". *Biology*, **9**(1):9. (ISI/SCIE, IF 3.796)
2. **Vu H-T**, Tran N, Nguyen T-D, Vu Q-L, Bui M-H, Le M-T & Le L* (2020). "Complete Chloroplast Genome of *Paphiopedilum delenatii* and its relationships to other *Paphiopedilum* species". *Plants*, **9**(1):61. (ISI/SCIE, IF 2.762)
3. **Vu H-T** & Le L (2019). "Bioinformatics Analysis on DNA Barcode Sequences for Species Identification: A Review". *Annual Research & Review in Biology*, **34**(1):1-12.
4. **Vu H-T**, Bui M-H, Vu Q-L, Nguyen T-D, Tran H, Khuat H-T & Le L* (2019). "Identification of Vietnamese *Paphiopedilum* Species Using Vegetative Morphology". *Annual Research & Review in Biology*, **34**(1):1-14.
5. **Vu H-T**, Huynh P, Tran H-D & Le L* (2018). "In Silico Study on Molecular Sequences for Identification of *Paphiopedilum* Species". *Evolutionary Bioinformatics*, **14**:117693431877454. (ISI/SCIE, IF 2.203)

### Domestic articles

6. Nguyễn Thanh Điềm, Lê Thị Lý, Nguyễn Hữu Thuần Anh, Vũ Quốc Luận, Nguyễn Thành Công & **Vũ Thị Huyền Trang*** (2020). "Xây dựng bản đồ bộ gen lục lạp hoàn chỉnh của loài lan Hài hồng (*Paphiopedilum delenatii* Guillaumin 1924) đặc hữu Việt Nam". *Tạp chí Công nghệ Sinh học*, **18**(1):87-102.
7. Đặng Văn Khải, Nguyễn Thị Nhã & **Vũ Thị Huyền Trang*** (2017). "Lựa chọn, thiết kế, thử nghiệm và ứng dụng một số mồi nhằm khuếch đại các vùng trình tự tiềm năng để nhận diện các loài lan Hài (*Paphiopedilum*) Việt Nam". *Tạp chí Nông nghiệp và Phát triển Nông thôn*, **2017**:113-118.
8. **Vu H-T**, Le L, Nguyen T-K, Tran D-D & Tran H-D* (2017). "Review on molecular markers for identification of Orchids". *Vietnam Journal of Science and Technology (version English)*, **59**(2):62-75.