Plant and Fungal Systematics 68(1): 223–231, 2023
DOI: https://doi.org/10.35535/pfsyst-2023-0003

ISSN 2544-7459 (print)
ISSN 2657-5000 (online)

In silico analysis of potential loci for the identification of *Vanda* spp. in the Philippines

Euricka Mae F. Rodriguez[1], Ma. Sophia O. Racelis[1], Anna Alliah D. Calonzo[1], Richard F. Clemente[1], Oliver R. Alaijos[1] & Christian Joseph N. Ong[2*]

Article info

Received: 28 Aug. 2022
Revision received: 24 Apr. 2023
Accepted: 30 May 2023
Published: 31 July 2023

Associate Editor

Marcial Escudero

Abstract. Difficulties in identifying *Vanda* species are still encountered, and the ambiguity in its taxonomy is still unresolved. To date, the advancement in molecular genetics technology has given rise to the molecular method for plant identification and elucidation. One hundred twenty-five (125) gene sequences of *Vanda* species from the Philippines were obtained from the NCBI GenBank. Four of the 25 loci were further examined using MEGA 11 software for multiple sequence alignment, sequence analysis, and phylogenetic reconstruction. The indel-based and tree-based methods were combined to compute the species resolution. The result showed that ITS from the nuclear region obtained the highest species resolution with 66.67%. It was then followed by *psbA-trnH*, *matK*, and *trnL-trnF* from the chloroplast genome with a species resolution of 60%, 40%, and 30.77%, respectively. ITS and *psbA-trnH* satisfied the ideal length for DNA barcoding as they have 655 bp and 701 bp, respectively. The locus *psbA-trnH* was also considered to have a higher potential to discriminate *Vanda* species since only a few sequences were tested for ITS. Furthermore, ITS and *trnL-trnF* have the highest variable rate, which is 2.9%, while *matK* and *psbA-trnH* have 2% and 1.3%, respectively. This showed the nature of the unique sequences of various species. In this study, the indel-based method provided better results than the tree-based method. It will help support further DNA barcoding studies and strengthen the conservation and protection of *Vanda* spp. in the Philippines.

Key words: DNA barcoding, in silico, ITS, *psbA-trnH*, *Vanda*

Introduction

Orchidaceae is one of the largest monocotyledonous families in the Philippines. Approximately 1,200 species are recorded, 85% of which are endemic to the country (Dizon et al. 2018). Orchids are widespread as they can either be aquatic, epiphytic, lithophytic, or terrestrial. They also have a variety of economic benefits including ornamental, medicinal, food, and cosmetic applications. Some orchid species are also ecologically significant as bioindicators of environmental health, and as habitats for microorganisms (Panal et al. 2015).

There is a lack of studies concerning the classification and identification of *Vanda* spp. in the Philippines. In Thailand, the use of in silico analysis of molecular data has broadened the study of *Vanda* spp. for identification and conservation. In silico analysis had provided new tools for plant molecular identification (Tanee et al. 2012). The genus *Vanda* is represented by 17 species, ten of which are endemic to the Philippine archipelago. As stated in the Department of Environment and Natural Resources Administrative Order No. 2017–11 (DENR Administrative Order 2017), *Vanda lamellata* Lindi. and *Vanda sanderiana* (Rchb.f.) Rchb.f are classified as critically endangered. Furthermore, *Vanda javierae* D.Tiu ex Fessel & Liickel, *Vanda luzonica* Lober ex Rolfe, *Vanda merrillii* Ames & Quisumb., and *Vanda scandens* Holttum are categorized as endangered. Biodiversity conservation and the sustainable use of orchid plant species rely on accurate species identification, which can be done by examining the species at a genetic level (Lahaye et al. 2008). DNA barcoding has increasingly become more accessible and is widely used to support conservation efforts.

DNA barcoding is a method employed to distinguish a species at any stage of development through a segment of DNA (Vu et al. 2017). It can be a cheaper alternative tool for promoting more standardized and reliable species identification (Vu et al. 2017). In plants, the universal barcode locus is still unknown. The Consortium for the

[1] Department of Biology, College of Science, Bulacan State University, Malolos, Bulacan, Philippines 3000
[2] Department of Biology, College of Science, De La Salle University, Manila, Philippines
ORCID: 0000-0002-0096-9773
[*] Corresponding author: e-mail: christian_joseph_ong@dlsu.edu.ph

Barcode of Life (CBOL) Plant Working Group (2009) included universality, sequence quality and coverage, and discrimination as the important criteria in evaluating candidate loci for it to be a standard DNA barcode for land plants.

There were several loci suggested for the family of orchids. The following are *rbcL*, *psaB*, *atpB*, and *matK* in the chloroplast genome, 18S and *Xdh* regions in the nuclear genome, and *nad1b-c* in the mitochondrial genome (Kim et al. 2013). In Thailand, *rbcL* + *matK* is the multi-locus barcode proposed for *Vanda* spp. (Kim et al. 2013; Tanee et al. 2012). The use of the mitochondrial genome in plants revealed adverse effects, such as influencing the rate of synonymous substitution to a low level, alteration in the genome structure, and the import of nucleus and chloroplast sequences (Vu et al. 2017). Despite the poor qualities of the mitochondrial genome regarding plant species discrimination, recent studies still consider utilizing it for further taxonomic and phylogenetic assessment.

The main objective of this in silico study was to evaluate the candidate locus that has the capacity to accurately identify *Vanda* spp. collected from the Philippines, which were available on the National Center for Biotechnology Information (NCBI) – GenBank. The study may contribute to the use of proper molecular sequences as effective barcoding markers for the identification of *Vanda* spp., for the purposes of breeding, conservation and diversity research of this orchid plant.

Material and methods

Sequence data acquisition

Only 9 of the 17 Philippine *Vanda* spp. have available accession numbers at NCBI GenBank Nucleotide Database. These are *V. furva*, *V. lamellata*, *V. luzonica*, *V. merrillii*, *V. miniata*, *V. roeblingiana*, *V. sanderiana*, *V. tricolor*, and *V. ustii*. The collection of the available DNA sequences of their synonyms and varieties were considered because a higher number of sequences is ideal for generating the necessary data for analysis. The additional accessions came from *V. aurantiaca* (synonym: *Ascocentrum aurantiacum*), *V. lamellata* var. *boxallii*, *V. lamellata* var. *remediosae*, *V. mariae* (synonym: *V. limbata*), *V. mindanaoensis* (synonym: *V. lindenii*), *V. miniata* (synonym: *Ascocentrum miniatum*), *V. sanderiana* (synonym: *Euanthe sanderiana*), and *V. tricolor* var. *suavis*. Upon initial collection, 105 GenBank accessions were acquired from these species. Each accession was then reinspected, and 125 DNA sequences were retrieved from the 25 loci that were located. The number of accession numbers and the number of species between each locus varied. Of the 25 loci, 21 have at least one to three sequences and were not used further in the analyses because bootstrapping requires four or more taxa to run. In this study, 90 DNA sequences from four loci underwent additional examination. However, the elimination of sequences per locus would still be possible depending on the parameters that would be evaluated.

Multiple sequence alignment

Multiple sequence alignment (MSA) has been conducted and utilized in the MEGA 11 software version 11.0.10. for this specific procedure. The unaligned sequences from each locus were imported into a blank alignment window where homologous characters were identified and aligned (Hall 2013) using the Multiple Sequence Comparison by Log Expectation (MUSCLE) algorithm and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering method. Both were carefully chosen as the results obtained from this procedure were used to assess the nucleotide polymorphism and construct the phylogenetic tree of each candidate locus.

In choosing the clustering method, UPGMA was employed as this gives a marginally higher benchmark score against Neighbor-joining (NJ) (Edgar 2004a, b). The MSAs obtained were then evaluated and 'cleaned'. Surfeit nucleotides were trimmed-off at the beginning and end regions of the sequences. If left untrimmed, the highly varied sequences might split conspecific sequences into the tree branches as they can be interpreted as polymorphisms.

Sequence variation and indel information analysis

The sequence variation was manually probed using the Sequence Data Explorer function in MEGA 11 to highlight and record constant sites, variable sites, parsimony informative sites, and singleton sites. The formula based on the recent study of Nguyen et al. (2020) was used in computing the rate for each site is as follows:

$$\frac{Pi, S, C \text{ or } V}{bp} \times 100$$

wherein:
Pi = parsimony-informative
S = singleton
C = conserved
V = variable
bp = alignment length

Gaps were also observed in the individually aligned loci. It represents the indels, carrying important evolutionary information (Chatzou et al. 2016). Most researchers are still debating whether to remove or retain the indels in the aligned sequences. Either option is crucial for reconstructing the phylogenetic tree because overlooked errors during the alignment will result in poor tree estimation (Tan et al. 2015). However, the indels contained in the aligned locus of *matK*, *trnH-psbA*, *trnL-trnF*, and ITS were reported to contribute to the enhanced resolution of phylogenetic analysis and species-level identification of plants (Sanyal et al. 2015).

Phylogenetic reconstruction analysis

Unrooted is the type of phylogenetic tree utilized in this study because it is advantageous when presenting sequences grouped according to their similarities or differences (Kinene et al. 2016). On the other hand, Neighbor-joining (NJ) is the statistical method used in its construction. A separate phylogenetic tree for each locus

was built to analyze the relationship among species. One thousand (1,000) bootstrap replicates were also conducted alongside the tree building to establish the confidence level of the species clustered together. A bootstrap value ≥90% is the objective, meaning that a particular clade or branch is strongly supported. To identify the most suitable nucleotide substitution model, the Best-fit Substitution Model analysis feature in MEGA 11 was used (Hall 2013). Among the 24, the T92 model obtained the lowest Bayesian Information Criterion (BIC) score and is posited to describe the best substitution pattern.

Computation of tree-based and indel-based species resolution

The species resolution could be measured based on sequence length, variable sites, and the combination of the tree-based method and indel fragments (Vu et al. 2018). All were evaluated in this study; however, the species resolution was more concentrated on the latter, as suggested by Vu et al. (2017). The species resolution was also calculated using the formula based on the recent study of Nguyen et al. (2020). Before adding the total number of successfully identified species from both the tree-based and the indel-based methods, it is important to remember that it would only be counted as one if a species is already considered identified either in one or both methods. The formula used is as follows:

$$\frac{T+I}{S} \times 100$$

wherein:
T = total number of identified species in the tree-based method
I = total number of identified species in the indel-based method
S = total number of the available species per locus

Results and discussion

Evaluation of the potential loci

Several factors must be considered to select the best loci for identifying Philippine *Vanda* spp. from the candidate loci. First, the nucleotide sequences must not contain ambiguous sites and must have a length ranging from 400 bp to 800 bp. As mentioned earlier, sequences that are too long might not be appropriate for extraction, amplification, and sequencing, while those that are too short might not have adequate divergence information. Next, the locus should have a good variable rate among the sequence variation, as this will aid the reliability of the data gathered from the indel-based method. Lastly, a high species resolution will serve as the fundamental criterion. However, the number of species assessed for every locus must be considered since they greatly vary for each locus.

Candidate loci for identifying Philippine *Vanda* spp.

The number of retrieved DNA sequences for each locus located in the varying genome was presented (Figs 1, 2). None were specifically found in the mitochondrial genome upon locating every locus within the acquired nucleotide sequences. Most were positioned either in the chloroplast

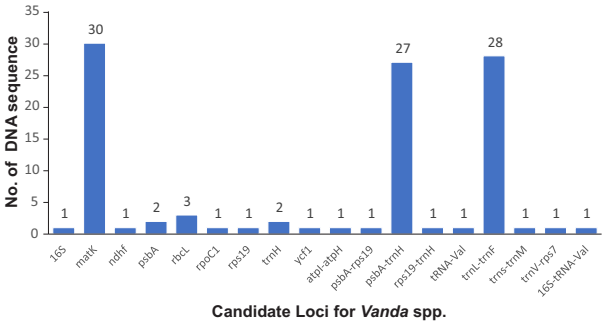

Figure 1. Number of DNA Sequences of the loci located in chloroplast genome of Philippine *Vanda* spp.

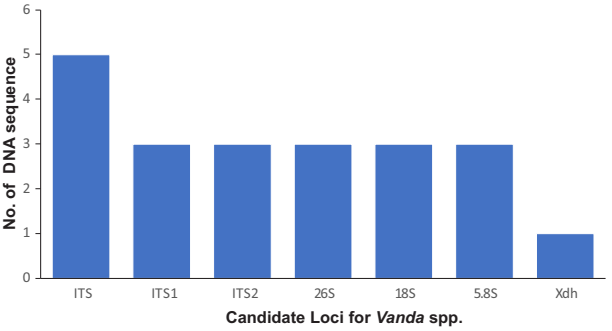

Figure 2. Number of DNA sequences of the loci located in nuclear genome of Philippine *Vanda* spp.

or nuclear genome. The loci with more than four nucleotide sequences underwent further examination. These were *matK*, *psbA-trnH*, and *trnL-trnF* in the chloroplast genome and the assemblage of ITS in the nuclear genome. The total number of nucleotide sequences for the four candidate loci shown is still subject to change. A set of parameters was regarded as necessary to select the best loci for discriminating *Vanda* species, and these were further discussed in the following sections.

Sequence length and alignment capability

One of the factors affecting the quality of the loci during the performance of MSA is the number of barcoded samples. The contrasting numbers of selected accessions and species were presented (Fig. 3). Then, the acquired alignment length of every locus was also analyzed (Fig. 4). Another factor that was examined in this study is the number of indels observed per locus (Fig. 5).

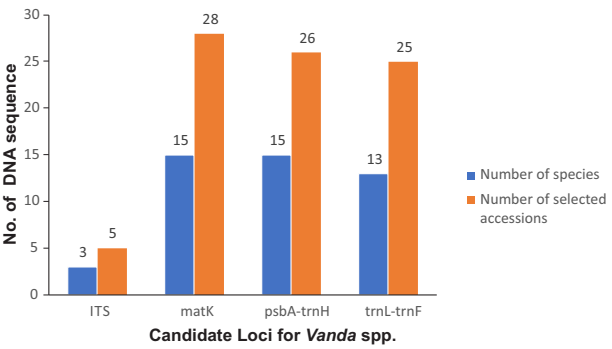

Figure 3. Number of species and selected accessions per locus of Philippine *Vanda* spp.

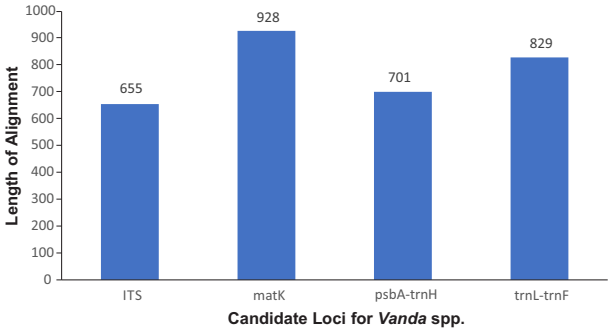

Figure 4. Alignment length per locus of Philippine *Vanda* spp.

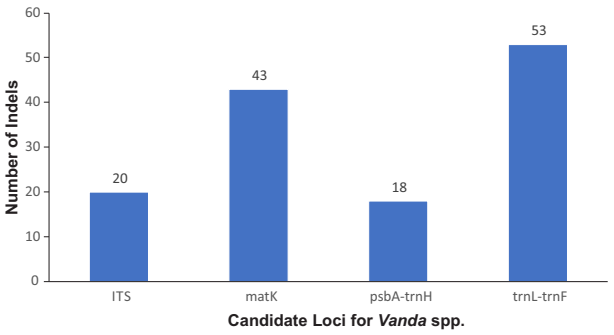

Figure 5. Indels per locus Philippine *Vanda* spp.

One assemblage of ITS1-5.8S-ITS2 and four assemblages of 18S-ITS1-5.8S-ITS2-28S, ranging from 654 bp to 937 bp, were analyzed in ITS. Sites 1 to 155 and 811 to 938 were deleted because these were not properly aligned. The inferred alignment length after MSA was 655 bp, with a total of 20 indels. Recent studies have recommended using of the ITS2 portion rather than the full-length ITS because the latter is more prone to complications when performing the required DNA barcoding protocols (Hollingsworth et al. 2011). However, an independent assessment for ITS2 was infeasible because of the locus's presence in every DNA sequence of Philippine *Vanda* spp. was sparse. In addition, there is still a need to evaluate if the reduction of characters could affect the ability of a locus to discriminate species (Hollingsworth et al. 2011).

Out of 30 sequences collected in *matK*, two nucleotide sequences were excluded preceding the MSA. The accession KP772698.1 contains N in sites 14 and 15 and it denotes an ambiguous nucleotide base that can either be A, G, C, or T (Tamura n.d.). Meanwhile, FR832764.1 carried M in site 53, which stands for either A or C nucleotide base (Tamura n.d.). Retaining these sequences may influence the accuracy of the interpretation because of their ambiguity. Since the length of the sequences was around 952 bp to 1682 bp and a substantial amount of unwanted nucleotide bases were observed, sites 1 to 422 and sites 1351 to 1682 were removed. Moreover, the length of the aligned sequences was 928 bp, and there were 43 indels found. Longer sequences of *matK* are usually suggested as they tend to improve species discrimination over shorter ones (Vu et al. 2018).

In the *psbA-trnH* locus, KC985290.1 was eliminated from the 27 obtained nucleotide sequences due to poor quality as it comprises N in site 725. The shortest sequence length in the accumulated nucleotide sequences was 712 bp, while the longest was 769 bp. Following the MSA, sites 1 to 22 and sites 724 to 791 were deleted because these were not entirely aligned. The sequence length after the alignment was 701 bp, and it only resulted in 18 indels.

The sequence length of *trnL-trnF* was reduced to 829 bp after removing sites 1 to 156, sites 185 to 418, and sites 1221 to 1348. Initially, it only spans from 972 bp to 1227 bp. GenBank accession EF670422.1 was omitted as it comprises N in site 54, M in site 253, and R in site 254. R indicates it can be a purine, A or G (Tamura n.d.). A similar character was also seen in site 159 of KC244663.1; thus – it was also excluded. Many ambiguous characters in sites 313, 320, 345, 348, 350, 354, and 374 were found in KC985396.1. In addition, Y was also seen in sites 840 and 972, representing pyrimidines like C or T (Tamura n.d.). There were 53 indels spotted in the 25 aligned nucleotide sequences.

Among the four loci, ITS and *psbA-trnH* met the criteria for the standard length required in DNA barcoding. In the assemblage of ITS, 18S and 26S are positioned at both ends, while 5.8S is in the middle (Vu et al. 2017). These highly conserved regions can provide better information for deriving relationships among similar sequences (Vu et al. 2017). It is also considered to be one of the loci that is complicated to align, perhaps due to a high number of indels or high variation in length (Vu et al. 2018). Meanwhile, the collected nucleotide sequences for *psbA-trnH* have a length approximately close to each other. It may have influenced the low number of indels obtained upon the alignment as compared to previous studies.

Longer sequences with high variation in length tend to be more challenging to align, as explicitly recognized in *matK* and *trnL-trnF*. Additionally, both loci acquired the highest number of indels. In this study, coding regions in the chloroplast genome like *matK*, are as challenging to align as the intergenic spacers in the chloroplast genome, in this case, *trnL-trnF*. It is probably because of the polynucleotide repeats in its sequences, aside from having many indels (Vu et al. 2018). Despite the complexity during the alignment, all four loci were aligned successfully. The characters within each locus's nucleotide sequences were useful in analyzing the nucleotide polymorphisms that could also help reveal the species resolution of a certain locus.

Sequence variation

Parsimony-informative sites have at least two nucleotides occurring twice (Tamura et al. 2021). In contrast, one type of nucleotide occurs multiple times in a singleton site. Sequences that are not lower than three and consist of unambiguous nucleotides are automatically recognized as singleton sites when using MEGA software (Tamura et al. 2021). The parsimony and singleton rate obtained for each locus were also presented (Fig. 6).

In this study, the chloroplast coding gene and intergenic spacers contained more parsimony-informative sites than the nuclear gene. Maturase k *(matK)* has a parsimony

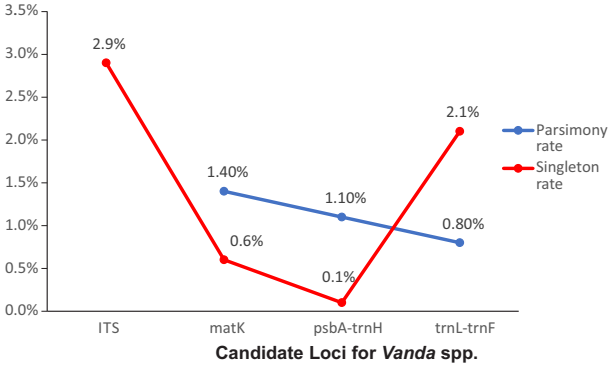

Figure 6. Parsimony rate and singleton rate per locus of Philippine *Vanda* spp.

rate of 1.4%, higher than *psbA-trnH* and *trnL-trnF* with 1.1% and 0.8%, respectively.

Meanwhile, the parsimony-informative site was absent in the aligned sequences of the ITS assemblage. On the other hand, the entire length of ITS has a higher singleton rate followed by *trnL-trnF*, *matK*, and *psbA-trnH*. The values obtained were 2.9%, 2.1%, 0.6%, and 0.1%, respectively. The conserved site, commonly called the constant site, carries identical nucleotides in all the existing sequences (Tamura et al. 2021). It is recognized by MEGA 11 if a minimum of two sequences are found to have unambiguous nucleotides. Conversely, the variable site is the opposite of the conserved or constant site. It comprises at least two types of nucleotides, which can either be parsimony-informative or singleton (Tamura et al. 2021). The conserved and variable rate acquired in this study is also shown (Fig. 7). The locus *psbA-trnH* was significantly conserved compared to the other loci as it has a rate of 98.7%. It was followed by *matK* (98%), ITS (96.9%), and *trnL-trnF* (95.9%). On the other hand, the locus with the least variable rate was *psbA-trnH* (1.3%). In contrast, the two loci with the gratest variable rate were ITS and *trnL-trnF* (2.9%). The variable rate of *matK* was 2%, which lies between the three loci.

The relationship between the conserved rate and variable rate was also emphasized (Fig. 7). The parsimony and singleton rates were no longer focused on since the combination of both also equals the variable rate. It was observed that the higher the conserved rate, the lower the variable rate and vice versa. The conserved rate depicts how similar the sequences of different species are and how

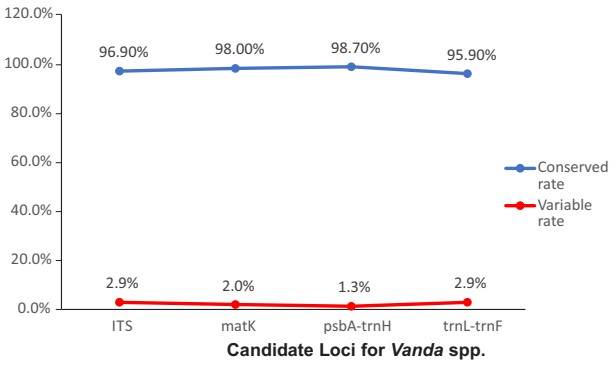

Figure 7. Conserved rate and variable rate per locus of Philippine *Vanda* spp.

relatively unchanged they are after several generations. It is also a major primer design quality (Vu et al. 2020). On the contrary, variable rates indicate how distinct the sequences of different species are. The unique character contained within the variable sites can distinguish species from one another. It is especially important to examine the species-specific SNP approach (Vu et al. 2020).

Tree-based species identification

The tree-based identification revealed that *V. tricolor* and *V. luzonica* formed a paraphyletic group (Fig. 8). The branching between EF670375.1 and EF670374.1 showed they were relatively similar. However, their bootstrap value was not supported, as it only has 47% support. Even though AY278111.1 is of different species, it has a well-supported bootstrap value of 87% with the previously mentioned accessions of *V. tricolor*. It revealed that there could be uncertainty in their relationship. A related problem may have occurred with EF670373.1 as no bootstrap value could support its relationship with *V. luzonica* and most notably with the other species of *V. tricolor*. Accession No. KC244655.1 was not supported by bootstrap value, which could also signify that the number of accessions for the species was too low. Despite these circumstances, it was still separated into a single monophyletic branch.

Several clades in the generated phylogenetic tree for the *matK* locus predominantly received weakly supported and not supported bootstrap values. Multiple accessions of different species were primarily clustered in monophyletic branches, except for one variations of *V. lamellata*, *V. limbata*, and *V. miniata*, along with *A. miniatum*, as each of them produced monophyletic branches separately. In the study by Gardiner et al. (2013), *V. tricolor* and *V. ustii* were also grouped with 63% support. The researchers proposed that both species belong to the section *Deltoglossa* as they have similar morphological features like 'cylindrical column with thickened base'. The grouping of *V. merrillii* and *V. luzonica* into a monophyletic branch could also be associated with the resemblance in their morphology that was shared among the species recommended to be categorized in section *Deltoglossa*. *V. lindenii* and *V. furva* were also clustered in one monophyletic branch. Both species were observed to have appendages on the midlobe of its labellum that was common to the species in section *Dactylolobata* (Gardiner et al. 2013). Lastly, most of the species of *V. lamellata* also formed

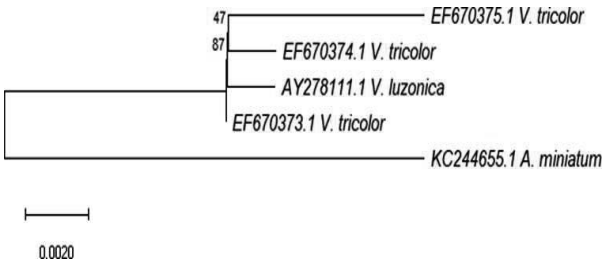

Figure 8. Neighbor-joining (NJ) tree of ITS. The NJ tree was inferred from ITS gene sequences using MEGA 11 with a 1,000 bootstrap replicates.

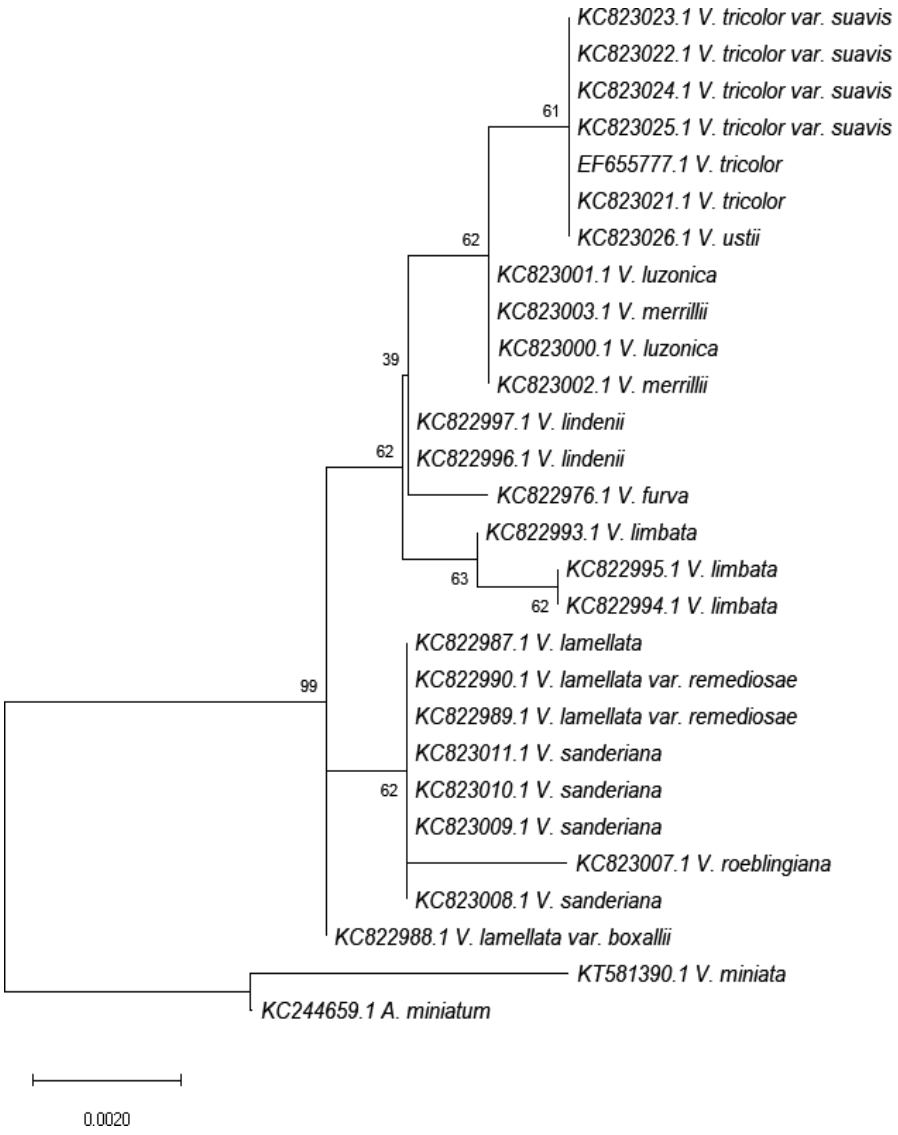

Figure 9. Neighbor-joining (NJ) tree of *matK*. The NJ tree was inferred from *matK* gene sequences using MEGA 11 with a 1,000 bootstrap replicates.

a monophyletic branch with *V. roeblingiana* and *V. sanderiana*. It could be due to their distinctive morphological characteristics, for instance, 'cylindrical column without thickened base' and 'usually modified or embellished labellum midlobe' as they were suggested to be classified in section *Roeblingiana* (Gardiner et al. 2013). However, *V. lamellata* var. *boxallii* was recognized from the other species of *V. lamellata* since it was clustered in a separate monophyletic branch.

The NJ tree inferred for *psbA-trnH* exhibited unresolved relationships for some species. Despite being different from each other, *V. merrillii*, *V. tricolor*, and *V. ustii* were clustered in a monophyletic branch. They were some of the species proposed to be categorized under section *Deltoglossa* (Gardiner et al. 2013). It has a 46% bootstrap value that is equivalent to not supported. Including *V. lindenii* might have affected it since they do not share a certain morphological character. It was suggested to be placed in the section *Dactylolobata* (Gardiner et al. 2013). Another identification failure was observed in *V. lamellata* and *V. luzonica* as they were split into polyphyly. However, it was noticeable that *V. lamellata* var. *remediosae* formed a distinct clade. Consequently, that variety could be considered identifiable. *V. roeblingiana*, *A. miniatum*, and *V. miniata* were also branched into a separate clade. Although *A. miniatum* is a homotypic synonym of *V. miniata*, they were individually identified using the *psbA-trnH* locus. Meanwhile, the species positioned in a paraphyletic group was *V. furva*. As presented in the tree, the node with an 81% bootstrap value implied that the few species recommended to be classified in section *Roeblingiana* were well-supported (Fig. 10). The other species that were grouped in a monophyletic branch that was successfully identified were *V. limbata* and *V. sanderiana*.

The clustering of species in the phylogenetic tree generated for *trnL-trnF* was intricate as well (Fig. 11). As mentioned earlier in *matK*, *V. furva* and *V. lindenii* might have been grouped in one monophyletic branch because they have common characteristics that are unique only to section *Dactylolobata*. Moreover, the species suggested to be classified under section *Deltoglassa*, including *V. limbata*, *V. luzonica*, *V. tricolor*, and *V. ustii*, were clustered in a monophyletic branch even if they were of varying species. As analyzed in the phylogenetic tree, *V. merillii*

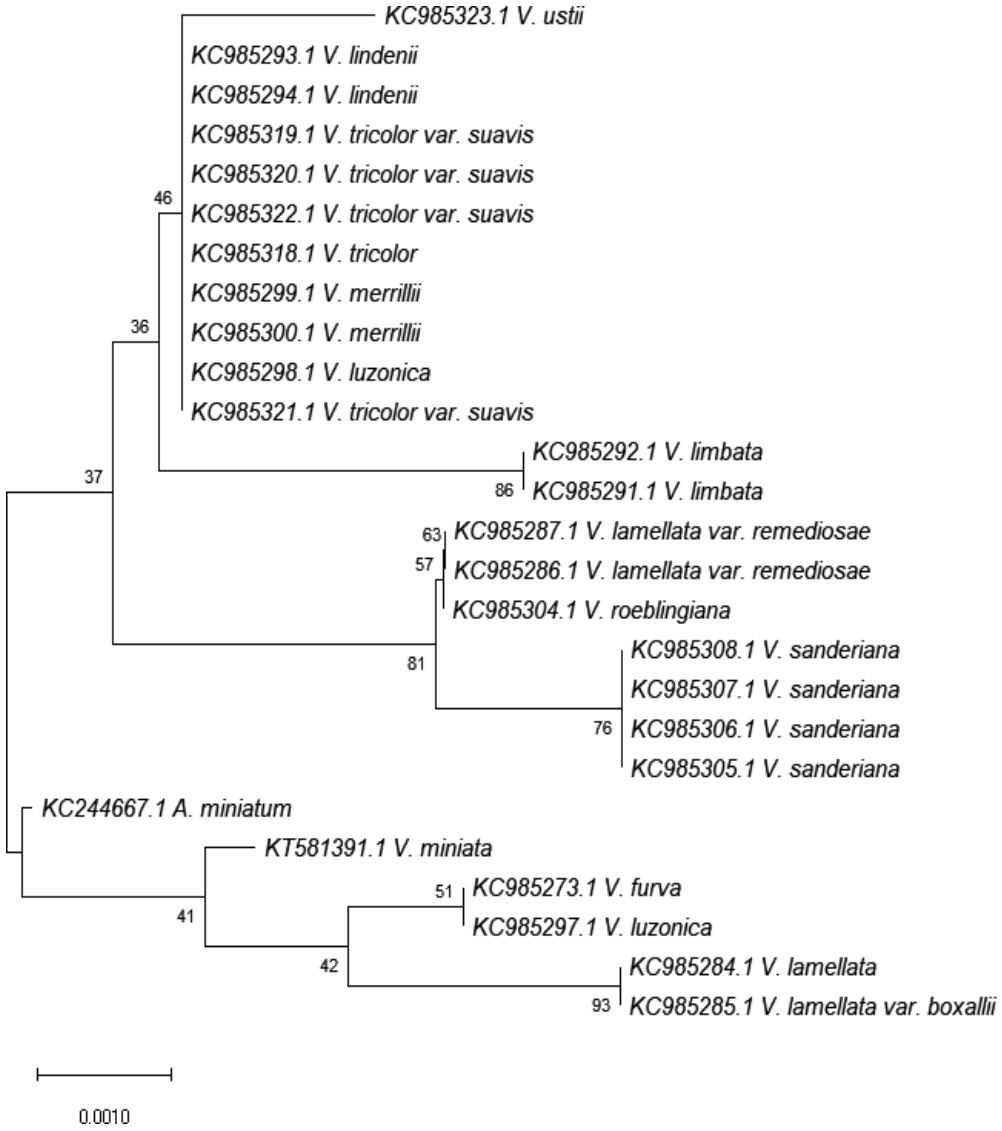

Figure 10. Neighbor-joining (NJ) tree of *psbA-trnH*. The NJ tree was inferred from *psbA-trnH* gene sequences using MEGA 11 with a 1,000 bootstrap replicates.

exhibits a polyphyletic group. *V. lamellata* var. *remediosae* and *V. sanderiana* were individually identified since they formed separate monophyletic branches. They acquired weakly supported bootstrap values, 62% and 69%, respectively. On the contrary, *V. roeblingiana* and the other species of *V. lamellata* remained unidentified in the monophyletic branch. It supports the unpublished data for the proposed species to be placed in the section *Roeblingiana* as they share a certain morphological character (Gardiner et al. 2013).

Indel-based species identification

The application of indel information is proposed to help identify distinctions among species that a phylogenetic tree sometimes fails to exhibit (Vu et al. 2020). Upon examining the indels within ITS, the three clones of *V. tricolor* share deletion in site 90 and insertion in site 636 with *A. miniatum*. The distinct insertions in sites 9, 63, 121, 420, 437, 446, 485, 490, 512, 604, 612, 637, and 640 made *A. miniatum* distinguishable from *V. tricolor*. The difficulty in identifying *V. tricolor* was possible because there were no certain homologous characters found within the clones. Meanwhile, *V. luzonica* was easily distinguished from both species because it has no deletion in site 90 and insertion in site 636.

In *matK*, *V. tricolor* and *V. ustii* were also indistinguishable using their indel information as both were seen to have an insertion in sites 635 and 778. Similarly, *V. luzonica* and *V. merrillii* also shared insertion in site 778. *V. furva* was easily recognized from the rest of the species since it does not have an insertion in site 697. On the other hand, there were no particular indels that could singly differentiate *V. lindenii*. Another species that was successfully identified was *V. limbate*, which carries a unique insertion in site 922. In this locus, *V. miniata* has an identical insertion in sites 16, 127, 175, 182, 378, 462, and 591 with *A. miniatum*. It was accepted as a correct identification since they are homotypic synonyms. However, *V. miniata* has its insertion at sites 919, 920, 922, and 927. *V. lamellata*, *V. roeblingiana*, and *V. sanderiana* were still inseparable because they have a similar insertion in site 787. However, *V. roeblingiana* was found to have

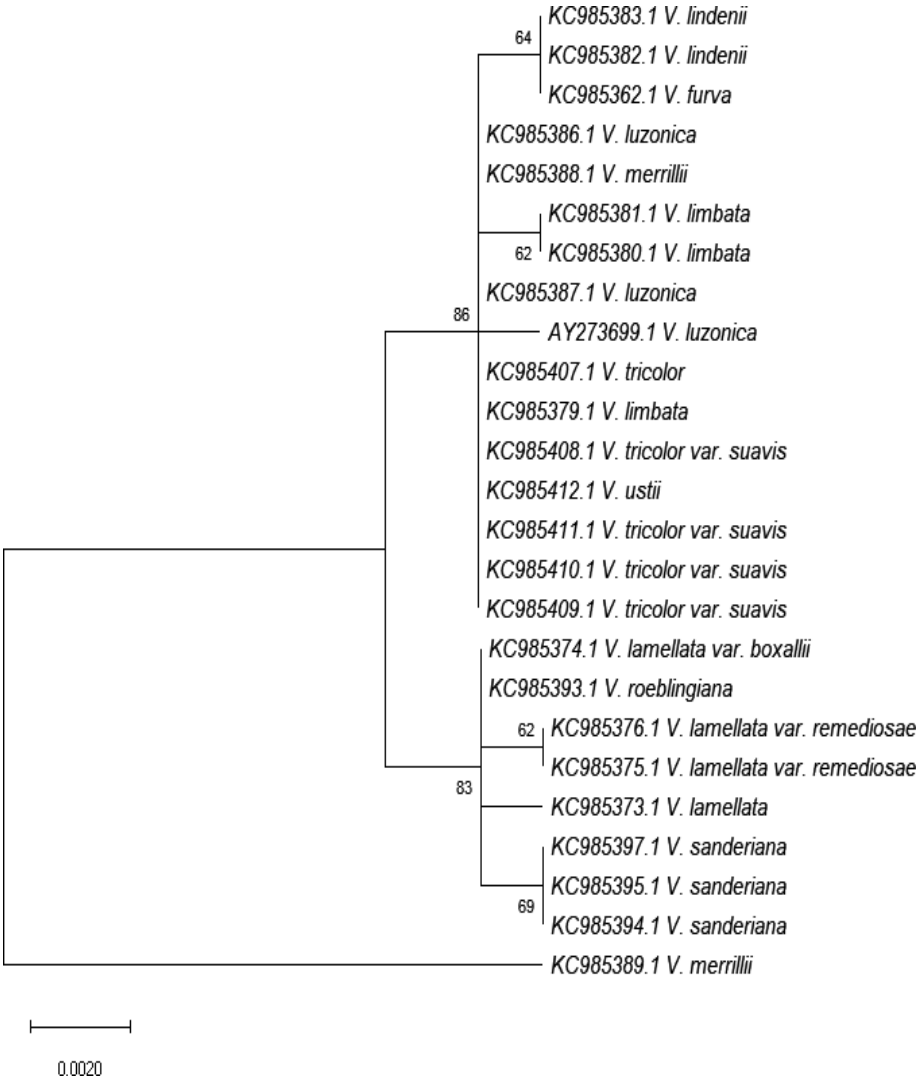

Figure 11. Neighbor-joining (NJ) tree of *trnL-trnF*. The NJ tree was inferred from *trnL-trnF* gene sequences using MEGA 11 with a 1,000 bootstrap replicates.

distinct insertion in sites 183 and 855, which helped it be recognized against the two species. It was also observed that it has a high species divergence in the monophyletic branch where it was clustered.

Indels in the aligned sequences of *psbA-trnH* provide better species resolution than the result of its ambiguous phylogenetic tree. The identification of *V. limbata* and *V. sanderiana* was consistent even with their indel information. *V. limbata* has an insertion in site 286, whereas *V. sanderiana* has distinct insertion in site 6. Four accessions of *V. lamellata* were determined as they have insertions in sites 622 to 631. In contrast, the rest of the species have deletions in that particular site. In addition, *V. lamellata* var. *remediosae* contained distinct insertion in site 631. It can be used to distinguish it from other species or variations of *V. lamellata*. *V. ustii* was also distinguished as it is comprised of distinctive insertion in site 329. A deletion in site 6 was observed in *A. miniatum* as well. Meanwhile, the remaining eight species did not contain sufficient indels to support individual identification.

The least homologous characters were observed to differentiate one species from another in the *trnL-trnF* locus. *V. lamellata*, *V. roeblingiana*, and *V. sanderiana* were indistinguishable as they all have an insertion in sites 37, 92, and 635. However, *V. lamellata* was recognizable since it has an insertion in site 345, as well as *V. lamellata* var. *remediosae* in site 811. *V. sanderiana* was also determined as an individual because its multiple accessions have homologous character in site 804. Only two out of three species were identified in *V. limbate*. At the same time, only one out of two was distinguished in *V. merrillii* upon checking their indels. Hence, it was still an identification failure. *V. tricolor* and *V. ustii* share insertion in sites 495 to 503, making them difficult to determine from one another. Lastly, *V. lindenii* was the only species with insertion in sites 226 to 230.

The species resolution of each locus was presented (Fig. 12). Overall, ITS gained the highest species resolution of 66.67%. It is followed by *psbA-trnH* with a species resolution of 60%. *matK* and *trnL-trnF* accumulated the least species resolution of 40% and 30.77%, respectively. Indel-based species identification gave a superior result compared to tree-based species identification. The clustering of species in the inferred phylogenetic tree for each locus was observed to be more grounded on their morphological characteristics.

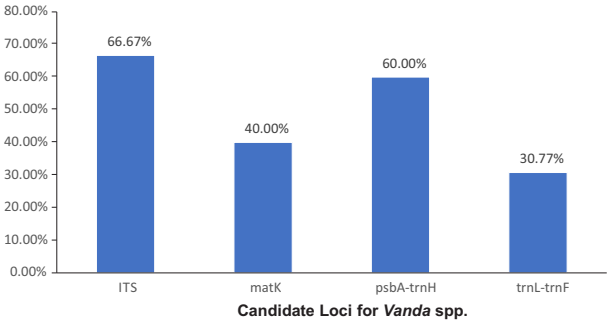

Figure 12. Species resolution per locus of Philippine *Vanda* spp.

This supports the previous literature by Vu et al. (2020) stating that indel information can be used to identify further interspecific and intraspecific dissimilarity that generated phylogenetic trees likely miss. Meanwhile, the insertions and deletions found in the aligned sequences could emphasize which sites a certain species was distinct from.

Conclusion

The present study revealed that ITS met the standard length for DNA barcodes and obtained the highest species resolution among the four loci. However, it should be noted that due to the low number of sequences that were examined in ITS, *psbA-trnH* was also considered to be one of the potential molecular markers for identifying *Vanda* spp. using in silico analysis. Current molecular and bioinformatics techniques, such as next-generation sequencing (NGS), can also be used as an in-silico approach to accurately identify *Vanda* spp. in the future. To validate the study results, it is recommended to do an actual sampling of *Vanda* spp. in the Philippines. Furthermore, a combination of morphological and molecular data would be recommended for a more comprehensive taxonomic and phylogenetic information on the *Vanda* species in the Philippines.

Acknowledgment

The authors would like to express their deepest gratitude to their alma mater, and the people who helped during the progress and publication of the manuscript. Special thanks to Dr. Joeselle M. Serrana of University of Ottawa in Canada and Dr. Mark Angelo O. Balendres of De La Salle University (DLSU) Manila for their valuable help during the content review of the manuscript.

References

Chatzou, M., Magis, C., Chang, J. M., Kemena, C., Bussotti, G., Erb, I. & Notredame, C. 2016. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics* 17(6): 1009–1023. https://doi.org/10.1093/bib/bbv099

DENR Administrative Order 2017. Updated national list of threatened Philippine plants and their categories. (DAO 2017-11). https://bmb.gov.ph/index.php/facts-and-figures-wild/national-list-of-threatened-flora. Access date: February 2022.

Dizon, S. A., Ocenar, A. P. & Naïve, M. A. K. 2018. Inventory of orchids in the Mount Hamiguitan Range Wildlife Sanctuary, Davao Oriental, Philippines. *Bio Bulletin* 4(1): 37–42.

Edgar, R. C. 2004a. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5): 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C. 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics* 5: 113. https://doi.org/10.1186/1471-2105-5-113

Gardiner, L. M., Kocyan, A., Motes, M., Roberts, D. L. & Emerson, B. C. 2013. Molecular phylogenetics of *Vanda* and related genera (*Orchidaceae*). *Botanical Journal of the Linnean Society* 173(4): 549–572. https://doi.org/10.1111/BOJ.12102

Hall, B. G. 2013. Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution* 30(5): 1229–1235. https://doi.org/10.1093/molbev/mst012

Hollingsworth, P. M., Graham, S. W. & Little, D. P. 2011. Choosing and using a plant DNA barcode. *PLOS One* 6(5): e19254. https://doi.org/10.1371/journal.pone.0019254

Kim, H. M., Oh, S. H., Bhandari, G. S., Kim, C. S. & Park, C. W. 2013. DNA barcoding of *Orchidaceae* in Korea. *Molecular Ecology Resources* 14(3): 499–507. https://doi.org/10.1111/1755-0998.12207

Kinene, T., Wainaina, J., Maina, S. & Boykin, L. M. 2016. Rooting trees, methods for. *Encyclopedia of Evolutionary Biology* 13: 489–493. https://doi.org/10.1016/B978-0-12-800049-6.00215-8

Lahaye, R., Van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G. & Savolainen, V. 2008. DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences* 105(8): 2923–2928. https://doi.org/10.1073/pnas.0709936105

Nguyen, N. H., Vu, H. T., Le, N. D., Nguyen, T. D., Duong, H. X. & Tran, H. D. 2020. Molecular identification and evaluation of the genetic diversity of dendrobium species collected in southern Vietnam. *Biology* 9(4): 76. https://doi.org/10.3390/biology9040076

Panal, C. L. T., Opiso, J. G. & Opiso, G. 2015. Conservation status of the Family *Orchidaceae* in Mt. Sinaka, Arakan, North Cotabato, Philippines. *Biodiversitas Journal of Biological Diversity* 16: 213–224. https://doi.org/10.13057/biodiv/d160217

Sanyal, G., Mahadani, A. K., Mahadani, P. & Bhattacharjee, P. 2015. Insertion-deletion as informative characters in DNA barcoding. *International Journal of Multimedia and Ubiquitous Engineering* 10(10): 67–74. https://doi.org/10.14257/ijmue.2015.10.10.07

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (n.d.). Molecular Evolutionary Genetics Analysis version 4. http://cda.psych.uiuc.edu/406_407_material/mega4.pdf. Access date: January 2022.

Tamura, K., Stecher, G. & Kumar, S. 2021. MEGA 11: Molecular Evolutionary Genetics Analysis version 11. *Molecular Biology and Evolution* 38(7): 3022–3027. https://doi.org/10.1093/molbev/msab120

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M. & Dessimoz, C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology* 64(5): 778–791. https://doi.org/10.1093/sysbio/syv033

Tanee, T., Chadmuk, P., Sudmoon, R., Chaveerach, A. & Noikotr, K. 2012. Genetic analysis for identification, genomic template stability in hybrids and barcodes of the *Vanda* species (*Orchidaceae*) of Thailand. *African Journal of Biotechnology* 11(55): 11772–11781. https://doi.org/10.5897/AJB11.3992

Vu, T. H. T, Le, T. L., Nguyen, T. K., Tran, D. D. & Tran, H. D. 2017. Review on molecular markers for identification of Orchids. *Vietnam Journal of Science, Technology and Engineering* 59(2): 62–75. https://doi.org/10.31276/VJSTE.59(2).62

Vu, H. T., Huynh, P., Tran, H. D. & Le, L. 2018. In silico study on molecular sequences for identification of *Paphiopedilum* species. *Evolutionary Bioinformatics*. https://doi.org/10.1177/1176934318774542

Vu, H. T., Vu, Q. L., Nguyen, T. D., Tran, N., Nguyen, T. C., Luu, P. N. & Le, L. 2020. Genetic diversity and identification of Vietnamese *Paphiopedilum* species using DNA Sequences. *Biology* 9(1): 9. https://doi.org/10.3390/biology9010009