Review Article

Bioinformatics analysis on DNA barcode sequences for species identification: a review

ABSTRACT:

Classification of organisms is the primary step in management of biodiversity, breeding, conservation and development of populations and distinguishing adulterant objects. There are many approaches in taxonomic identification, from morphological, PCR-based to sequence-based techniques. Molecular methods give more accurate results than morphological comparisons and are independent of plant stages. PCR-based methods are low-cost but their limited information gives less reproducibility and can only distinguish samples among determined groups. In contrast, in sequence-based methods each nucleotide site is considered as genetic information hence a sequence of nucleotide represents large data, which is highly specific and more stable than PCR bands. Establishment of worldwide DNA library for barcoding is essential. There were previous reviews on screenings and applications of barcodes in different taxa. In this review we discussed common bioinformatics analyses as well as some new improved techniques relying on barcoding approaches.

Key words: molecular classification, bioinformatics tools, DNA barcoding, sequence analysis, identification technique

1. INTRODUCTION

Collection of genetic information, for looking up the origin of a wide range of organisms linked all over the world, is an advanced and essential idea for the protection of species, phylogenetic inference, management and development of genetic diversity [1, 2]. Morphological methods show limitations of accuracy and high reliance on reproductive organs. PCR-based methods can overcome these above problems with just a small piece of sample. However, amplification techniques can only be effectively applied to samples of a defined group using RADP, RFLPs, AFLPs [3] or to samples containing specific genes using species-specific PCR [4-6]. An unknown taxon can not be determined using PCR-band techniques.

Since each site of 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. Specific and stable features of monomers are useful in evaluating genetic relationship of a new query sample based on the available sequence library and thus allow to consult origin of the unknown homologous taxon. DNA polymorphism may even provide more information than proteins due to the degradation of genetic code and the presence of large non-coding stretches [7]. DNA fragments represented for the organism can be used as an identifying sequence like the human fingerprint (Figure 1). This is the most common method in molecular identification strategy today. In animal, the highly conserved sequence of mitochondrial cytochrome c oxidase subunit I (COI), which relates to oxidative phosphorylation for metabolism, was effectively used as barcode in diverse animal species [8, 9]. Nevertheless there was no such effective barcode for plants.

A numerous studies of barcoding for plants have been conducted. To date, sequencing techniques for determining species using next generation sequencing (NGS) in plants are limited primarily to agricultural crops [10-13] due to their high cost and time. Therefore short sequences are still considered as convenient and effective toosl due to the quick and accurate sequencing [14]. A number of studies have reviewed on finding and applying of DNA barcoding in which the efficiency of different biomarkers have been summarized [15-22]. Others discussed the effectiveness of different barcoding techniques, criteria and measurements [7, 23, 24]. In this review, we discussed some bioinformatics tools in identification analysis of barcoding studies. We also presented some prospects and developed techniques relied on barcoding approaches.

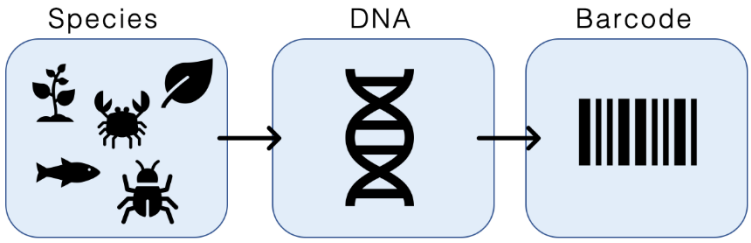

Figure 1. DNA barcoding scheme (https://en.wikipedia.org/wiki/DNA_barcoding)

2. COMMON BIOINFORMATICS ANALYSES IN IDENTIFICATION TECHNIQUES USING SEQUENCES

Bioinformatics procedure for species identification using sequences comprises two basic steps: First, the sequence alignment should be conducted on the basis of a comparative step. This alignment can be performed by two methods. The novel sequences are pairwise aligned against the available sequences in certain databases (similarity search) [25, 26] or studied sequences are aligned against each other in a specific set of data (many-against-each other search) [2, 27, 28]. Following, based on these alignment data, similarity or variation are investigated. This step is performed by evaluating such parameters as GC% content [27, 29], genetic distance [30-33], variable sites [32, 34], indel appearances [35], or monophyletic clusters [36-38], which can indicate typical characters of the examined sequences. These measurements vary from different studies as they depend on alignment and identification methods (Table 1). Final target of this step is to decide whether the species are different or not. In some studies, species resolution was calculated by counting the number of identified species out of the total number of examined species [2, 31, 39].

Table 1. List of common sequence-based identification methods and identification criteria

Alignment Methods	Species Identification Methods		Identification Criteria
Similarity search	BLAST		Correct Identification Ambiguous Identification Incorrect Identification
Many-against-each other search	Genetic distance-based	Nearest distance	Correct Identification Ambiguous Identification Incorrect Identification
Many-against-each other search		Best match	Correct (match) Ambiguous Incorrect (mismatch)
		Best close match	Correct (match) Ambiguous Incorrect (mismatch) Nomatch (under Threshold (%)
		All species barcodes	Correct (match) Ambiguous Incorrect (mismatch)
Many-against-each other search		Barcoding gap	Intra-specific genetic distance Inter-specific genetic distance Barcoding gap
Many-against-each other search	Tree-based		Monophylectic Paraphylectic Polyphylectic
Many-against-each other search	Character-based (nucleotide polymorphism)		Variable sites Indels Sequence lengths GC% contents

In the alignment step, sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). For BLAST (Basic Local Alignment Search Tool)

(https://blast.ncbi.nlm.nih.gov/Blast.cgi) and FASTA [40] programs, each examined sequence is considered a query sequence. A local pairwise alignment is running between two biological sequences: the query sequence and each of database sequences . In contrast to similarity search, alignment process in many-against-each other search is based on two stages, the pairwise alignments followed by multiple alignments. In multiple alignment, whether global alignment or local alignment should be applied depending on similarity level and difference in sequence lengths. ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences [41]. However, for MAFFT [42] and MUSCLE [43] programs, users can select suitable aligning strategies. Global alignment is effective when the input sequences share global homology, and the similarity level is high. On the other hand, with fragmentary and divergent sequences, local alignment would be the better option [44].

For "BLAST" method, the target is searching for the best homologous sequences (best hits) from the available databases (GenBank, BOLD, others). Based on this background, "Correct Identification" means that: the best hit is the sequence with species name as expected. "Ambiguous Identification": the best hits are some sequences belong to different species including the expected species. "Incorrect Identification": the best hit does not match with expected species [25, 26, 28, 45]. However, because "BLAST" depends on available sequences reported in databases, query species must be already included in a database otherwise the result will be a failure, therefore a negative "Incorrect Identification" will occur.

For genetic distance-based methods, the "best hit" feedback of "Correct", "Ambiguous" and "Incorrect" criteria are similar to that in "BLAST", but based on the smallest genetic distances ("Nearest Distance") [28] or the most similarity ("Best match", "Best close match") [46-49]. The query sequence is compared with each other in the given data set. "Best close match" differs from "Best match". A similarity threshold value (e.g. 95%) of all intraspecific distances is established to determine how similar a barcode match needs to be and the results under this similar value (No match) would be removed before identification step.

For "All species barcodes" methods, a list of sequences sorted by similarity to each query using the same threshold as for "Best close match" are assembled. If the query is followed by all conspecific barcodes with at least two ones then the species is identified. If the query is followed by only one or some of conspecific barcodes then the identification is ambiguous. If the query is followed by none of conspecific barcodes but other species then misidentification occurs [47].

Barcoding gap method analyses the divergence between intra-specific and inter-specific genetic distances of each query *versus* other conspecific and hetero-specific sequences in a data set [25-28, 30-33, 50, 51]. However this method may be not precise in case of using mean instead of smallest inter-specific distances versus largest intra-specific distances [52]. If barcoding gap exists, the species is successfully identified.

Another method is the use of nucleotide polymorphism features, such as variable sites, sequence length variation, indel information, GC% content [4, 19, 32, 53, 54] as tools of identification. This approach is known as character-based method in which each nucleotide is consider as the fifth character beside four traditional characters A, T, C and G.

Regardless of which method is used, the core property of the identification strategy is that all conspecific sequences should be grouped together without blending with any other species. To address this issue, the tree-based method is a simple and visualized approach that is most common in such classification studies [2, 25-27, 29, 34, 36-39, 45, 46, 48, 49, 55]. Two species are completely separated when all sequences of one species are clustered in a monophyletic branch [56]. There are some problems that should be noticed to avoid errors in identification process. Small sample size [4, 26, 29, 36, 51, 57-59] or a wide range but less con-level of taxonomic samples [27, 60, 61] may lead to over-fitting [56].

3. COMMON BIOINFORMATICS TOOLS IN IDENTIFICATION TECHNIQUES USING SEQUENCES

For various measurements, different bioinformatics tools could be used also. The BLAST method is presented by BLASTn program from NCBI. Intra- and inter-specific genetic distances, matching sequences, and clustering sequences based on pairwise distances can be calculated in "Best match", "Best close match" and "All species barcodes" methods by Species Identifier tool using

TaxonDNA software. When multiple regions are compared for selection of optimal biomarkers, TaxonGap program is used to infer intra- and inter-specific distance for “Nearest” method in high-throughput sequencing researches [62].

For tree-based reconstruction, 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. Which method should be used in different studies is still a problem that need to be noticed for obtaining the most accurate results. Some studies performed comparisons on different methods [27, 45]. Although algorithms are inferred from suppositions, having a thorough understanding of our algorithms and data is the best way to achieve the highest efficiency. The standard principal of tree building is to examine all possible topologies or certain topologies that represent the true structure. Neighbor-joining performance [39, 46] is a time-consuming method in reconstructing phylogenetic trees. It finds pairs of operational taxonomic units (OTUs) that minimize the total branch length at each stage of OTUs clustering and starts with a star-like tree [63]. However the reliability of Neighbor-joining tree is a problematic issue [47] as it quickly generates only one phylogenetic tree while others may be more fitting. In contrast, Maximum Likelihood (ML) accounts the probability for all events that can happen simultaneously and the best tree is supported at a higher probability. Hence ML become a powerful and professional method in phylogenetic algorithm [64, 65] although it requires significant running time for optimal tree especially with large data [66, 67].

The Maximum parsimony method is based on the least character state changes required to infer a tree. In case of heterogeneous evolution, maximum parsimony (MP) is strongly biased towards recovering an incorrect tree. However this method outperforms ML over a wide range of conditions, including low and moderate heterogeneity [68].

While both ML and MP use the probabilities called likelihood, the Bayesian (BA) technique represents the posteriori probability (Bayes’ rule). A known theory called prior is implemented. The posteriori is in direct proportion with the product of likelihood and prior [67]. Bayesian posterior probability gives more-generous estimates of subtree reliability than the maximum likelihood analysis, particularly when using the gamma distribution modeling [65, 67]. When the size of data is small, the probabilities inferred from ML may be over-fitting. Maximum a posteriori (MAP) can be taken accounts to solve this problem.

Regarding genetic distance and tree-based methods, MEGA is the most popular software used due to its friendly interface and optimal analysis time. Since genetic variations have to be inferred from evolutionary distance matrices, MEGA versions have integrated these evolutionary models into their program along with different algorithms, i.e. Neighbor-joining, Maximum Likelihood, Maximum Parsimony and UPGMA [69]. However, the number of models in MEGA is pruned for its convenience. For more accurate and reliable results, PAUP* and MRBAYES are often used although it takes much more time. Maximum Likelihood and Maximum Parsimony can also be calculated using PAUP* [70]. Data was estimated for revolutionary models using software jModelTest [71] before running in PAUP*. The software MRBAYES analyses the Bayesian Inference [72]. However, the use of PAUP* and MRBAYES needs bioinformatics skill to run the program, which is quite complex for some researchers.

4. RELATION BETWEEN MOLECULAR IDENTIFICATION AND APPLICATION FOR PHYLOGENETIC STUDY

For discrimination of species, tree-based method seems to be the most common technique in many different studies on variety taxa. Phylogenetic tree can be presented for both phylogenetic and barcoding studies. However phylogenetic analysis represents the measurement and estimation of evolutionary past. Whereas barcoding analysis is used to identify taxa in certain taxonomic groups [2, 39, 48] or to determine a new taxon [60, 73] by DNA sequence comparison. In barcode-phylogenetic tree, the differences between nucleotide characters are more important than the way they form through the evolutionary time. This means that taxa of the same species should be clustered in a monophyletic branch and the different ones should be distributed in separated clades [45, 49]. The length of branches and the members of clusters are not strictly evaluated. Therefore, barcode-phylogenetic tree is not really a phylogenetic tree. However, as the trees are built based on specific DNA sequences, information from them can be used for phylogenetic investigations. Barcoding and phylogenetic relationships of species have also been studied in combination

previously [28, 29, 36, 74]. For phylogenetic relationship analyses, homologous variations at each alignment site are considered. In this case, Maximum Likelihood or Maximum Parsimony is used instead of Neighbor-joining [27].

Yang *et* al. (2013) used six data sets including sequences of whole complete (cp) genomes, protein-coding exons, large single-copy region, small single-copy region, inverted repeat region, introns and spacers for phylogenetic tree establishments to indicate congruent among different data partitions (Figure 2). Only cp genome tree and the intron and intergenic spacer tree gave genetic similarity. The relationships between seven species *C. danzaiensis*, *C. pitardii*, *C. impressinervis*, *C. cuspidata*, *C. taliensis* 7, *C. taliensis* 8 and *C. yunnanensis* were not consistent in other trees [75]. This result posed a question: whether a partial DNA sequence region is sufficient to represent species and the relationship between them. The great data of whole genome might be more reliable for estimating the evolution compared to shorter barcodes.

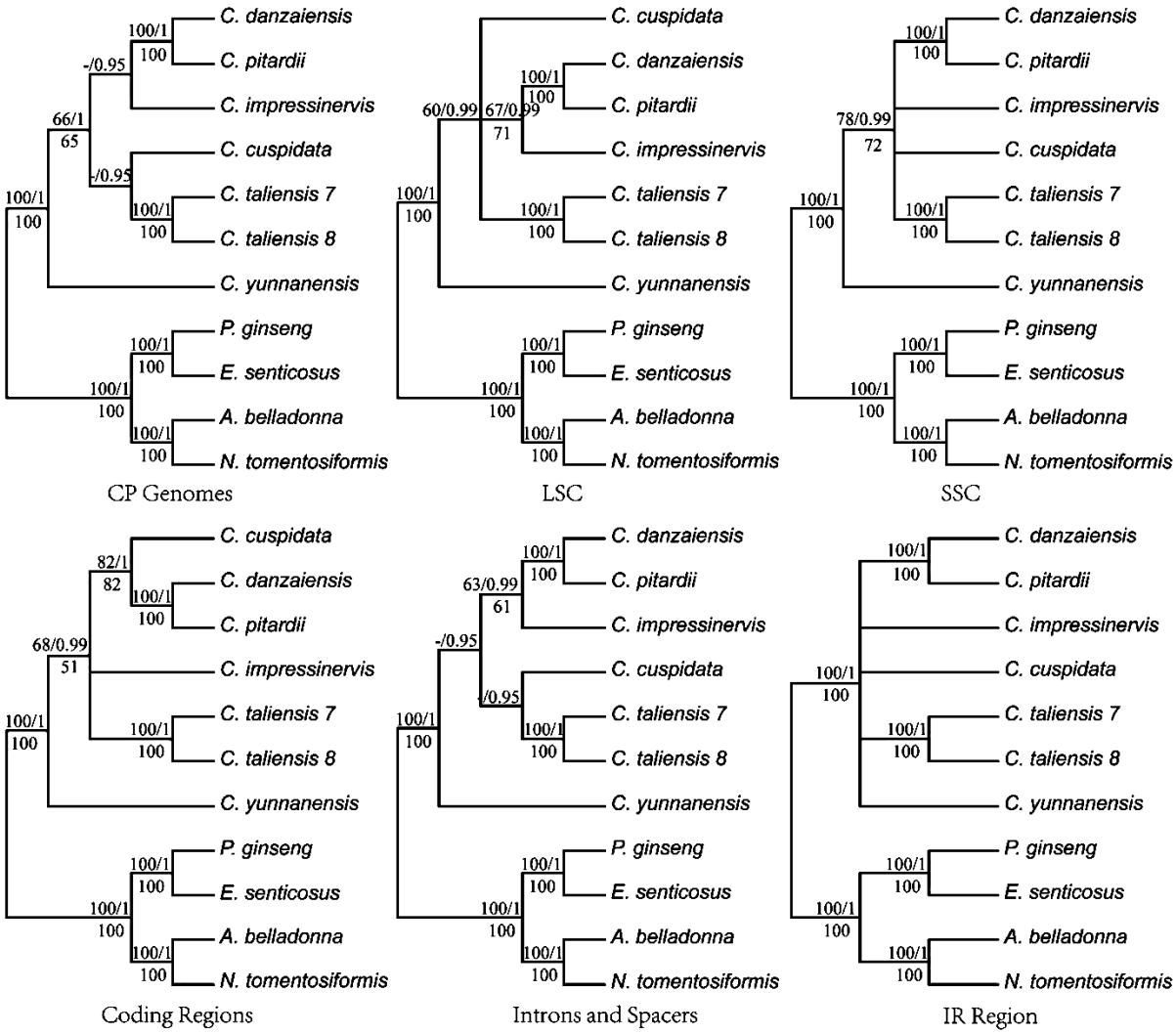

Figure 2. Phylogenetic trees constructed from different data partitions from whole chloroplast genome, with all clades were absolutely separated by high genetic variation [75]

(CP Genome: complete genome; LSC: large single-copy regions; SSC: small single-copy regions; IR Region: inverted repeat region; numbers above the lines on the left indicate the maximum parsimony bootstrap of each clade >50%; numbers above the lines on the right indicate the Bayesian posterior probabilities; numbers below each branch are the maximum likelihood bootstrap of each clade >50%)

5. THE USE OF COMPLETE CHLOROPLAST GENOME AS ULTRA BARCODE (SUPER-BARCODE)

Short DNA sequences can solve most questions in identification at species and above-species levels but still have some limitations with closely related taxa. Some researchers have suggested

the way of serving complete chloroplast genome as a single barcode in plants [31, 76]. This method was called ultra-barcoding by Kane *et* al. when they compared nine complete plastid genotypes of *Theobroma cacao* and one related species *T. grandiflorum* with a control GenBank accession [77]. The complete chloroplast DNA (cpDNA) could absolutely separate all examined intra-species in the study. This approach promise new effective applications in identification at species and below-species level [78]. *Fritillaria*, a popular herbal medicine genus in China, experienced some efforts in characterizing closely related species using universal molecular markers (ITS, *trn*L-*trn*F ect.) but could not be distinguished entirely [79-81]. Using complete chloroplast sequences, phylogenetic analyses of eight *Fritillaria* species were well resolved in the study of Bi *et* al. [82]. In other studies, even no potential regions in cp genome were proposed but the complete genome itself had a capability in distinguishing samples as a single barcode [83]. The interesting thing is that you can also use this meta-data to develop potential mini-barcodes which are high variability for quick authentication of certain taxa [84-88].

This genomic strategy not only allows the use of complete cpDNA as a single barcode, but also facilitate the utilization of other data partitions to identify plants. Protein-coding exons, large single-copy region, small single-copy region, inverted repeat region, introns and intergenic spacers (Figure 2) [75], pseudogenes [89] or the differences between size, number of annotated genes [86] could be taken into accounts in phylogenetic analyses. Length-variations which based on the differences of indels (insertions and deletions) at certain locations of cpDNA genome were also considered as effective barcodes [35, 90].

Genetic information of this super-data is definitely great, enough to avoid the analytical limitation by different bioinformatics methods such as Maximum Likelihood (ML), Maximum Parsimony (MP) or Bayesian (BA). This means that the topologies of phylogenetic trees based on these three algorithms, which be usually congruent using short DNA sequences, are highly similar in the case [75].

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 there to $1,000 in 2016, a series of studies on plant barcoding was published in the next two years 2017 and 2018. Along with sequencing and assembling techniques, whole-plastome barcode may offer more informative sites and is considered as accurate and effective single barcode for identification in plants.

6. APPLICATION OF BARCODES TO DEVELOP OTHER IDENTIFICATION TECHNIQUES

Nucleotide information can be used to develop some species-specific amplification markers for quick and cheap identification of specific subjects. PCR primers derived from these techniques only react upon annealing to specific DNA sequences and give more specific and reproducible results than random amplifications. SCAR (sequence characterized amplified regions) technique succeeded in authentication three species of *Paphiopedilum armeniacum*, *Paphiopedilum micranthum*, *Paphiopedilum delenatii* and their hybrids by developing three species-specific primer pairs from ITS sequences [6]. Some studies focused on developing and comparing simple sequence repeat (SSR) system among screened taxa [86, 91, 92]. Kim *et* al. designed new primers for ARMS (amplification refractory mutation system) technique based on specific insertion in sequence of *Cypripedium macranthos* and SNPs (single nucleotide polymorphisms) in sequences of *Cypripedium japonicum* and *Cypripedium formosanum*) located inside *atp*F-*atp*H barcode. These three primer pairs can be used in combination to distinguish four *Cypripedium* species with different-size bands on the electrophoresis gel [4]. RFLP (restriction fragment length polymorphism) technique is a type of random PCR-based method in identification of subjects. However, RFLP based on the combination of ITS and some chloroplast sequences gave more reproducible and successful identification of native *Dendrobium* species in Thailand by Peyachoknagula *et* al. [93]. Therefore using known barcodes to develop other time- and cost-saving methods can also support molecular identification.

Furthermore, metabarcoding using high-through put sequencing can help identify a variety of species in multiple samples. A detection of species was performed simultaneously in 55 commercial salep products based on ITS barcode. Each sample was found to contain 1-55 species [94]. *Rbc*L, *mat*K and ITS were also used in combination to determine 16 orchid species as components of a common food named *Chikanda* in Zambia [95]. Based on this foundation, the authors alerted the

over-harvesting condition and called for the conservation of these rare orchids. This technique also opened a new trend in application of DNA barcodes.

7. CONCLUSIONS

Barcoding technique had been shown to be useful in many practical applications and classification studies. This method promises more optimal results than the PCR method especially when the price of the sequencing is decreasing. In barcoding technique using sequences, the chosen loci and algorithms are directly related to the identification results.

Whole genome comparison based on next generation sequencing and high-throughput sequencing has been shown to be more convenient and contain greater data than traditional barcoding. However, in terms of price, it is about $200 per sample up to now, which is 20 times more expensive than mini-barcodes ($11-13). In terms of technology, a powerful computer, which is not available in small laboratories, is needed to manage large genome data. Besides, it also takes a considerable time from the sequencing to the analyzing. For those reasons, traditional barcoding methods are still popular and effective in many cases. Hence these two barcoding trends can be performed according to the demands and conditions of different laboratories. In parallel, the sequencing techniques and analyzing tools should be improved to make it simpler and more convenient for researchers. The next orientation of identification should be a species identification gadget that is portable and requires no amplification step.

ACKNOWLEDGEMENT

The study was supported by Air Force Office of Scientific Research and Asian Office of Aerospace Research and Development (grant number FA23861514119). All authors declared no other competing financial interests.

ABBREVIATION

AFLP	Amplified Fragment Length Polymorphism
ARMS	Amplification Refractory Mutation System
BA	Bayesian
BLAST	Basic Local Alignment Search Tool
cp	complete
cpDNA	chloroplast DNA
IR	Inverted Repeat
LSC	Large single-copy region
MAP	Maximum a posteriori
ML	Maximum Likelihood
MP	Maximum Parsimony
NGS	Next Generation Sequencing
NJ	Neighbor-joining
OTU	Operational Taxonomic Units
PCR	Polymerase Chain Reaction
RADP	Random Amplified Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
SCAR	Sequence Characterized Amplified Region
SNP	Single Nucleotide Polymorphism
SSC	Small single-copy region
SSR	Simple Sequence Repeat
UPGMA	Unweighted Pair Group Method with Arithmetic mean

REFERENCES

1. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(23):8369-8374.

2. Kim HM, Oh SH, Bhandari GS, Kim CS, Park CW. DNA barcoding of Orchidaceae in Korea. Molecular Ecology Resources. 2014;14(3):499-507.
3. Abbas B, Dailami M, Listyorini FH, Munarti. Genetic Variations and Relationships of Papua's Endemic Orchids Based on RAPD Markers. Natural Science. 2017;9:377-385.
4. Kim JS, Kim HT, Son S-W, Kim J-H. Molecular identification of endangered Korean lady's slipper orchids (Cypripedium, Orchidaceae) and related taxa. Botany. 2015;93(9):603-610.
5. Peyachoknagul S, Mongkolsiriwatana C, Wannapinpong S, Huehne PS, Srikulnath K. Identification of native Dendrobium species in Thailand by PCR-RFLP of rDNA-ITS and chloroplast DNA. ScienceAsia. 2014;40:113–120.
6. Sun YW, Liao YJ, Hung YS, Chang JC, Sung JM. Development of ITS sequence based SCAR markers for discrimination of *Paphiopedilum armeniacum, Paphiopedilum micranthum, Paphiopedilum delenatii* and their hybrids. Scientia Horticulturae. 2011;127(3):405-410.
7. Pereira F, Carneiro J, Amorim A. Identification of species with DNA-based technology: current progress and challenges. Recent patents on DNA and gene sequences. 2008;2(3):187-99.
8. Yang F, Ding F, Chen H, He M, Zhu S, Ma X, Jiang L, Li H. DNA Barcoding for the Identification and Authentication of Animal Species in Traditional Medicine. Evidence-Based Complementary and Alternative Medicine. 2018;2018:18.
9. Waugh J. DNA barcoding in animal species: progress, potential and pitfalls. Bioessays. 2007;29(2):188-97.
10. Li X, Wu L, Wang J, Sun J, Xia X, Geng X, Wang X, Xu Z, Xu Q. Genome sequencing of rice subspecies and genetic analysis of recombinant lines reveals regional yield- and quality-associated loci. BMC Biology. 2018;16(1):102.
11. Ray S and Satya P. Next generation sequencing technologies for next generation plant breeding. Frontiers in plant science. 2014;5:367-367.
12. Thottathil GP, Jayasekaran K, Othman AS. Sequencing Crop Genomes: A Gateway to Improve Tropical Agriculture. Tropical life sciences research. 2016;27(1):93-114.
13. Zhou X, Bai X, Xing Y. A Rice Genetic Improvement Boom by Next Generation Sequencing. Current Issues in Molecular Biology. 2018;27:109-126.
14. Hebert PD, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proceedings of the Royal Society. B, Biological sciences. 2003;270(1512):313-21.
15. Das S and Deb B. DNA barcoding of fungi using Ribosomal ITS Marker for genetic diversity analysis: A Review. International Journal of Pure and Applied Bioscience. 2015;3(3):160-167.
16. Selvaraj D, Park JI, Chung MY, Cho YG, Ramalingam S, Nou I-S, Utility of DNA Barcoding for Plant Biodiversity Conservation. Vol. 1. 2013.
17. Techen N, Parveen I, Pan Z, Khan IA. DNA barcoding of medicinal plant material for identification. Current Opinion in Biotechnology. 2014;25:103-110.
18. Teixeira da Silva JA, Jin X, Dobranszki J, Lu J, Wang H, Zotz G, Cardoso JC, Zeng S. Advances in Dendrobium molecular research: Applications in genetic variation, identification and breeding. Molecular Phylogenetics and Evolution. 2016;95:196-216.
19. Vu H-T, Huynh P, Tran H-D, Le L. In Silico Study on Molecular Sequences for Identification of *Paphiopedilum* Species. Evolutionary Bioinformatics. 2018;14:117693431877454.
20. Fiser Pecnikar Z and Buzan EV. 20 years since the introduction of DNA barcoding: from theory to application. Journal of Applied Genetics. 2014;55(1):43-52.
21. Hollingsworth PM. Refining the DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(49):19451-2.
22. Saddhe AA and Kumar K. DNA barcoding of plants: Selection of core markers for taxonomic groups. Plant Science Today; Vol 5 No 1 (2018). 2017.
23. Duan H, Chen F, Liu W, Zhou C, Zhou Y. Research and Applications of DNA Barcode in Identification of Plant Species. Research in Plant Biology. 2014;4(3).
24. Ganie SH, Upadhyay P, Das S, Prasad Sharma M. Authentication of medicinal plants by DNA markers. Plant Gene. 2015;4:83-99.

25. Parveen I, Singh HK, Malik S, Raghuvanshi S, Babbar SB. Evaluating five different loci (rbcL, rpoB, rpoC1, matK, and ITS) for DNA barcoding of Indian orchids. Genome. 2017;60(8):665-671.
26. Rajaram MC, Yong C, Azlan GJ, Go R. DNA Barcoding of Endangered *Paphiopedilum* species (Orchidaceae) of Peninsular Malaysia. Phytotaxa. 2019;387(2):94-104.
27. Chattopadhyay P, Banerjee G, Banerjee N. Distinguishing Orchid Species by DNA Barcoding: Increasing the Resolution of Population Studies in Plant Biology. Omics. 2017;21(12):711-720.
28. Feng S, Jiang Y, Wang S, Jiang M, Chen Z, Ying Q, Wang H. Molecular Identification of *Dendrobium* Species (Orchidaceae) Based on the DNA Barcode ITS2 Region and Its Application for Phylogenetic Study. International journal of molecular sciences. 2015;16(9):21975-21988.
29. Wu C-T, Gupta SK, Wang AZ-M, Lo S-F, Kuo C-L, Ko Y-j, Chen C-L, Hsieh C-C, Tsay H-S. Internal Transcribed Spacer Sequence Based Identification and Phylogenic Relationship of Herba Dendrobii. Journal of Food and Drug Analysis. 2012;20(1):143-151.
30. Ginibun FC, Saad MRM, Hong TL, Othman RY, Khalid N, Bhassu S. Chloroplast DNA Barcoding of *Spathoglottis* Species for Genetic Conservation. Acta Hortuculturae. 2010;878:453-460.
31. Singh HK, Parveen I, Raghuvanshi S, Babbar SB. The loci recommended as universal barcodes for plants on the basis of floristic studies may not work with congeneric species as exemplified by DNA barcoding of *Dendrobium* species. BMC Research Notes. 2012;5:42.
32. Tanee T, Chadmuk P, Sudmoon R, Chaveerach A, Noikotr K. Genetic analysis for identification, genomic template stability in hybrids and barcodes of the *Vanda* species (Orchidaceae) of Thailand. African Journal of Biotechnology. 2012;11(55):11772-11781.
33. Yao H, Song JY, Ma XY, Liu C, Li Y, Xu HX, Han JP, Duan LS, Chen SL. Identification of *Dendrobium* species by a candidate DNA barcode sequence: the chloroplast psbA-trnH intergenic region. Planta Medica. 2009;75(6):667-9.
34. Givnish TJ, Spalink D, Ames M, Lyon SP, Hunter SJ, Zuluaga A, Iles WJD, Clements MA, Arroyo MTK, Leebens-Mack J, Lorena E, Ricardo K, Kurt MN, Whitten WM, Norris HW, Kenneth MC. Orchid phylogenomics and multiple drivers of their extraordinary diversification. Proceedings of the Royal Society B. 2015;282:20151553.
35. Santos C and Pereira F. Identification of plant species using variable length chloroplast DNA sequences. Forensic Science International: Genetics. 2018;36:1-12.
36. Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M. Identification of medicinal *Dendrobium* species by phylogenetic analyses using matK and rbcL sequences. Journal of Natural Medicines. 2010;64(2):133-8.
37. Tang Y, Yukawa T, Bateman RM, Jiang H, Peng H. Phylogeny and classification of the East Asian *Amitostigma alliance* (Orchidaceae: Orchideae) based on six DNA markers. BMC Evolutionary Biology. 2015;15:96.
38. Tran DD, Khuat HT, La TN, Nguyen TTT, Pham BH, Nguyen TK, Tran HD, Do MT, Tran DK. Identification of Vietnamese Native *Dendrobium* Species Based on Ribosomal DNA Internal Transcribed Spacer Sequence. Advanced Studies in Biology. 2018;10(1):1-12.
39. Poovitha S, Stalin N, Balaji R, Parani M. Multi-locus DNA barcoding identifies matK as a suitable marker for species identification in *Hibiscus* L. Genome. 2016;59(12):1150-1156.
40. EMBL-EBI. FASTA Help and Documentation. 2019. Accessed https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/FASTA+Help+and+Documentation.
41. Lloyd A. The Clustal W WW server at the EBI. 1997. Accessed http://www.ebi.ac.uk/embnet.news/vol4_3/clustalw1.html.
42. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods in molecular biology (Clifton, N.J.). 2009;537:39-64.
43. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.

44. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. 2017;20(4):1160-1166.
45. Ghorbani A, Gravendeel B, Selliah S, Zarre S, de Boer H. DNA barcoding of tuberous Orchidoideae: a resource for identification of orchids used in Salep. Molecular Ecology Resources. 2017;17(2):342-352.
46. Guo YY, Huang LQ, Liu ZJ, Wang XQ. Promise and Challenge of DNA Barcoding in Venus Slipper (*Paphiopedilum*). PLOS ONE. 2016;11(1):e0146880.
47. Meier R, Shiyang K, Vaidya G, Ng PK. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology. 2006;55(5):715-28.
48. Xiang XG, Hu H, Wang W, Jin XH. DNA barcoding of the recently evolved genus *Holcoglossum* (Orchidaceae: Aeridinae): a test of DNA barcode candidates. Molecular Ecology Resources. 2011;11(6):1012-21.
49. Xu S, Li D, Li J, Xiang X, Jin W, Huang W, Jin X, Huang L. Evaluation of the DNA Barcodes in *Dendrobium* (Orchidaceae) from Mainland Asia. PLOS ONE. 2015;10(1):e0115168.
50. Gao T, Yao H, Song J, Liu C, Zhu Y, Ma X, Pang X, Xu H, Chen S. Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. Journal of Ethnopharmacology. 2010;130(1):116-21.
51. Siripiyasing P. DNA barcoding of the *Cymbidium* species (Orchidaceae) in Thailand. African journal of agricultural research. 2012;7(3):393-404.
52. Meier R, Zhang G, Ali F. The Use of Mean Instead of Smallest Interspecific Distances Exaggerates the Size of the "Barcoding Gap" and Leads to Misidentification. Systematic Biology. 2008;57(5):809-813.
53. Guo Y-Y, Luo Y-B, Liu Z-J, Wang X-Q. Evolution and Biogeography of the Slipper Orchids: Eocene Vicariance of the Conduplicate Genera in the Old and New World Tropics. PLOS ONE. 2012;7(6):e38788.
54. Shaw J, Lickey EB, Schilling EE, Small RL. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany. 2007;94(3):275-88.
55. Trung KH, Khanh TD, Ham LH, Duong TD, Khoa T. Molecular Phylogeny of the Endangered Vietnamese *Paphiopedilum* Species Based on the Internal Transcribed Spacer of the Nuclear Ribosomal DNA. Advanced Studies in Biology. 2013;5(7):337 - 346.
56. Meyer CP and Paulay G. DNA barcoding: error rates based on comprehensive sampling. PLoS biology. 2005;3(12):e422-e422.
57. Gigot G, Van Alphen-Stahl J, Bogarin D, Warner J, Chase MW, Savolainen V. Finding a suitable DNA barcode for Mesoamerican orchids. Lankesteriana International Journal on Orchidology. 2007;7(1-2):200-203.
58. Parveen I, Singh HK, Raghuvanshi S, Pradhan UC, Babbar SB. DNA barcoding of endangered Indian *Paphiopedilum* species. Molecular Ecology Resources. 2012;12(1):82-90.
59. Yukawa T, Kinoshita1 A, Tanaka N. Molecular Identification Resolves Taxonomic Confusion in *Grammatophyllum speciosum* Complex (Orchidaceae). Bulletin of the National Museum of Nature and Science, Series B. 2013;39(3):137–145.
60. Chen L-J, Liu Z-J, Li Y-Y, Li L-Q. A new orchid *Paphiopedilum guangdongense* and its molecular evidence. Journal of Systematics and Evolution. 2010;48(5):350-355.
61. Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V. DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(8):2923-8.
62. Slabbinck B, Dawyndt P, Martens M, De Vos P, De Baets B. TaxonGap: a visualization tool for intra- and inter-species variation among individual biomarkers. Bioinformatics. 2008;24(6):866-7.

63. Nei M and Saitou N. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;4(4):406-425.
64. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981;17(6):368-76.
65. Penny D. Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts. Systematic Biology. 2004;53(4):669-670.
66. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 1985;39(4):783-791.
67. Mar JC, Harlow TJ, Ragan MA. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC evolutionary biology. 2005;5:8-8.
68. Kolaczkowski B and Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 2004;431:980.
69. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics. 2008;9(4):299-306.
70. Swofford DL. PAUP*. Phylogenetic analysis using parsimony and other methods. Version 4.0. 2003.
71. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25(7):1253-6.
72. Huelsenbeck JP and Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754-5.
73. Xu Q, Zhang G-Q, Liu Z-J, Luo Y-B. Two new species of *Dendrobium* (Orchidaceae: Epidendroideae) from China: evidence from morphology and DNA. Phytotaxa. 2014;174(3):15.
74. Xiang X-G, Jin W-T, Li D-Z, Schuiteman A, Huang W-C, Li J-W, Jin X-H, Li Z-Y. Phylogenetics of Tribe Collabieae (Orchidaceae, Epidendroideae) Based on Four Chloroplast Genes with Morphological Appraisal. PLOS ONE. 2014;9(1):e87625.
75. Yang J-B, Yang S-X, Li H-T, Yang J, Li D-Z. Comparative chloroplast genomes of *Camellia* species. PloS one. 2013;8(8):e73053-e73053.
76. Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal. 2011;9(3):328-33.
77. Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JM, Cronk Q. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. American Journal of Botany. 2012;99(2):320-9.
78. Parker J, Helmstetter AJ, Devey D, Wilkinson T, Papadopulos AST. Field-based species identification of closely-related plants using real-time nanopore sequencing. Scientific Reports. 2017;7(1):8345.
79. Day PD, Berger M, Hill L, Fay MF, Leitch AR, Leitch IJ, Kelly LJ. Evolutionary relationships in the medicinally important genus *Fritillaria* L. (Liliaceae). Molecular Phylogenetics and Evolution. 2014;80:11-9.
80. Turktas M, Aslay M, Kaya E, Ertugrul F. Molecular characterization of phylogenetic relationships in *Fritillaria* species inferred from chloroplast trnL-trnF sequences. Turkish Journal of Biology. 2012;36(5):552-560.
81. Khourang M, Babaei A, Sefidkon F, Naghavi MR, Asgari D, Potter D. Phylogenetic relationship in *Fritillaria* spp. of Iran inferred from ribosomal ITS and chloroplast trnL-trnF sequence data. Biochemical Systematics and Ecology. 2014;57:451-457.
82. Bi Y, Zhang MF, Xue J, Dong R, Du YP, Zhang XH. Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on *Fritillaria.* Scientific Reports. 2018;8(1):1184.
83. Chen X, Zhou J, Cui Y, Wang Y, Duan B, Yao H. Identification of *Ligularia* Herbs Using the Complete Chloroplast Genome as a Super-Barcode. Frontiers in pharmacology. 2018;9:695-695.

84. Lee J, Chon J, Lim J, Kim E-K, Nah G. Characterization of Complete Chloroplast Genome of *Allium victorialis* and Its Application for Barcode Markers. Plant Breeding and Biotechnology. 2017;5(3):221-227.
85. Dong W, Liu H, Xu C, Zuo Y, Chen Z, Zhou S. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: a case study on ginsengs. BMC genetics. 2014;15:138-138.
86. Zhou Y, Nie J, Xiao L, Hu Z, Wang B. Comparative Chloroplast Genome Analysis of Rhubarb Botanical Origins and the Development of Specific Identification Markers. Molecules. 2018;23(11).
87. Curci PL, De Paola D, Danzi D, Vendramin GG, Sonnante G. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. PLoS One. 2015;10(3):e0120589.
88. Yi D-K, Choi K, Joo M, Yang JC, Mustafina FU, Han J-S, Son DC, Chang KS, Shin CH, Lee Y-M. The complete chloroplast genome sequence of *Abies nephrolepis* (Pinaceae: Abietoideae). Journal of Asia-Pacific Biodiversity. 2016;9(2):245-249.
89. Zhang Y, Guan W, Zhang X, Li L. The Complete Chloroplast Genomes of Asteraceae Species. Research & Reviews: Journal of Botanical Sciences. 2016;5(1):24-28.
90. Jheng CF, Chen TC, Lin JY, Chen TC, Wu WL, Chang CC. The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis* orchids. Plant Science. 2012;190:62-73.
91. Lin J-Y, Lin B-Y, Chang C-D, Liao S-C, Liu Y-C, Wu W-L, Chang C-C. Evaluation of chloroplast DNA markers for distinguishing *Phalaenopsis* species. Scientia Horticulturae. 2015;192:302-310.
92. Yu XQ, Drew BT, Yang JB, Gao LM, Li DZ. Comparative chloroplast genomes of eleven *Schima* (Theaceae) species: Insights into DNA barcoding and phylogeny. PLOS ONE. 2017;12(6):e0178026.
93. Peyachoknagul S, Mongkolsiriwatana C, Wannapinpong S, Huehne P, Srikulnath K. Identification of native Dendrobium species in Thailand by PCR-RFLP of rDNA-ITS and chloroplast DNA. ScienceAsia. 2014;40(2):113–120.
94. Boer HJ, Ghorbani A, Manzanilla V, Raclariu AC, Kreziou A, Ounjai S, Osathanunkul M, Gravendeel B. DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. Proceedings of the Royal Society B: Biological Sciences. 2017;284(1863).
95. Veldman S, Kim SJ, van Andel TR, Bello Font M, Bone RE, Bytebier B, Chuba D, Gravendeel B, Martos F, Mpatwa G, Ngugi G, Vinya R, Wightman N, Yokoya K, de Boer HJ. Trade in Zambian Edible Orchids-DNA Barcoding Reveals the Use of Unexpected Orchid Taxa for Chikanda. Genes (Basel). 2018;9(12).