The use of Barcoding Sequences for the construction of phylogenetic trees in the Lauraceae

Lalit Kumar Dangol

Master's Thesis Faculty of Forest Sciences and Forest Ecology

Department of Forest Genetics and Forest Tree Breeding

Georg-August University of Göttingen

The use of Barcoding Sequences for the construction of phylogenetic trees in the Lauraceae

Submitted by

Lalit Kumar Dangol

Supervisor

Prof. Dr. Oliver Gailing

Co-supervisor

Prof. Dr. Konstantin Krutovsky

Master's thesis at the Faculty of Forest Sciences and Forest Ecology

Georg-August-Universität Göttingen

December, 2018

Georg-August-Universität Göttingen

Die Verwendung von Barcoding-Sequenzen für die Konstruktion phylogenetischer Bäume in den Lauraceae

Eingereicht von

Lalit Kumar Dangol

Betreuer

Prof. Dr. Oliver Gailing

Zweitbetruer

Prof. Dr. Konstantin Krutovsky

Masterarbeit an der Fakultät für Waldwissenschaften und Waldökologie

Georg-August-Universität Göttingen

Dezember 2018

Acknowledgements

I would like to express my deepest gratitude to all of them who were involved directly or indirectly to complete this dissertation. I am sincerely grateful to my supervisor Prof. Dr. Oliver Gailing for his support, helpful advice and patience guidance throughout my study and giving me the opportunity to work and learn under his supervision. I am extremely indebted to Prof. Dr. Konstantin V. Krutovsky for being my co-advisor and also for his intellectual guidance, precious suggestion throughout my thesis preparation.

Continuous and strong support, supervision, valuable comments, suggestions and scholar advice of all of them are the key to accomplish the preparation of this thesis. I would also like to thanks to Dr. Carina Moura for her continuous support during data analysis and results preparation. I am sincerely grateful to Larisa Kunz for teaching and supporting me during my computer work.

My grateful thanks go to the Department of Forest Genetics and Forest Tree breeding, University of Götingen and its members for supporting and providing all the facilities. I also gratefully acknowledge the (EFForTS) for providing the DNA data for barcoding and further analysis. I am deeply grateful to the International Office of the Georg-August-University for providing the financial aid during my thesis preparation period.

My friends Sudha Parajuli Dhungana, Bikash Kharel, Saurav Nepal and all Nepalese friends of Göttingen deserves vote of thanks for their support, motivation and kindness throughout the study period.

Last but not least, my huge and heartfelt gratitude is expressed to my beloved wife Sirjana Shrestha for her incessant support, huge love and patience during my study and stay in Germany. I am indebted to my parents, sisters, nephews, nieces and entire family members in Nepal for their countless gratitude and endless love without which my study would have been difficult that paved the way to complete my study.

Table of Contents

List of Figures

List of Tables

Zusammenfassung

DNA-Barcoding ist das Verfahren, das zur schnellen Identifizierung einer Spezies verwendet wird, basierend auf der Extraktion einer DNA-Sequenz aus einer beliebigen lebenden oder toten Gewebeprobe eines Organismus, der auf verschiedenen Gebieten, einschließlich der Erforschung von Blumen, weit verbreitet ist. Die Identifizierung war schwierig, da keine universellen Gene für alle Pflanzenarten verfügbar sind. matK und rbcL werden trotz Debatten über geeignete Gene für Pflanzen als Kernbarcodes für Pflanzen ausgewählt. Neben der traditionellen taxonomischen Klassifizierung dient das DNA-Barcoding als Ergänzung zur traditionellen Taxonomie und zur Beschleunigung des Identifizierungsprozesses.

Diese Studie wurde mit dem Ziel durchgeführt, DNA-Barcodes für Pflanzenarten der Laureceae-Familie in Sumatra unter Verwendung der beiden Kern-Barcodes für die Pflanze (rbcL und matK) zu generieren. Diese beiden Barcodes wurden auf der Grundlage ihrer Leistung bei der Identifizierung von Arten und der Überprüfung der morphologischen Identifizierung von Laureceae-Proben bewertet. Zweiundfünfzig Blattproben wurden von zweiunddreißig Parzellen gesammelt, die in vier verschiedenen Landnutzungssystemen des Nationalparks Bukit Duabelas und des Harapan-Regenwaldes verteilt wurden. Jede gesammelte Probe wurde von Taxonomen durch Vergleiche mit Referenzgutscheinen im Herbarium Bogoriensis und BIOTROP Herbarium Bogor, Indonesien, klassifiziert. Alle Laborverfahren wurden an der Abteilung für Waldgenetik und Waldbaumzucht der Georg-August-Universität durchgeführt.

Die Sequenzbearbeitung wurde sorgfältig mit der CodonCode Aligner ™ -Software für alle erfolgreich generierten Barcodes durchgeführt. Nach der Bearbeitung der Sequenzen wurden die molekularen Identifikations- und Phylogenese-Bäume aufgebaut. Die molekulare Identifizierung wurde durchgeführt, indem die erzeugten Barcodes in den Nucleotid-Datenbanken abgefragt wurden, d. H. NCBI GenBank, während die Stammbäume unter Verwendung der MEGA7-Software erstellt wurden. Gyrocarpus americanus subsp africanus aus der Familie der Herendiaceae (Ordnung Lauraels) als Außengruppe zur Verwurzelung der Stammbäume.

Im Vergleich zu matK wurde im Ergebnis für rbcL eine einfachere Amplifikation und Sequenzierung beobachtet. Getrocknete Blattproben wurden zur Extraktion von DNA-Materialien verwendet. Insgesamt wurden 52 Proben für die PCR-Amplifikation und Sequenzierung verwendet. Davon waren 45 erfolgreich amplifiziert (87%) und nur 43 wurden

erfolgreich für rbcL sequenziert (83%), im Vergleich dazu wurden nur 27 Proben erfolgreich amplifiziert (52%) und 23 erfolgreich sequenziert (44%) für matK.

Die molekulare Identifizierung nicht identifizierter Proben unter Verwendung von BLASTn für jeden Barcode ergab viele mehrdeutige Ergebnisse mit unterschiedlichen Spezies, die dieselben Identitätsprozentsätze und E-Werte (0,0) aufwiesen. Die Kombination beider Marker konnte jedoch den besten Treffer mit einer Sequenzidentität nahe 100% und einem E-Wert von 0,0 finden.

In dieser Studie wurden sechs phylogenetische Bäume mit zwei verschiedenen Methoden (Neighbor Joining und Maximum Likelihood) unter Verwendung von rbcL, matK und der Kombination beider Marker erstellt. Bäume, die mit beiden Methoden erstellt wurden, zeigten ähnliche Topologien mit geringfügigen Änderungen in der Position der Kladen- und Bootstrap-Werte. Die rbcL, matK-Sequenzen aus gesammelten Proben gruppierten sich korrekt zusammen mit den Genbank-Sequenzen, die die gleichen Gattungen darstellen. Es bildeten sich jedoch nur sehr wenige Sequenzen mit den Genbank-Sequenzen, die dieselbe Art repräsentieren. Einige der Probensequenzen, die morphologisch als Laureceae-Proben identifiziert wurden, gruppierten sich mit Arten der Rutaceae und anderen Familien, die das Ergebnis morphologischer Fehlklassifizierung und Fehlmarkierung und Kontamination während Laborverfahren sein können.

Aus dieser Studie kann geschlossen werden, dass zwei Barcode-Regionen, d. H. MatK und rbcL, nicht zufriedenstellend waren, aber als die Kern-Barcodes waren diese beiden Marker für die Verwendung bei der Identifizierung von Pflanzenspezies mindestens bis zur Gattungsebene wirksam. Die Kombination von matK und rbcL erwies sich jedoch als höher diskriminierend.

Schlüsselwörter: Barcode, phylogenetischer Baum, Laureceae

Summary

DNA barcoding is the process applied for the rapid identification of a species based on the extraction of DNA sequence from any living or dead tissue sample of any organism which has been widely used in the different fields including floral exploration. Identification has been difficult because of unavailability of universal genes that can work for all plant species. *matK* and *rbcL* are selected as the core barcodes for plants in spite of debates about suitable genes for plants. Along with the traditional taxonomic classification, DNA barcoding serves as a complement to traditional taxonomy and to accelerate the identification process.

This study was carried out with the aim to generate DNA barcodes for plant species of the Laureceae family in Sumatra using the two core barcodes for the plant (*rbcL* and *matK*). These two barcodes were evaluated based on their performance in identifying species and verification of morphological identification of Laureceae samples. Fifty two leaves samples were collected from thirty-two plots distributed in four different land use system of Bukit Duabelas National Park and the Harapan rainforest. Each collected sample was classified by taxonomists by making comparisons with reference vouchers at the Herbarium Bogoriensis and BIOTROP herbarium Bogor, Indonesia. All laboratory procedures have been done at the Department of Forest Genetics and Forest Tree Breeding, Georg-August-University.

Sequence editing was carefully done using the CodonCode Aligner™ software to all successfully generated barcode. After the editing of the sequences, the molecular identification and phylogenetic trees were constructed. Molecular identification was conducted by inquiring the generated barcodes to the nucleotide databases i.e. NCBI GenBank while the phylogenetic trees were constructed using MEGA7 software. *Gyrocarpus americanus* subsp *africanus* of Herendiaceae family (order Lauraels*)* as an outgroup to root the phylogenetic trees.

Easier amplification and sequencing were observed in the result for *rbcL* in compared to *matK*. Dried-leaf specimens were used to extraction of DNA materials. Altogether, 52 samples were used for PCR amplification and sequencing. Out of which, 45 were successfully amplified (87%) and only 43 were successfully sequenced (83%) for *rbcL* in comparision to only 27 samples being amplified successfully (52%) and 23 successfully sequenced (44%) for *matK*.

The molecular identification of unidentified samples using BLASTn for each barcode gave many ambiguous results with different species showing the same identity percentages

and E-values (0.0). However, the combination of both markers was successful in finding the best hit with a sequence identity close to 100% and E-value equal to 0.0.

Six phylogenetic trees were constructed in this study using two different methods (Neighbor Joining and Maximum Likelihood) using *rbcL*, *matK* and the combination of both markers. Trees constructed by both methods showed similar topologies with slight change in the position of clade and bootstrap values. The *rbcL*, *matK* sequences from collected samples correctly clustered together with the Genbank sequences representing the same genera. However, only very few sequences clustered together with the Genbank sequences representing the same species. Some of the samples sequences which were morphologically identified as Laureceae samples clustered together with species of the Rutaceae and other families which can be the result of morphological misclassification and mislabeling and contamination during laboratory procedures.

It can be concluded from this study that two barcode regions, i.e. *matK* and *rbcL,* were not satisfying but as the core barcodes, these two markers were effective to be used in plant species identification at least up to genus level. The combination of *matK* and *rbcL*, however, was proven to have a higher level of discriminatory power.

Key words: Barcoding, Phylogenetic tree, Laureceae

1. INTRODUCTION

1.1 DNA Barcoding

DNA barcoding is the process applied for the quick identification of a species based on the extraction of the DNA sequence from any living or dead tissue sample of any organism. It is one of the most efficient methods for correct identification of any plant or animal species in a simple, rapid, repeatable and reliable way (Walker, 2009). This method also helps to discover new species and to classify puzzling species (Ward et al 2008). DNA barcoding has to be used in combination with other traditional taxonomic method for describing species as it alone cannot describe new species (Prendini, 2005). When used with other sources of information, such as morphological data, bar coding is the most successful method in species description (Goldstein and DeSalle 2011). The process of DNA barcoding can be accomplished in two steps: a) establishing barcoding libraries of known species and b) matching or assigning barcode sequence of unidentified/unknown samples against the library for successful identification (Walker, 2009).

1.2 DNA barcodes

DNA barcode is defined as short genomic sequence extracted from a standardized portion of genome (Walker, 2009). Apart from species identification, DNA barcodes improve or supplement traditional taxonomy based on morphological characters (Hebert & Gregory, 2005). An ideal barcode must fulfill at least three criteria a) universality (simplicity in sequencing and amplification) b) quality of sequence and c) discriminatory power (P. M. Hollingsworth, Graham, & Little, 2011). It is very much challenging to find a universal DNA barcode for plants in comparison to animals (Cowan and Fay 2012). Few exceptions in specific taxa, much lower base substation rates, frequent genome rearrangements and transfer of genes between different genomes and across species in plants are different to high base substitution rate, highly conserved gene content and order in animals(Palmer et al 2000). Amplification across all taxa using standardized primers and better quality in sequencing are the most important characteristics of a universal barcode (Chase et al 2007).

The DNA barcode, “Mitochondrial gene cytochrome oxidase c subunit (COI)” is the most successful and common molecular genetic marker used for DNA barcoding in animals (Hebert et al., 2004). However, it has low discriminatory power in plants species and is not used for the plant barcoding (Cho et al., 2004; Fazekas et al., 2008). The search for the

universal and consistent DNA barcoding markers is proven to be difficult in plant species (Hollingsworth et al., 2011). As a result, many plant DNA barcodes with different efficiency for different plant species such as nuclear internal transcribed spacer (*ITS1* and *ITS2*), chloroplast intergenic spacers (*trnH-psbA*, *atpF-aptH*, etc.) and chloroplast coding regions (*rbcL*, *matK* , etc.) (CBOL Plant Working Group et al., 2009). Among these plant barcodes, *rbcL* and *matK* and their combination are suggested and employed as the main barcodes for plant species (CBOL Plant Working Group et al., 2009).

The reasons for choosing *rbcL* and *matK* were 1) *rbcL* is able to track evolutionary relationship of plant species and is easy to be amplified and sequenced (Hollingsworth et al., 2009) and 2) both have a high discriminatory power (Hollingsworth et al., 2011). Although *matK* has a higher discriminatory power than *rbcL*, it is more difficult to amplify across distantly related species (Hollingsworth et al., 2011).

1.2.1 *matK* gene

matK gene, that encodes a maturase enzyme, evolves rapidly (Hilu et al 1997) and is regarded to one of the most informative loci for determining phylogenetic relationships(Hilu et al 2003). At the center of the gene, the chloroplast *matK* marker consists of ca. 841base pairs(bp), located between bp 205-1046 (including primer sites) in the complete *Arabidopsis thaliana* plastid genome sequence(Hollingsworth et al 2011). Primers of *matK* needs to be optimized to be adapted to specific taxonomic groups.

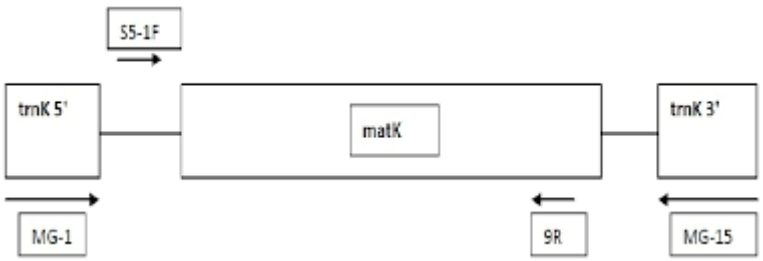


Figure 1: Diagram showing *matK* gene nested between *trnK* introns. The arrows indicate the positions of PCR and Sequencing primers (adapted from Hilu et al., 1999)

1.2.2 *rbcL* gene

The chloroplast rbcL marker consists of a 599 bp region at the 5` end of the gene, located at bp 1-599(including primer sites) in the complete *Arabidopsis thaliana* plastid genome sequence (Hollingsworth et al 2011). It is the first gene to be sequenced in plants which exists as a single copy and contains no introns(Zurawsky et al 1981)

It is the mostly used gene for retracing the evolutionary relationshipof plant groups diverged over historical time because of being one of the most conserved genes in the chloroplast genome. It is easily amplified and sequenced in most land plants, but showed too little variation to enable identifying all plant species (e.g., Hollingsworth et al 2009). Combination with more variable regions increase the power of this gene for phylogenetic purposes (Vijayan and Tsou 2010).

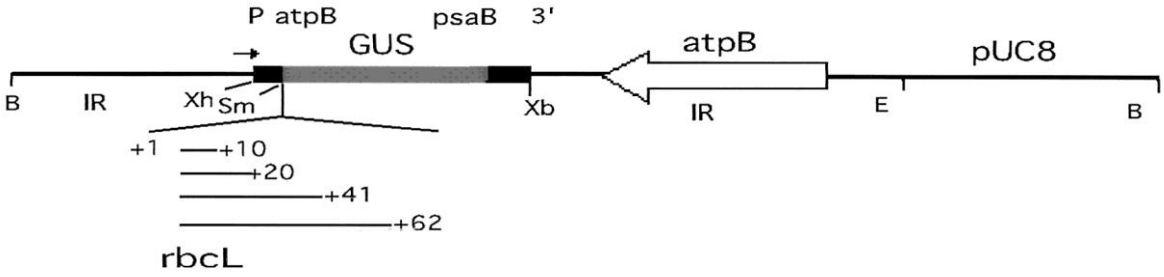


Figure 2: A diagram showing the organization of reporter genes containg *rbcl*

1.3 Use of DNA barcoding in species identification

Out of approximately 8.7 million species on earth (Mora et al., 2011), only 1.7 million species have been identified and described (List, 2011). An experienced taxonomist can identify a few hundreds to few thousands species in his lifetime, in this way to identify the remaining 7 million unidentified species, at least 8,700 additional taxonomists would be required, but, the number of professional taxonomists around the world is limited to 5,000-7,000 (Haas et al., 2005). Because of several limiting factors in tropical forests, species identification is slower in comparison to the species loss. Along with lack of number of specialist in the field of taxonomy, inaccessibility in taxonomic literature and inadequate herbarium collections are the problems in species exploration in tropical forest (Kiew 2002, Meyer and Paulay 2005).

Morphological species identification is time-consuming and unreliable. At the mean time, climate change, growth of human population, habitat destruction, pollution and many other detrimental factors have resulted in the rapid decline of the species. Many species are vulnerable or endangered and may become extinct even before they are discovered or scientifically explored. Reliable and efficient methods of species identification using molecular traits are the current necessity to speed up the species exploration in tropical forest (Finkeldey et al 2009).

1.4 Biodiversity in Sumatra

Indonesia is the land of some of the most magnificent tropical forest in the world harboring 10% of world's flowering plant species (about 25,000 flowering plants, 55% endemic), 16% of world's reptiles (781 species), 17% of birds (1,592 species) and 12% of worlds mammals (515 species) (CBD Secretariat (2016b)). Indonesia has the second largest rainforest in the world after Brazil (Hansen et al., 2009) and has 3% of world's total forest area (UN FAO, 2015).

Sumatra is the largest island in Indonesia and sixth largest island in the world. It is home to a rich flora and fauna. The natural area of the island has about 5,680,000 ha of Montane forest, 16,493,000 ha of tropical evergreen lowland forest and 25,154,000 ha of tropical evergreen lowland forest (Whitten et al., 2000). It has more than 10,000 plant species, 201 species of mammals, 580 bird species and has one of the largest tropical lowland forest areas in the world (Whitten et al., 2000). The biological diversity of tree species is extremely high in the Sumatran lowland forest. In spite of its richness in biodiversity, the central portion of island needs to be explored for its floristic diversity (Laumonier 1997).

1.4.1 Deforestation and forest degradation in Sumatra

The problem of deforestation and forest degradation exists all around the world. In tropical countries like Indonesia, the problem of mass forest destruction and forest degradation has been a great concern for years. Among the tropical countries, Indonesia alone accounts for approximately 12.8% of forest destruction (Hansen et al., 2008). In the nineties, the rate of forest clearing in Indonesia was the highest in the world (FAO, 2001). Expansion of palm oil cultivation and forest fires are the main reasons for the high scale mass clearing of the forests in 2001, the World Bank reported that the loss of forest areas in Sumatra was estimable in 7 million hectares from 1985 to 1997. From 2000 to 2012, about 1.21 million ha of lowland forest in Sumatra have been lost due to deforestation (Margono et al., 2014).

1.5 Lauraceae Family

Lauraceae are one of the basal angiosperm families with fossils dating back to the mid-Cretaceous (Drinnan et al. 1990). This family of flowering plants comprises about 2850 known species in about 45 genera worldwide (Christenhusz & Byng 2016). They are mainly distributed throughout the tropical and subtropical regions. Most of the Lauraceae are evergreen trees in habit. Exceptions include some two dozen species of *Cassytha*, all of which are obligately parasitic vines. The fruits of Lauraceae are drupes, one-seeded fleshy fruit with a hard layer, the endocarp, surrounding the seed. However, the endocarp is very thin, so the fruit resemble a one-seeded berry. (Little, S.A.; Stockey, R.A.; Penner, B. 2009)

1.5.1 Classification of Lauraceae

Classification within the Lauraceae is still not fully resolved. Although multiple classifications based on a different morphological and anatomical characteristics have been proposed, but none are fully accepted(Judd et al. 2007). Lauraceae is divided into two subfamilies, Cassythoideae and Lauroideae. *Cassytha* as a single genus, defined by herbaceous, parasitic habit Cassythoideae is classfied under Cassythoideae subfamily. The Lauroideae are then divided into three tribes: Laureae, Perseeae, and Cryptocaryeae (van der Weff and Richter 1996). Embryological studies had only been completed on individuals from 26 genera yielding a 38.9% level of knowledge (Kimoto,Y.,and H. Tobe 2001). A major challenge for developing a reliable classification is the large amount of variation (H van der Werff; J.G. Richter 1996).

1.6 The EFForTS-Project

The interdisciplinary research project “Ecological and Socioeconomic Functions of Tropical Lowland Rainforest Transformation Systems in Sumatra, Indonesia” (EFForTS) that focuses on ecological and socioeconomic effects of rainforest conversion on three different agricultural land-use systems (rubber plantation, oil palm plantation, jungle rubber agroforestry) in Jambi province, Indonesia (Drescher et al., 2016) . Based on three major lines of research (i) environmental processes, (ii) biota and ecosystem services, and (iii) human dimensions(Drescher et al., 2016)), this projects major objective is to facilitate in-depth understanding of the consequences of rainforest transformation to functional diversity of that area. The project area covers two landscapes in Jambi which are characterized by two different land systems namely Bukit Deuabelas National Park and Harapan Rainforest (Drescher et al. 2016). A core plot design was used to collect data regarding ecological dimensions while socioeconomic surveys design are used to collect data regarding human dimensions (Drescher et al. 2016). In each landscape, four core plots measuring 50m x 50m in

each of the four land-use systems were established in 2012, resulting in a total of 16 plots per landscape and 32 core plots in the overall project area (Drescher et al. 2016).

1.7 Objectives

The specific objective was to evaluate use of barcoding sequences for the construction of phylogenetic trees in the Lauraceae. To achieve this objective, the following tasks have been completed:

- Assessment of barcode universality of DNA barcode from Sumatra's Lauraceae samples
- Evaluation of DNA barcoding performance in species identification
- Phylogenetic analysis to compare molecular and morphological identification.

2. MATERIALS AND METHODS

2.1 Study sites

This study was conducted in two landscapes (Bukit Duabelas National Park and the Harapan Rainforest) in the Jambi Province- Sumatra, Indonesia that represents the remaining rainforests in Sumatra. The Bukit Duabelas National Park (1º51'S102º39'E) lies in the center of Jambi province which is a small national park with an area of 605 km². The area of the park is mainly covered by secondary forest while the northern part consists of primary rainforest with varying topography from flat land (164 meter in altitude) to slightly hilly area (438 meters in altitude). Meanwhile, Harapan Rainforest (2º14'S103º19'E) covering an area of 98,555 ha of rainforest in Jambi Province is one of the most biodiversity rich forests representing 20% of remaining lowland forest of Sumatra. The forest is managed by the NGOs groups i.e. Burung Indonesia, Birdlife International and Royal Society for the Protection of Birds (IUCN, 2018)

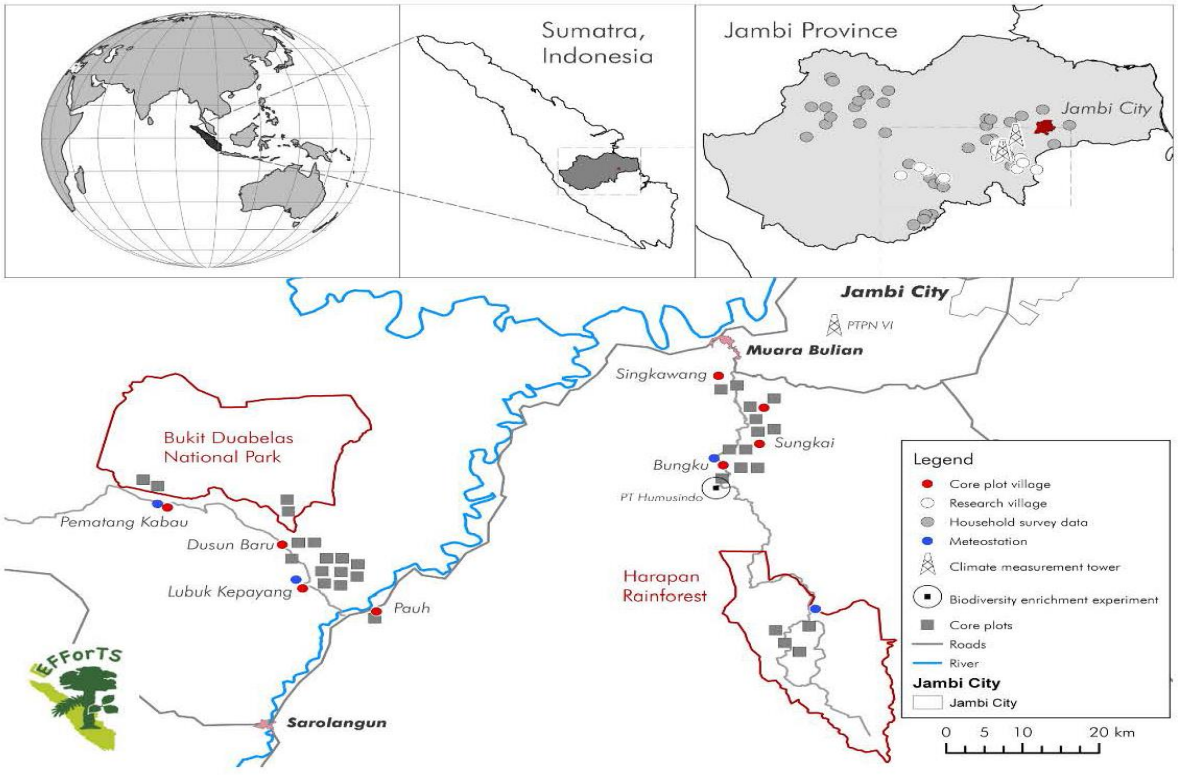


Fig. 2. Location of the area of CRC 990 showing the location of villages that are part of the province-level household surveys and the two landscapes (Bukit Duabelas National Park and Harapan Rainforest) where the core plot based design has been implemented. (Drescher, Rembold et al., 2016)

Figure 3: Map showing two study sites: Bukit Duabelas National Park and Harapan Rainforest respectively (Drescher, Rembold et al., 2016)

2.2 Study plots

In each landscape, four core plots were established representing the four different land use systems (lowland forest, jungle rubber, rubber monoculture plantation and oil palm monoculture plantation). Eight study plots were constructed in each of the four land use systems (8*4= 32 plots in total). Each plot was sized 50*50 m and contained a sub-plot of 5*5 m.

2.3 Specimen collection

Specimens were collected from all the study plots. Big trees (DBH≥30 cm) specimens were collected from each plot and under-story specimen were collected from the sub-plots. Each species was sampled at least 3 times. After that, leaf tissues of approx. 2cm2 size were collected and dried in silica-gel for DNA analysis. The specimens were identified by Hardianto Mangopo, JJ Afriastini, Katja Rambold, Iqbal Moh and Pak Ruspandi between 04/07/2013 to 07/20/2014.

Herbarium vouchers were prepared and stored in Bogoriensis and BIOTROP herbarium in Bagor Indonesia. All the collected samples were marked with a unique sample ID. Samples from the families were collected, out of which 52 specimens were provided for the analysis from Lauraceae family. Ninteen specimens were collected from plots of Jungle rubber and remaining from the Forest Plot. Information about all specimens regarding Sample ID, Core plots, Field name, and Species name is given in the (Annex 1) below.

2.4 Morphological identification of species

Each collected sample was classified by taxonomists by making comparison with reference vouchers at Herbarium Bogoriensis and BIOTROP herbarium Bogor, Indonesia. The morphological identification was then compared with molecular identification.

2.5 DNA analysis

All laboratory procedures have been done in the Department of Forest Genetics and Forest Tree Breeding, Georg August University.

2.5.1 DNA extraction

DNA extraction was done on the healthy dried leaf tissue from all the samples following the protocol of DNeasy Plant Mini Kit. Agarose electrophoresis gel (0.8-1%) with Lambda DNA size marker (Roche) (Sambrook et al., 1989) was used for checking the concentration and quality of the extracted DNA. It was then visualized by UV illumination using a Polaroid camera after staining in ethidium bromide.

2.5.2 Polymerase Chain Reaction (PCR), DNA amplification

After extraction of DNA, *rbcL* and *matK* markers were amplified performing Polymerase Chain reaction (PCR) using the universal primers used for the amplification are listed in (Table 1) below:

Table 1: List of Primers used

Region	Primer	Primer sequence(5'-3')	Reference
rbcL	*rbcL*a_f	ATGTCACCACAAACAGAGACTAAAGC	Kress & Erickson, 2007
	*rbcL*_r2	GAAACGGTCTCTCCAACGCAT	Fazekas et al., 2008
matK	MatKnewF	GTTCAAACTCTTCGCTACTGG	(Kress et al., 2009),(Yu et al., 2011)
	MatKnewR	GAGGATCCACTGTAATAATGAG	
	3Fkim(matK)	CGTACAGTACTTTTGTGTTTACGAG	
	1Rkin(matK)	CGTACAGTACTTTTGTGTTTACGAG	

PCR was done in the, Peltier Thermal Cyler PTC-200 (MJ Research Inc.) with a reaction mixture volume of 15 µl reaction mixture and 1 µl diluted sample for both markers used. Reaction mixture of PCR reagents is listed in the (Table 2) below:

Table 2: Reaction mixture of PCR reagents

Reagents	Volume (15ul)
H_2O	6.8
PCR Buffer	1.5
$Mgcl_2$	1.5
dNTPs	1.0
Primer F(5pmol/ml)	1.0
Primer R (5pmol/ml)	1.0
Taq polymerase	0.2

The PCR protocol consisted of an initial denaturation at 95°C for 15 min, followed by 35 cycles of denaturation at 94° C for 1 min, annealing at 50°C for 1 min, extension at 72°C for 1 min and a final extension at 72°C for 20 min. It is presented in (Table3) below:

Table 3: PCR protocol

Steps	Conditions
Step 1	Denaturation at 95ºCfor 15 minutes
Step 2	35 cycles of: ➢ Denaturation at 94ºc for 1 minute ➢ Annealing at 50ºC for 1 min ➢ Extension at 72ºC for 1:30 minutes
Step 3	Final extension at 72ºC for 20 minutes

Amplification success rates were calculated for both *rbcL* and *matK*. For this, the ratio of the number of successfully amplified samples in relation to the total number of PCRs using the corresponding marker was calculated.

2.5.3 DNA sequencing

The PCR reactions were purified using the innuPREP Gel Extraction Kit Protocol (Analytikjena, Jena, Germany) in order to obtain DNA for sequencing, after that by electrophoresis amplified fragments were separated in agarose gels. With the help of a razor, DNA fragments were excised from the gel and purified using the GENECLEAN ® Kit (MP Biomedicals, Illkirch, France).

Sequencing reactions were performed using the ABI PrismTM Big DyeTM Terminator Cycle Sequencing Ready Reaction Kit v1.1 (Applied Bio systems), based on the principle recommended by Sanger et al., (1977). Data from capillary electrophoresis on an ABI Prism 3100® Genetic Analyzer with the Sequence Analysis Software v3.1 (Applied Biosystems) were collected. Each DNA sample was sequenced in both directions separately with forward and reverse primers, respectively. The sequencing reaction mixture and protocol PCR are presented in (Table 4 and 5) respectively:

Table 4: Reaction mixture of PCR sequencing reagent

Reagent	Volume (µl)
H_2O	4.5
Barcoding	Dye 0.5
Buffer 5X	2.0
Primer F/R (5pmol/ml)	1

Table 5: Sequencing reaction protocol

Step	Condition
1	Initial denaturation for 1 min at 96ºC
2	35 cycles of ➢ Denaturation for 10 minutes at 96ºC ➢ Annealing for 10 minutes at 45ºC ➢ Elongation for 4 minutes at 60ºC
3	Final extension for 20 minutes at 72ºC

Sequencing success rates were calculated for each marker. The ratio of the number of bi-directional consensus sequences that were successfully obtained compared to the total number of successfully amplified samples was used for obtaining sequencing rate. The numbers of repetitions were excluded to obtain successful sequences.

2.6 DNA Sequence analysis

2.6.1 DNA Sequence editing

CodonCode Aligner™ software was used to align and edit sequence as accurate as possible by trimming the low quality nucleotides at the ends (the first and last 20 bp should contain less than 2 nucleotides showing quality values (QV) less than 20) of the forward and reverse sequences of investigated samples. Both strand traces, were visually checked for mismatches and manually edited by correcting sequencing errors and high quality consensus sequences were generated and saved for the further multiple sequence alignments and phylogenetic analysis. The low quality sequences which failed to assemble in bi-directional consensus sequences were removed from the data set. The resulting sequences were saved under the original ID and sample name. The stored names consisted of the original name assigned during species morphological identification, followed by the DNA extraction plate number and then field sample ID number.

2.6.2 DNA Sequence alignment

Sequences obtained from the collected samples and sequences downloaded from the NCBI GenBank were aligned for each marker. Samples with sequences generated or downloaded for both markers were compared using the multiple sequence alignment program MUSCLE (Edgar, 2004) embedded in CodonCode Aligner™. The results of the alignment were manually corrected and both ends were trimmed if needed to generate equal length multiple sequence alignments. The aligned *rbcL* and *matK* sequences were concatenated for

the same samples using Sequence Matrix software (Vaidya et al., 2011) and then the concatenated alignments were exported as NEXUS files.

2.7 Sequences from the NCBI Genbank

Sequence data from related species of the Lauraceae family were retrieved from the NCBI GenBank website. The homologous searches for the best matching sequences available in GenBank were done using the Basic Local Alignment System Tools for the nucleotides (BLASTn). The BLASTn program uses the query sequence and searches for the best matching highly similar and supposedly homologous sequences in the GenBank nucleotide sequence database. The sequences retrieved from the NCBI database were then aligned with sequences of collected samples and trimmed to make equal length multiple sequence alignments across the samples. The CodonCode MUSCLE aligner was used for the multiple alignments (Edgar, 2004).

2.8 Identification and verification of Barcode sequence using Nucleotide BLAST tools (BLASTn)

The BLASTn analysis was conducted to identify the unidentified samples and verify the questionable samples. The BLASTn analysis was done online on the NCBI website and the *rbcL* and *matK* sequences from collected samples. The best matching sequences based on E-value and percentage of maximum identity were downloaded and used further in multiple sequence alignments and phylogenetic analysis.

2.9 Construction of Phylogenetic trees

Phylogenetic trees were reconstructed, based on the aligned sequences from the laboratory (Marked with I) and sequences retrieved from the NCBI database. Phylogenetic trees were generated using MEGA7 (Kumar et al., 2016) for each marker, *rbcL* and *matK* separately and for concatenated sequences containing both markers. *Gyrocarpus americanus* subsp *africanus* of Herendiaceae family Lauraels order as an outgroup to root the phylogenetic tree.

The Neighbor Joining and Maximum Likelihood methods were used to generate phylogenetic trees. The Neighbor Joining method builds a tree based on the matrix of pair-wise genetic distances between samples studied and downloaded (Gascuel & Steel, 2006) while Maximum Likelihood uses evolutionary models to find evolutionary trees with the highest likelihood probability of explaining the sequence relationships (Felsenstein, 1981). For both the trees bootstrap support was computed using 1000 replicates.

3. RESULTS

3.1 Morphological identification of samples

At the begining, when samples were provided for the analysis, there were 61 samples. But because of misclassification (9 samples from other families), were mislabelled as Lauraceae (Annex 2). Hence, the number of samples is regarded to be 52 out of them 3 samples (Sample ID 1409, 4194, 4359) were not classified morphologically (Table 6).

Table 6: Identification fact sheets

S.no	Parameters	Numbers	Remarks
1	Numbers of Sample	52	
2	Number of unidentified samples	3	Sample ID, 4194, 1409, 4359
3	Number of identified samples	49	

3.2 Morphological classification of Samples

Samples collected from all the study plots were morphologically identified by a professional taxonomist matching the correspondent herbarium vouchers available at Herbarium Bogoriensis and BIOTROP Herbarium, Bogor, Idonesia. The morphological classification assigned samples to 1 subfamily Lauroideae and 13 genera in the Lauraceae family.

Table 7: Sample Composition of Lauraceae family samples

Subfamily	Genus	Number of samples
Lauroideae	*Phoebe*	4
	Persea	1
	Lindera	2
	Neolitsea	1
	Beilschmiedia	4
	Litsea	18
	Dehassia	3
	Cinamomum	1
	Cryptocaria	9
	Alseodaphne	2

	Ocotea	1
	Actinodaphne	2
	Endiandra	1

3.3 DNA amplification and Sequencing rates

Dried-leaf specimens were used for the extraction of DNA materials. Using the *rbcL* marker amplification success rate was 87% and success rate of sequencing was 83%. At the mean time, for the *matK*, amplification success rate was 52% and success rate of sequencing was 44% (Table 8).

Table 8: Amplification and Sequencing data for *rbcL* and *matK*

s.no	Parameters	*rbcL*	*matK*
1	Number of Samples	52	52
2	Successful amplification number	45	27
3	Successful amplification (%)	87%	52%
4	Successful Sequencing number	43	23
5	Successful Sequencing (%)	83%	44%

3.4 Barcoding Marker

3.4.1 *rbcL* barcoding marker

The amplification of this region was successful for most of the leaf samples. Altogether, 52 samples were used for PCR amplification and sequencing. Out of which, 45 were successfully amplified and only 43 were sequenced successfully. Along with 20 sequences from the NCBI Genbank and 36 lab sequences, 56 were used for further analysis. The final length of the multible sequence alignment was 509 bp.

3.4.2 *matK* barcoding marker

In comparision to rbcL marker, amplification and sequencing success rates of matK was very low. Out of 52 samples, only 27 samples amplified successfully and 23 were successfully sequenced. Forty Six sequences including 20 lab sequences and 26 downloaded from the NCBI Genbank were used for the further analysis. The final length of the multiple sequence was 545bp.

3.5 Use of Nucleotide BLAST (BLASTn) for sample identification and barcode analysis

During the morphological identification, 3 samples were unnamed because of different circumstances. Out of these 3 samples, sample (ID 4359) was not amplified and sequenced. Only one sample (ID 4194) amplified and sequenced successfully. Amplification and sequencing of samples (ID 1409) was successful in *rbcL* only. Using, *rbcL* and *matK* and combination of both markers, molecular identification of these unidentified samples were done (Table 9).

The molecular identification based on the *rbcL* region produced many ambiguous results with different species showing the E-values 0 and the same identity percentage. Query results of sample ID 4194 were very ambiguous as it gave variation in the genera for a sample. Species from *Laurus, Litsea, Machilus* genus were revealed.

Table 9: Identification of unidentified samples

s.no	Sample ID	Field name	*rbcL*			*matK*			*matK+rbcL*		
			Best match	E.Value	Identity (%)	Best match	E.Value	identity (%)	Best match	E.Value	Identity (%)
1	1409	Lauraceae sp. 07	*Pouteria campechiana*	0	99	No amplification and sequencing					
			Manilkara subsericea	0	99						
2	4194	Litsea sp. 26	*Laurus nobilis*	0	99	Litsea grandis	0	100	*Litsea firma*	0	100%
			Litsea verticillata	0	99	Litsea resinosa	0	99	*Litsea castanea*	0	100%
			Machilus thunbergii	0	99				*Litsea castanea*	0	100%
3	4359	Litsea sp. 27	No amplification and sequencing			No amplification and sequencing					

3.6 Use of Nucleotide BLAST (BLASTn) for barcode analysis

3.6.1 *rbcL*

Based on blast results of 36 *rbcL* sequences, none of the sequences found best match with sequences of same species to support morphological identification. Twenty sequences found best match with the species of same genus. Seven query sequences found best match

from species belonging to different genera. Sequences belonging to *Persea rimosa* and *Litsea noronhae* found best matches with the species belonging to the genera *Litsea* and *Cryptocarya* respectively.

Seven query sequences of sample ID found their best match with the species belonging to different families. Most of the query sequences showed ambiguous results while E-Value for all the sequences was 0.0. All the result of the BLASTn of *rbcL* sequences are presented in (Annex 3).

3.6.2 *matK*

Based on blast result of 20 *matK* sequences, one of the sequences (sample ID1836) had the best match with sequences of same species to conform the morphological identification. Twelve sequences found best matches with the species of the same genus. Six query sequences found best match from species belonging to different genera. Sequences belonging *Cryptocarya densiflora* (Sample ID 4556) found best match with the species belonging to the genera *Litsea*. (Table 10) shows the result of BLASTn based on *matK* marker.

One query sequences of sample (ID 4618) found their best match with the species belonging to a different family Rutaceae. Most of the query sequences showed ambiguous results while E-Value for all the sequences was 0.0.

Table 10: The homologous sequences best matching the *matK* sequences based on the BLASTn analysis

s.no	sample ID	Field name	Name of species	NCBI Gen bank best match	e value	Identity	Accession
1	1092	Lauraceae sp. 05	*Phoebe grandis*	*Phoebe lanceolata*	0	99%	LC388289.1
				Apollonias barbujana	0	99%	KJ189037.1
				Phoebe neurantha	0	99%	MH394355.1
2	1458	cf. Actinodaphne sp. 03	*Neolitsea cinnamomea*	*Neolitsea javanica*	0	99%	AB259096.1
				Actinodaphne trichocarpa	0	99%	KX546064.1
3	1835	Litsea sp. 10	*Litsea elliptica*	*Litsea elliptica*	0	98%	KJ708983.1
				Litsea firma	0	98%	KJ708984.1
				Litsea salicifolia	0	98%	KX546038.1
4	1836	Litsea sp. 10	*Litsea elliptica*	*Litsea elliptica*	0	99%	KJ708983.1
				Litsea firma	0	99%	KJ708984.1

				Litsea lancifolia	0	99%	KX545886.1
5	1874	Litsea sp. 11	*Litsea monopetala*	*Litsea firma*	0	98%	AB259078.1
				Lindera metcalfiana	0	98%	KX545869.1
				Litsea garrettii	0	98%	KR531071.1
6	2899	Litsea sp. 18	*Dehaasia cf. Firma*	*Alseodaphne semecarpifolia*	0	99%	NC_037491.1
				Phoebe lanceolata	0	99%	LC388289.1
				Phoebe neurantha	0	99%	MH394355.1
7	3348	Lauraceae sp. 19	*Litsea castanea*	*Litsea castanea*	0	98%	KJ708980.1
				Litsea elongata	0	98%	KR531068.1
8	3389	Litsea sp. 21	*Phoebe grandis*	*Phoebe zhennan*	0	99%	MF315089.1
				Phoebe neurantha	0	99%	MH394355.1
				Persea borbonia	0	99%	MF349978.1
9	4194	Litsea sp. 26		*Litsea grandis*	0	100%	AB259082.1
				Litsea resinosa	0	99%	AB259089.1
10	4201	Lauraceae sp. 33	*Phoebe grandis*	*Phoebe zhennan*	0	99%	MF315089.1
				Apollonias barbujana	0	99%	KJ189037.1
11	4518	Litsea sp. 31	*Litsea cubeba*	*Litsea costalis*	0	100%	AB259072.1
				Litsea sarawacensis	0	99%	AB259091.1
				Litsea erectinervia	0	99%	AB259075.1
12	4545	Litsea sp. 32	*Cryptocarya ferrea*	*Cryptocarya mannii*	0	99%	HG314990.1
				Machilus chrysotricha	0	99%	HQ427399.1
				Cryptocarya concinna	0	99%	KJ510890.1
13	4556	Cryptocaria cf. Laevigata	*Cryptocarya densiflora*	*Litsea grandis*	0	99%	AB259082.1
				Cinnamomum camphora	0	98%	LC228240.1
14	4618	Litsea sp. 33	*Litsea cf. Machilifolia*	*Melicope pteleifolia*	0	99%	KP093332.1
				Euodia simplicifolia	0	99%	FJ716733.1
				Acronychia pedunculata	0	98%	KJ510923.1
15	4835	Lauraceae sp. 40	*Dehaasia incrassata*	*Alseodaphne semecarpifolia*	0	99%	NC_037491.1
				Phoebe zhennan	0	99%	MF315089.1

				Apollonias barbujana	0	99%	KJ189036.1
16	4835	Lauraceae sp. 40	*Dehaasia incrassata*	*Alseodaphne huangliangshanensis*	0	100%	NC_037490.1
				Phoebe bournei	0	99%	MF315088.1
17	4842	Litsea sp. 35	*Litsea lanceolata*	*Litsea sarawacensis*	0	99%	AB259091.1
				Litsea garciae	0	99%	AB259081.1
				Cinnamomum bejolghota	0	99%	GQ248098.1
18	4911	Beilschmiedia sp. 02	*Endiandra rubescens*	*Beilschmiedia dictyoneura*	0	100%	HG314957.1
				Sinopora hongkongensis	0	99%	HG315005.1
19	4933	Litsea sp. 37	*Cryptocarya ferrea*	*Cryptocarya mannii*	0	99%	HG314990.1
				Cryptocarya gracilis	0	99%	HG314987.1
				Machilus chrysotricha	0	99%	HQ427399.1
20	4948	Lauraceae sp. 38	*Beilschmiedia cf. Madang*	*Beilschmiedia dictyoneura*	0	99%	HG314957.1
				Sinopora hongkongensis	0	99%	HG315005.1
				Beilschmiedia miersii	0	99%	AJ627916.1

3.6.3 Combination of *rbcL* and *matK* markers

Based on blast result of 13 common *rbcL and matK* sequences, (sample ID 3348) *Litsea castanea* found best match with sequences of same species conforming morphological identification. Four sequences found best match with the species of same genus.

Six query sequences found best match from species belonging to different genera. Interestingly, all Sequences belonging to *Phoebe grandis* found best match with the species belonging to the genera *Machilus*. (Sample ID 2899 and 4835) *Dehaasia cf. firma* and *Dehaasia incrassate* matched the best sequence from *Alseodaphne semecarpifolia*.

A single query sequences of (sample ID 4618) found their best match with the species belonging to Rutaceae family.

Table 11: The homologous sequences best matching the *rbcL* and *matK* sequences based on the BLASTn analysis

s.no	Sample ID	Morphological name	NCBI best match	E-value	Identity	Accession
1	1092	*Phoebe grandis*	*Machilus japonica*	0	100%	MF651954.1
			Machilus thunbergii	0	100%	NC_038204.1
			Machilus pauhoi	0	100%	NC_038203.1
2	1458	*Neolitsea cinnamomea*	*Neolitsea javanica*	0	100%	AB259096.1
			Neolitsea cassia	0	99%	AB259095.1
			Neolitsea parvigemma	0	99%	MF651959.1
3	1836	*Litsea elliptica*	*Litsea elliptica*	0	99%	KJ708983.1
			Litsea firma	0	99%	KJ708984.1
4	1874	*Litsea monopetala*	*Litsea mappacea*	0	99%	AB259086.1
			Litsea firma	0	99%	AB259078.1
			Litsea salicifolia	0	99%	KX546038.1
5	2899	*Dehaasia cf. firma*	*Alseodaphne semecarpifolia*	0	99%	NC_037491.1
			Alseodaphne semecarpifolia	0	99%	MG407595.1
			Alseodaphne gracilis	0	99%	NC_037489.1
6	3348	*Litsea castanea*	*Litsea firma*	0	99%	KJ708984.1
			Litsea castanea	0	99%	KJ708981.1
			Litsea castanea	0	99%	KJ708980.1
7	3389	*Phoebe grandis*	*Machilus japonica*	0	100%	MF651954.1
			Machilus thunbergii	0	100%	NC_038204.1
8	4194		*Litsea firma*	0	100%	KJ708984.1
			Litsea castanea	0	100%	KJ708981.1
			Litsea castanea	0	100%	KJ708980.1
9	4201	*Phoebe grandis*	*Machilus japonica*	0	100%	MF651954.1
			Machilus thunbergii	0	100%	NC_038204.1
			Machilus pauhoi	0	100%	NC_038203.1
			Machilus duthiei	0	100%	LC388311.1
10	4518	*Litsea cubeba*	*Litsea sp.*	0	100%	MH332646.1
			Litsea sp.	0	100%	MF419078.1
			Litsea sp.	0	100%	MF419075.1

11	4556	*Cryptocarya densiflora*	*Litsea elliptica*	0	100%	KJ708983.1
			Litsea firma	0	99%	KJ708984.1
			Litsea castanea	0	99%	KJ708981.1
12	4618	*Litsea cf. machilifolia*	*Melicope pteleifolia*	0	99%	KR531174.1
			Melicope pteleifolia	0	99%	KP093332.1
			Melicope pteleifolia	0	99%	KP093331.1
13	4835	*Dehaasia incrassata*	*Alseodaphne semecarpifolia*	0	100%	NC_037491.1
			Alseodaphne semecarpifolia	0	100%	MG407595.1
			Alseodaphne huanglianshanensis	0	100%	NC_037490.1

3.7 Phylogenetic analysis

Six phylogenetic trees were constructed based on the multiple sequence alignment of *rbcL*, *matK* and both *rbcL* and *matK* markers using two different methods: Neighbor joining (NJ) and Maximum likelihood (ML). In all the phylogenetic trees, the laboratory sequences are named according to the morphologically identified name followed by sample ID and (I) in the parentheses while species name are given for the sequences downloaded from NCBI GenBank. Similar topologies were observed among these trees with slight changes in bootstrap values. All branches were displayed and clades with less than 50% bootstrap support were presented.

3.7.1 *rbcL*

Neighbor Joining Method

The evolutionary history was inferred using the Neighbor-Joining method (Saitou et al. 1987). The optimal tree with the sum of branch length = 0.54903875 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsentein, 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2004) and are in the units of the number of base substitutions per site. The analysis involved 56 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 509 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2015).

Gyrocarpus americanus subsp *africanus* of Herendiaceae family Lauraels order as an outgroups, using Neighbor joining method the *rbcL* sequences from collected samples correctly clustered together with the Gene Bank sequences representing the same genera. However, only a few of the sequences clustered together with the Gene Bank sequences representing the same species (sample ID 1874) with bootstrap value of 55% (as shown in Fig 4). The result showed various clusters belonging to different tribes.

Sample ID 1092 and 4556 morphologically identified as *Phobe grandis* and *Cryptocarya densiflora* best matched with *Laurus* genera respectively formed clades with the Laureae tribe. The *Dehaasia* genera sequences (Sample ID 2899, 4835) formed a clade with the species of Perseae tribe including *Aleseodaphne* genera (sample ID 4250) and *Perseae* genera (Sample ID 1330). Within that clade, *Perseae* genera (Sample ID 1330) form a sub clade with *Ocotea veraguensis* with a high bootstrap value of 91%.

Sequences of *Cryptocarya* genera (Sample ID 4545, 4202, 4606, and 4804), *Litsea* genera (Sample ID 3061), *Ocotea* genera (Sample ID 4488) formed a clade with the Cryptocaryeae tribe with a boot strap value of 61%. The morphologically unclassified samples (sample IDs 4194) formed a cluster with Laureae tribe whereas (sample ID1409) clustered with species of another family.

The interesting result in this phylogenetic tree is that the sequence morphologically classified as *Cryptocarya pulchrinervia* (sample ID 4301) did not clustered with the species of the Laureceae family. Also, Sample ID (4249, 1409, 1458, 5060, 4655, and 1482) formed another clade different to 3 tribes mentioned above. Inclusion of species from one genus in another tribe, clustering with another family shows that samples could have been morphologically misclassified as Laureceae family and might have gone through contamination during laboratory proedure.

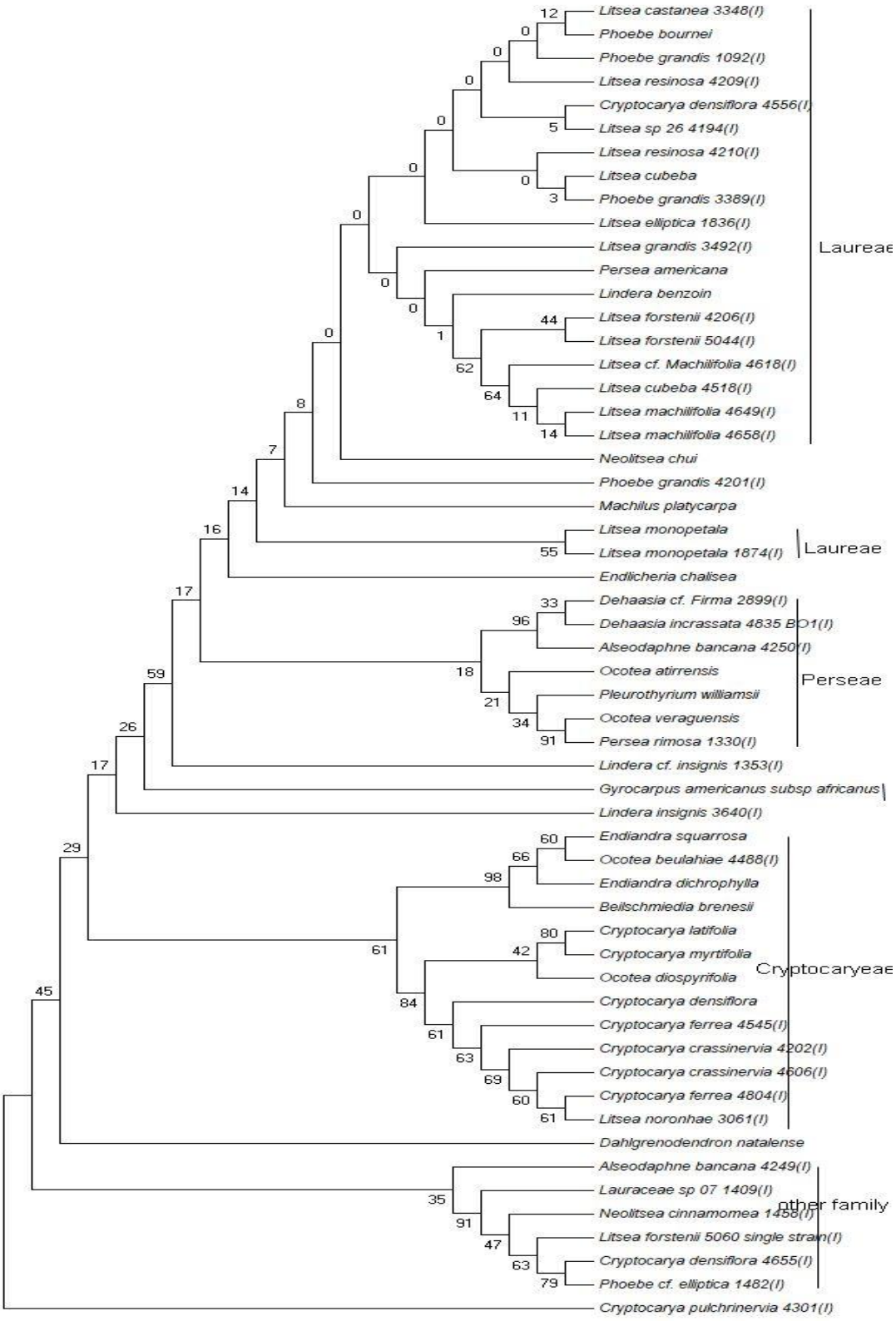


Figure 4 Phylogenetic tree constructed by Neighbor joining method of the samples representing the Laureceae plant family based on *rbcL* gene sequences. Samples marked by (I) in the parentheses represents sequences of collected samples

Sequences of *Cryptocarya* genera (Sample ID 4545, 4202, 4606, and 4804), *Litsea* genera (Sample ID 3061), *Ocotea* genera (Sample ID 4488) formed a clade with the Cryptocaryeae tribe with a boot strap value of 61%. The morphologically unclassified samples (sample IDs 4194) formed a cluster with Laureae tribe whereas (sample ID1409) clustered with species of another family. Sample ID (4249, 1409, 1458, 5060, 4655, and 1482) formed another clade different to 3 tribes mentioned above. Inclusion of species from one genus in another tribe, clustering with another family shows that samples could have been morphologically misclassified as Laureceae family and might have gone through contamination during laboratory proedure.

Maximum likelihood

The phylogenetic tree reconstructed by this method showed similar topology to the tree constructed by the Neighbor joining method with little bit change in position of clades and bootstrap values(Annex 4). This method was also successful in correctly differentiating the sequences representing colleceted samples according to the species and genus level. *Litsea monopetala* (Sample ID 1874) that formed a separate clade of Laureae tribe in the previous method clustered with bigger clade of Laureae tribe.

The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura et al.,1993). The tree with the highest log likelihood (-2310.3925) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The analysis involved 56 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 498 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2015).

3.7.2 *matK*

Phylogenetic trees constructed based on the *matK* marker using neighbor joining and maximum likelihood methods show similar topologies with slight change in the position of clade and bootstrap value for both methods. Both methods distinguished different tribes with high bootstrap support in comparison to *rbcL* marker. Fourty Six nucleotide sequences including 20 sequences from samples and 26 downloaded were used for further analysis.

Neighbor Joining Method

Gyrocarpus americanus subsp *africanus* of Herendiaceae family Lauraels order as an outgroup, Using the Neighbor joining method the *matK* sequences from collected samples correctly clustered together with the Gene Bank sequences representing the same genera.

Sequences of genus Litsea (Sample ID 1835, 1836, 3348, 1874, 4518, 4842) along with genus Cryptocarya (sampleID 4556) formed clade with the sequences of Laureae tribe with a bootstrap value of 64%. Within that clade, *Perseae* genera (Sample ID 1330) form a sub clade with *Ocotea veraguensis* with a high bootstrap value of 91 %(as shown in fig. 5).

Sequences from genus Dehaasia, Phoebe were clustered in the clade with Perseae tribe with bootstrap value of 54%. Sample ID 4545 and 4933 of Cryptocarya genera and Belilschmiedia (Sample ID 4948) were clustered in the Cryptocaryeae tribe with high bootstrap value of 90%. Unidentified sample ID 4194 clustered with Laureae tribe while the sample ID 4618 was unrooted (0% bootstrap value) with Laureceae family.

The evolutionary history was inferred using the Neighbor-Joining method (Saitou N. Et al., 1987). The optimal tree with the sum of branch length = 0.34064875 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein J. 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2004) and are in the units of the number of base substitutions per site. The analysis involved 46 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 545 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

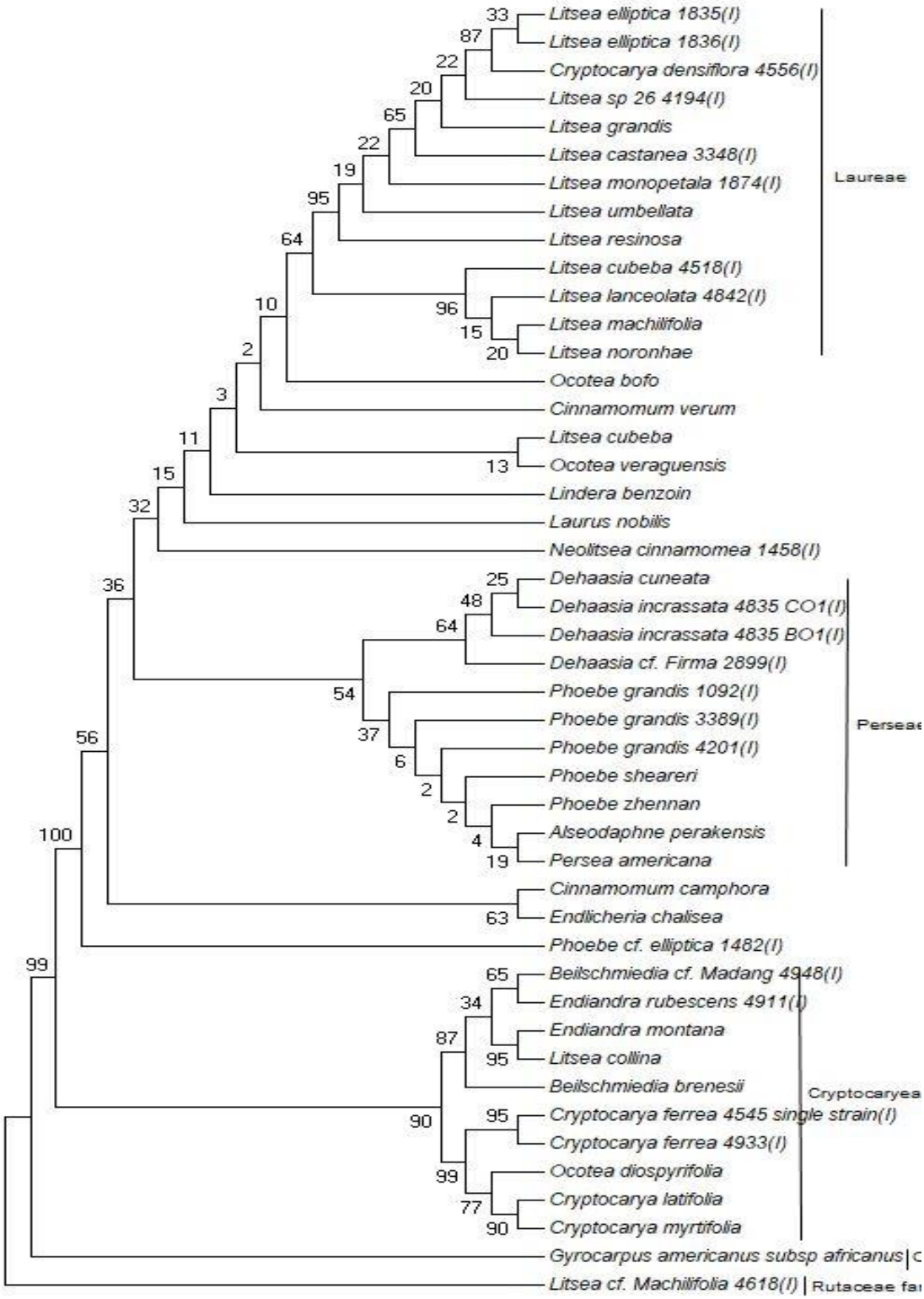


Figure 5:Phylogenetic tree constructed by Neighbor joining method of the samples representing the Laureceae plant family based on *matK* gene sequences. Samples marked by (I) in the parentheses represents sequences of collected samples

Maximum Likelihood Method

The phylogenetic tree reconstructed by this method showed similar topology to the tree constructed by the Neighbor joining method with little bit change in position of clades and bootstrap values(Annex 5). This method was also successful in correctly differentiating the sequences representing colleceted samples according to the species and genus level. Sequences of genus *Litsea* (Sample ID 1874 and3348) formed an individual sub clade within Laureae tribe.

The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura et al., 1993). The tree with the highest log likelihood (-1786.7471) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The analysis involved 46 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 535 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2015)

3.7.3 Combination of *rbcL* and *matK*

Neighbour Joining Method

Gyrocarpus americanus subsp *africanus* of Herendiaceae family Lauraels order as an outgroups, Using Neighbor joining method the rbcL and *matK* sequences from collected samples correctly clustered together with the Gene Bank sequences representing the same genera and formed 2 clades(Fig. 6).

Sequences from collected samples of Litsea, Neolitsea genera clustered in this clade of the Laureae tribe. *Cryptocarya densiflora* (Sample ID 4556) morphologically identified as a member of Cryptocaryeae tribe also clustered in this clade. This shows clearly that the samples might have undergone misidentification during morphological classification. Unidentified sample ID 4194 also clustered in this clade.

The evolutionary history was inferred using the Neighbor-Joining method (Saitou et al., 1987). The optimal tree with the sum of branch length = 0.28172178 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al.,

2004) and are in the units of the number of base substitutions per site. The analysis involved 22 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 1108 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2015).

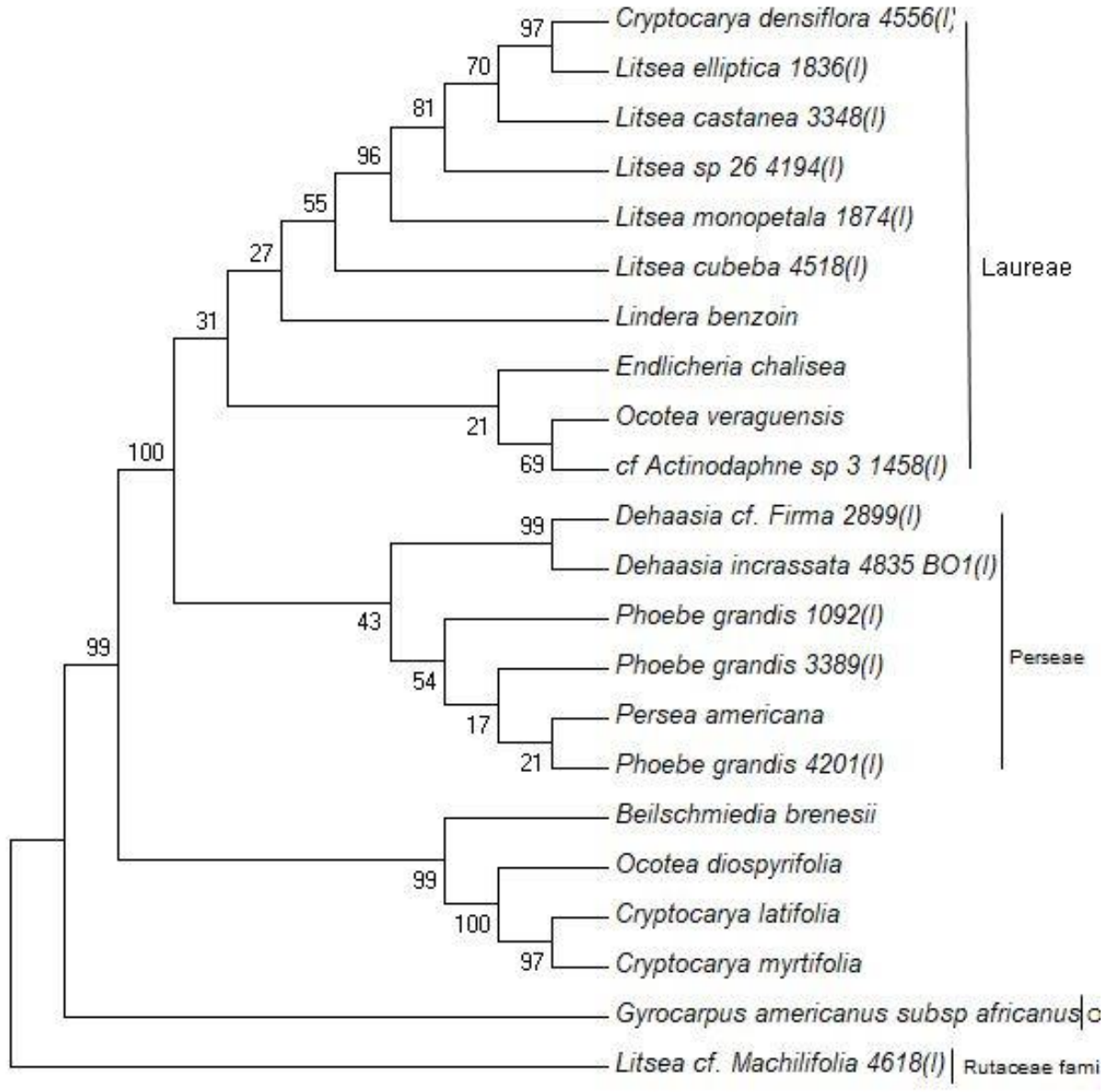


Figure 6: Phylogenetic tree constructed by the Neighbor joining method of the samples representing the Laureceae plant family based on *rbcL* and *matK* gene sequences.Samples marked by (I) in the parentheses represents sequences of collected samples

Maximum Likelihood Method

The phylogenetic tree reconstructed by this method showed similar topology to the tree constructed by Neighbor joining method with little bit change in position of clades and bootstrap values (Annex 6).

The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura et al., 1993). The tree with the highest log likelihood (-3143.6400) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The analysis involved 22 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1092 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2015).

4. DISCUSSION

4.1 Universality of DNA barcodes

Amplification and sequencing success rate are the most important criteria to evaluate DNA barcoding for plant identification (CBOL Plant Working Group et al., 2009; Hollingsworth et al., 2009). This study focused on the calculation of amplification and sequencing success rate for two main primers pairs for *rbcL* and *matK* as it can give better evaluation of DNA barcoding in plant identification. The result of this study concluded that *rbcL* has relatively higher amplification (87%) and sequencing (83%) success rate compared to those for *matK* (52% and 44% respectively). This result is in agreement with similar studies performed by Kress and Erickson (2007), Chen et al (2010), CBOL Plant Working Group (2009) and Hollingsworth et al (2009a, b) that also observed the lower amplification and sequencing success rate for *matK* was lower compared to *rbcL*.

It is also supported by the study of the tree species in tropical cloud forest of Bawangling resulted in low success rates of amplification and sequencing of *matK* fragment of $57.24 \pm 4.42\%$ and $50.82 \pm 4.36\%$ in comparision to the high success rate of amplification and sequencing ($75.26\% \pm 3.65\%$ and $63.84\% \pm 4.32\%$, respectively) found for *rbcL* (Kang Y et al. 2017). Recoverability of DNA sequences for *rbcL* was high in amplification and sequencing success was high 96.91% and 94.66%. At the same time, the amplification and sequencing success using the primer of *matK* was moderately success 79.05% and 65.81% for 2,590 sample included in the study (Fitri Yola, 2015).

Our results vary with the study which shows that the core barcodes (*rbcL* and *matK*) for Lauraceae plants had the best performance in PCR amplification and sequencing rate of 92.5% (Liu Z et al. 2017). Compared to the above core barcodes, *ITS* had a relatively low sequencing success rate of 39.1%, because of the lack of universal primers. Another study conducted in Lauraceae concluded that *psbA-trnH*, *matK*, and *rbcL* sequences were successfully amplified and sequenced at 100% in comparision to poor PCR amplification efficiency of the *ITS* region (Liu Zhen et al. 2012).

Many studies showed that it was almost impossible to calculate the amplification success rate of *matK* even by using the universal primers and the application of various conditions and dilutions. Again in case of tropical flora amplification seemed very difficult using *matK* as shown in a study of (Gonzalez et al., 2009) and compared to temperate flora (de Vere et al., 2012), (Bruni et al., 2010). The low success rate of amplification and

sequencing of *matK* fragments probably shows that it has a poor universality. This is possibly caused by in-sufficient number of primer pairs selected, and can be solved by using more and diverse primers (Kang Y et al. 2017).

Many other studies showed that *matK* marker can be useful as DNA barcode when it is used in certain taxa such as spices (De Mattia et al., 2011), tea plants (Stoeckle et al., 2011) and palms (Jeanson et al., 2011). Also, the success rate of *matK* can be high when it is used in combination with specific taxa primers (1). Thus, rather than considering matK as a difficult and less efficient marker for DNA barcode, its efficiency should be tested for various ranges of taxa as well as in combination with other taxa-specific primers.

The low rate of amplification and sequencing of tropical cloud forest species can be partly explained by Lauraceae and Fagaceae species with large amount of secondary metabolites (such as polysaccharides and phenolic compounds) effecting negatively on the extraction of high quality DNA (Kang Y et al. 2017). The poor success in amplification and sequencing by primers is probably due to the problem of secondary structure formation resulting in poor quality sequence data, multiple copy numbers. For a more conclusive and strong results, higher number of representations is recommended for further studies. It is very challenging to use barcoding successfully because of substantial phylogeographic structure of tropical flora. Use of taxonomically broad analysis and study beyond the focal geographic region so that evaluation and discrimination of potential sister taxa can be carried out for a better result.

An increasing number of molecular tests are being developed for plant species identification based on exploiting the physical and chemical properties of DNA using techniques such as DNA amplification, gene cloning, and nucleotide sequencing to deal with uncertainity in the field of biodiversity scientific researches. Primer universality is an important criterion for a useful DNA barcode. Thus, there is a challenge for plant DNA barcoding to find the most suitable markers to identify plants because of higher uncertainty.

4.2 Species identification success

4.2.1 Identification of samples using BLASTn compared to morphological identification

Among 52 samples of Laureceae, 49 samples were morphologically classified this is presented in the (Annex 1). Morphological identification becomes difficult because of invisibility of some feature followed by not well developed specimens which can make the identification even impossible. In order to identify these 3 samples that were not classified, they need to be compared with reference sequences. Because of the absence of this reference sequence, we use a local alignment tool known as BLAST algorithm (Altschul et al., 1997) which has been very popular for sequence analysis in barcoding in recent years (Ford et al., 2009).

In order to develop sequences of these 3 unidentified samples, sample ID 4194 with field name *Litsea sp*. 26 could undergo with the process of amplification and sequencing for both *rbcL* and *matK* markers whereas (sample ID 4359) could not go with both process for both markers. Meanwhile, (sample ID 1409) field named Lauraceae sp. 07 caould be amplified and sequence only for *rbcL*. Although there are no statistical methods that can measure the accuracy of identifications by BLAST (Munch et al., 2008), E-value and maximum identity are two statistics that can be used as measures of the likeliness of an identification being correct. Simply, the closer a hit is to 100% in sequence identity (and an E-value of 0), the more likely it is to have been correctly identified to species as well.

Sample ID 1409 when analyzed using rbcL marker as it could be sequenced successfully was from genus of another family Sapotaceae. Several factors could be the reason of misidentification like misidentification of the voucher by taxonomist, mislabeling or contamination during specimen collection and laboratory procedures like DNA extraction, amplification and sequencing. According to (Liu et al. 2017), who studied the phylogenic relationships of 409 individuals representing 133 species of 12 genera in the Lauraceae using DNA barcoding, 44 individuals from the ten species were misclassified by taxonomists. This supports the fact that the morphological features of the Lauraceae are indeed complex to be classified unambiguously.

Sample ID 4194 showed that it matches with different genera of Laureceae based on *rbcL* and matched with different species of same genera based on *matK*. Combined *rbcL* and *matK* also matched with different species of same genera. This clarifies that the barcode database is lacking information about the species level. The reference sequences stored in the database show high levels of incorrect species assignment. Use of different markers would

give better results in the accomplishment of the species level identification for uncertain samples. Availability of nucleotide data of the corresponding samples in the DNA sequence data base like Genebank and BOLD also affects the success of species identification.

4.2.2 Identification success according to the Best-Close Hit Match Analysis

Correct identification of the species at the genus level was satisfactory with *rbcL*, *matK* and combination of these barcode markers. However, very few sequences found best match with the sequences of same species. Matching with the species from other families, ambiguous and incorrect identification was also observed more for *rbcL* barcode in comparison to *matK*. It shows that *matK* perform better in species identification. The result is similar with the result of the study on ability of DNA barcoding to confirm the identities of 14 endangered endemic vascular plant species in Trinidad with the result of a higher proportion of correctly identified species obtained with *matK* sequences compared with *rbcL* sequences. *rbcL* sequences had a higher proportion of ambiguously classified sequences compared to *matK* sequences(Hosein et al. 2017).

The species identification success by DNA barcoding depends on investigated taxa and the markers used. We should also consider the best marker to be used according to the taxa in interest to achieve high success of DNA barcoding. An ideal DNA barcode must combine conserved regions for universal primer design, which show high rates of PCR amplification and sequencing and should also provide a high rate of success for species discrimination and identification which can be increased with combination of DNA barcodes(Liu Zhen et al. 2012).

We also profusely observed in this study matching with species of different genera in both markers. This shows that BLASTn matches alone cannot be used for sample identification at species level. This can happen due to several reasons: i) lack of enough sequences of our concerned species in NCBI GenBank database ii) lack of enough sequence variation in our barcode regions. Thus increase of nucleotide databases, use of other barcode regions might be helpful in the reliable species identification at the species level.

4.3 Phylogenetic analysis and comparison of molecular and morphological identification

The phylogenetic analysis was conducted in this study to see if *matK* and *rbcL* barcodes resolve the investigated species into appropriate taxonomical grouping. The topologies of the six phylogenetic trees (discussed above) were reconstructed in this study based on two methods (Neighbour joining and Maximum likelihood). All branch lengths and clades with less than 50% bootstrap support were presented. *Gyrocarpus americanus* subsp *africanus* of Herendiaceae family Lauraels order was downloaded for the purpose of using as outgroup. Beside some differences in the clade position and bootstrap values, the phylogenetic trees in both methods were seen more or less in consensus with each other. The phylogenetic tree reconstructed helped us in the correct identification of the misclassified samples.

One of the sequence morphologically identified as *Litsea cf. machillifolda*(Sample ID 4618), but in the phylogenetic tree based on matk and combined rbcl and matK, the sample was found to be in the clade belonging to Rutaceae family which was also confirmed by the BLASTn result of the sequences of that sample. Unidentified sample ID 4194 clustered with the clade of Laureae tribe. In our study, samples from genera *Litsea, Lindera, Laurus* formed a cluster within in Laureae tribe which is similar to the resultof phylogenetic analysis supporting the grouping of all Lindera species with three Litsea species and Laurus nobilis (Zhao M, et. al., 2018). The sequence of the morphologically identified as *Cryptocarya densiflora*(Sample ID 4556) which taxonomically belong to Cryptocaryeae tribe in Lauraceae family clustered together in the clade of Laureae tribe.

5. CONCLUSION

The use of barcoding sequences for the construction of phylogenetic trees in the Laureceae was performed in this thesis work. *rbcL* and *matK* were the two barcodes used for the analysis to make comparisons between morphological classification and classification using barcoding and understand phylogenetic relationships of species in the Laureceae family. Effectiveness of barcodes can be measured by the barcode quality and identification success. Comparisions were made between *rbcL, matk* and the combination of these two DNA barcodes.

DNA barcodes used in the analysis were from the tropical plants. *rbcL* was comparatively easier to amplify and sequence than *matK*, though without enough information to discriminate the samples at the species level. Combination of different primers can play a key role to improve amplification and sequencing success of *matK*. It was much more difficult to generate a good quality barcode using *matK* which was also not that successful in discriminating the samples at the species level. Morphologically identified samples were again cross checked with the molecular identification which indicated that the combination of both *rbcL* and *matK* barcode markers can perform better in species classification. The use of other barcode regions like *cpDNA* and *psbA* can give better results in the identification of samples at species level as the most suitable marker can solve the problems in plant DNA barcoding. Molecular identification based on the nucleotide database, Genbank (NCBI), were not clear enough for species identification and barcode analysis of two regions resulting more ambiguity and difficulty in proper species identification and barcode analysis. Incorrect morphological identification, mislabeling and contamination in the laboratory procedure can also be the cause behind ambiguous result.

Six phylogenetic trees were reconstructed based on *rbcL*, *matK* and combination of both using two different phylogenetic tree construction methods (Neighbor Joining and Maximum Likelihood) to make comparison between molecular and morphological identification. This analysis showed that neither *matK* and *rbcL* alone nor the combined marker were able to give a better identification at the species level. For most or all species sequences are missing in the database. However, the phylogenetic trees were successful in discriminating samples at the genus and other higher taxonomic levels.

Taking measures to minimize the misidentification can enhance quality of morphological identification. In combination to this, avoiding mislabeling and contamination

during lab work can bring effectiveness in correct molecular identification. Selection of the most suitable markers, robust primer combination, and utilization of supplement markers with *matk* and *rbcL* and expansion of nucleotide database can help in the success of DNA barcoding.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389–3402.

Bruni, I., De Mattia, F., Galimberti, A., Galasso, G., Banfi, E., Casiraghi, M., & Labra, M. (2010). Identification of poisonous plants by DNA barcoding approach. International Journal of Legal Medicine, 124(6), 595–603.

CBOL Plant Working Group, Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M.,Ratnasingham, S., Little, D. P. (2009). A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America, 106(31), 12794–12797.

Cho, Y., Mower, J. P., Qiu, Y.-L., & Palmer, J. D. (2004). Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proceedings of the National Academy of Sciences, 101(51), 17741–17746.

Davis, C. C., Latvis, M., Nickrent, D. L., Wurdack, K. J., & Baum, D. A. (2007). Floral gigantism in rafflesiaceae. Science, 315(5820), 1812.

De Mattia, F., Bruni, I., Galimberti, A., Cattaneo, F., Casiraghi, M., & Labra, M. (2011). A comparative study of different DNA barcoding markers for the identification of some members of Lamiacaea. Food Research International, 44(3), 693–702.

de Vere, N., Rich, T. C. G., Ford, C. R., Trinder, S. A., Long, C., Moore, C. W., Wilkinson, M. J. (2012). DNA barcoding the native flowering plants and conifers of wales. PLoS ONE, 7(6),1–12.

Drescher, J., Rembold, K., Allen, K., Beckscha, P., Buchori, D., Clough, Y., Scheu, S. (2016). Ecological and socio-economic functions across tropical land use systems after rainforest conversion. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* vol. 371,1694 (2016): 20150275.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797.

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Barrett, S. C. H. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE, 3(7).

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap Joseph Felsenstein. Evolution, 39(4), 783–791.

Ford, C. S., Hoot, S. B., Cowan, R. S., Gardens, R. B., & Wilkinson, M. J. (2009). Selection of candidate DNA barcoding regions for use on land plants, *Botanical Journal of the Linnean Society*, Volume 159, Issue 1, 1 January 2009, Pages 1–11.

Gascuel, O., & Steel, M. (2006). Neighbor-joining revealed. Molecular Biology and Evolution, 23(11), 1997–2000.

Gillman, L. N., Keeling, D. J., Gardner, R. C., & Wright, S. D. (2010). Faster evolution of highly conserved DNA in tropical plants. Journal of Evolutionary Biology, 23(6), 1327–1330.

Gonzalez, M. A., Baraloto, C., Engel, J., Mori, S. A., Pétronelli, P., Riéra, B., Chave, J. (2009). Identification of amazonian trees with DNA barcodes. PLoS ONE, 4(10).

H. M. Mahbubur Rahman, A., & Iffat Ara Gulshana, M. (2014). Taxonomy and Medicinal Uses on Amaranthaceae Family of Rajshahi, Bangladesh. Applied Ecology and Environmental Sciences, 2(2), 54–59.

Haas, F., Häuser, C. L., & Rica, C. (2005). Global Taxonomy Initiative, 240903.

Hansen, M. C., Stehman, S. V., Potapov, P. V., Loveland, T. R., Townshend, J. R. G., DeFries, R. S DiMiceli, C. (2008). Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. Proc. Natl. Acad. Sci. U.S.A, (105), 9439–9444.

Hansen, M. C., Stehman, S. V., Potapov, P. V., Arunarwati, B., Stolle, F., & Pittman, K. (2009). Quantifying changes in the rates of forest clearing in Indonesia from 1990 to 2005 using remotely sensed data sets. Environmental Research Letters, 4(3).

Hao, D. C., Chen, S. L., & Xiao, P. G. (2010). Sequence characteristics and divergent evolution of the chloroplast psbA-trnH noncoding region in gymnosperms. Journal of Applied Genetics, Journal of Applied Genetics September 2010, Volume 51, Issue 3, pp 259–273

Hebert, P. D. N., & Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. Systematic Biology, 54(5), 852–859.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. PLoS Biology, 2(10).

Hilu, K. W., & Liang, H. (1997). The matK gene sequence variation and application in plantsystematics. American Journal of Botany, 84(6), 830–839.

Hollingsworth, M. L., Andra Clark, A., Forrest, L. L., Richardson, J., Pennington, R. T., Long, D. G.,… Hollingsworth, P. M. (2009). Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources, 9(2), 439–457.

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. PLoS ONE, 6(5).

Hosein, F. N., Austin, N., Maharaj, S., Johnson, W., Rostant, L., Ramdass, A. C., & Rampersad, S. N. (2017). Utility of DNA barcoding to identify rare endemic vascular plant species in Trinidad. *Ecology and evolution*, *7*(18), 7311-7333.

Hwang, SW., Kobayashi, K., Zhai, S. et al.(2018). Automated identification of Lauraceae by scale-invariant feature transform Journal of Wood Science April 2018, Volume 64, Issue 2, pp 69–77

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution, 17(6), 368-376.

Jeanson, M. L., Labat, J. N., & Little, D. P. (2011). DNA barcoding: A new tool for palm taxonomists? Annals of Botany, 108(8), 1445–1451.

Kim, S.-C., Crawford, D. J., Jansen, R. K., & Santos-Guerra, A. (1999). The use of a non-coding region of chloroplast DNA in phylogenetic studies of the subtribe Sonchinae (Asteraceae:Lactuceae). Plant Systematics and Evolution, 215, 85–99.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), 111–120.

Kreft, H., & Jetz, W. (2010). A framework for delineating biogeographical regions based on species distributions. Journal of Biogeography, 37(11), 2029–2053.

Kress, W. J., & Erickson, D. L. (2007). A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region. PLoS ONE, 2(6): e508

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E.(2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences, 106(44), 18621–18626.

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America, 102(23), 8369–8374.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution, 33(7), 1870–1874.

List, R. (2011). Table 1 : Numbers of threatened species by major groups of organisms (1996 –2011) NOTES (for rows and columns as indicated by the superscripted numbers): Sources for Numbers of Described Species : Vertebrates. *World*, 9(April 2003), 2009–2010.

Little, D. P., & Little, D. P. (2007). A comparison of algorithms for the identi cation of specimens using DNA barcodes: examples from gymnosperms. *New York*, 23, 1–21.

Liu ZF, Ci XQ, Li L, Li HW, Conran JG, Li J (2017) DNA barcoding evaluation and implications for phylogenetic relationships in Lauraceae from China. PloS ONE 12(4), e0175788

Margono, B. A., Potapov, P. V., Turubanova, S., Stolle, F., & Hansen, M. C. (2014). Primary forest cover loss in indonesia over 2000-2012. Nature Climate Change, 4(8), 730–735.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on earth and in the ocean? PLoS Biology, 9(8), 1–8.

Moritz, C., & Cicero, C. (2004). DNA barcoding: Promise and pitfalls. PLoS Biology, 2(10).

Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., & Nielsen, R. (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. Systematic Biology, 57(5), 750–757.

Mwine, T. J., & Van Damme, P. (2011). Why do Euphorbiaceae tick as medicinal plants?: a review of Euphorbiaceae family and its medicinal features. Journal of Medicinal Plants Research, 5(5), 652–662.

Nei, M. (1987). The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees'. Science, 4(4), 406–425.

Rydbert, A. (2010). DNA barcoding as a tool for the identification of unknown plant material A case study on medicinal roots traded in the medina of Marrakech, 24.

IUCN, (2018) Saving the rainforest with a groundbreaking protected area management model | IUCN. (n.d.).

Saitou N. and Nei M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

Sen, L., Fares, M. A., Liang, B., Gao, L., Wang, B., Wang, T., & Su, Y. (2011). Molecular evolution of *rbcL* in three gymnosperm families : identifying adaptive and coevolutionary patterns, Biology Direct, 6, art no. 29

Simpson, M. G. (2010). Plant Molecular Systematics. Plant Systematics, 585–601.

Stepanović, S., Kosovac, A., Krstić, O., Jović, J., & Toševski, I. (2016). Morphology versus DNA barcoding: two sides of the same coin. A case study of Ceutorhynchus erysimi and C. contractus identification. Insect Science, 23(4), 638–648.

Stoeckle, M. Y., Gamble, C. C., Kirpekar, R., Young, G., Ahmed, S., & Little, D. P. (2011). Commercial teas highlight plant DNA barcode identification successes and obstacles. Scientific Reports, 1, 1–7.

Tamura, K., Nei, M., & Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences, 101(30),11030–11035.

Tamura K. and Nei M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10:512-526.

Tokuoka, T. (2007). Molecular phylogenetic analysis of Euphorbiaceae sensu stricto based on plastid and nuclear DNA sequences and ovule and seed character evolution. Journal of Plant Research, 120(4), 511–522.

Journal of Plant Research, 119(6), 599–616.

Vaidya, G., Lohman, D. J., & Meier, R. (2011). SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics, 27(2), 171–180.

Von Rintelen, K., Arida, E., & Häuser, C. (2017). A review of biodiversity-related issues and challenges in megadiverse Indonesia and other Southeast Asian countries. Research Ideas and Outcomes, 3, e20860.

Walker, J. M. (2009). IN M OLECULAR B IOLOGY TM Series Editor. Life Sciences (Vol. 531).

Yu, J., Xue, J. H., & Zhou, S. L. (2011). New universal matK primers for DNA barcoding angiosperms. Journal of Systematics and Evolution, 49(3), 176–181.

Annex

Annex1: List of Specimen collected

S.no	Sample No	Core plots	Plot	Subplot	Field name	Species	Family
1	1092	Forest	BF4	BF4e	Lauraceae sp. 05	*Phoebe grandis*	Lauraceae
2	1330	Jungle rubber	BJ5	-	Lauraceae sp. 06	*Persea rimosa*	Lauraceae
3	1353	Jungle rubber	BJ5	-	Litsea sp. 09	*Lindera cf. insignis*	Lauraceae
4	1409	Jungle rubber	BJ5	-	Lauraceae sp. 07		Lauraceae
5	1458	Jungle rubber	BJ5	BJ5b	cf. Actinodaphne sp. 03	*Neolitsea cinnamomea*	Lauraceae
6	1481	Jungle rubber		BJ5b	cf. Endiandra sp. 01	*Lithocarpus blumeanus*	Fagaceae
7	1482	Jungle rubber	BJ5	BJ5b	Persea sp. 02	*Phoebe cf. elliptica*	Lauraceae
8	1795	Jungle rubber	BJ3	BJ3d	cf. Lauraceae sp. 08	*Beilschmiedia maingayi*	Lauraceae
9	1835	Jungle rubber	BJ3	BJ3e	Litsea sp. 10	*Litsea elliptica*	Lauraceae
10	1836	Jungle rubber	BJ3	BJ3e	Litsea sp. 10	*Litsea elliptica*	Lauraceae
11	1874	Jungle rubber	BJ4	-	Litsea sp. 11	*Litsea monopetala*	Lauraceae
12	1899	Jungle rubber	BJ4		Persea sp. 03	*Terminalia subspathulata*	Combretaceae
13	2899	Forest	BF2	BF2a	Litsea sp. 18	*Dehaasia cf. firma*	Lauraceae
14	3061	Forest	BF2	BF2d	Litsea sp. 19	*Litsea noronhae*	Lauraceae
15	3348	Jungle rubber	HJ4	-	Lauraceae sp. 19	*Litsea castanea*	Lauraceae
16	3389	Jungle rubber	HJ4	HJ4b	Litsea sp. 21	*Phoebe grandis*	Lauraceae
17	3437	Jungle rubber	HJ4	HJ4d	Lauraceae sp. 22	*Litsea umbellata*	Lauraceae
18	3461	Jungle rubber	HJ4	HJ4d	Endiandra sp. 04	*Lithocarpus blumeanus*	Fagaceae
19	3462	Jungle rubber	HJ4	HJ4d	Endiandra sp. 04	*Lithocarpus blumeanus*	Fagaceae
20	3492	Jungle rubber	HJ4	HJ4e	Lauraceae sp. 23	*Litsea grandis*	Lauraceae
21	3604	Jungle rubber	HJ3	HJ3c	Cinnamomum sp. 02	*Cinnamomum iners*	Lauraceae
22	3640	Jungle rubber	HJ2	HJ2e	Litsea sp. 22	*Lindera insignis*	Lauraceae
23	4194	Forest	HF1	HF1a	Litsea sp. 26		Lauraceae
24	4201	Forest	HF1	HF1a	Lauraceae sp. 33	*Phoebe grandis*	Lauraceae
25	4202	Forest	HF1	HF1a	Litsea sp. 27	*Cryptocarya crassinervia*	Lauraceae
26	4206	Forest	HF1	HF1a	Litsea sp. 28	*Litsea forstenii*	Lauraceae
27	4209	Forest	HF1	HF1a	Litsea sp. 29	*Litsea resinosa*	Lauraceae
28	4210	Forest	HF1	HF1a	Litsea sp. 29	*Litsea resinosa*	Lauraceae
29	4249	Forest	HF1	HF1a	Lauraceae sp. 35	*Alseodaphne bancana*	Lauraceae
30	4250	Forest	HF1	HF1a	Lauraceae sp. 35	*Alseodaphne bancana*	Lauraceae
31	4301	Forest	HF1	HF1b	Cryptocaria sp. 02_1	*Cryptocarya pulchrinervia*	Lauraceae
32	4359	Forest	HF1	HF1c	Litsea sp. 27		Lauraceae
33	4360	Forest	HF1	HF1c	Litsea sp. 27	*Cryptocarya crassinervia*	Lauraceae
34	4390	Forest	HF1	HF1c	Cryptocaria sp. 03	*Litsea grandis*	Lauraceae
35	4397	Forest	HF1	HF1c	Lauraceae sp. 36	*Beilschmiedia madang*	Lauraceae
36	4488	Forest	HF1	HF1e	Beilschmiedia sp. 02	*Ocotea beulahiae*	Lauraceae
37	4509	Forest	HF1	HF1e	Lauraceae sp. 37	*Xanthophyllum rufum*	Polygalaceae
38	4518	Forest	HF2	-	Litsea sp. 31	*Litsea cubeba*	Lauraceae
39	4532	Forest	HF2	-	Lauraceae sp. 35_2	*Xanthophyllum rufum*	Polygalaceae
40	4545	Forest	HF2	-	Litsea sp. 32	*Cryptocarya ferrea*	Lauraceae
41	4547	Forest	HF2	-	Lauraceae sp. 35_2	*Xanthophyllum rufum*	Polygalaceae

42	4556	Forest	HF2	-	Cryptocaria cf. laevigata	*Cryptocarya densiflora*	Lauraceae
43	4606	Forest	HF2	-	Cryptocaria sp. 02_2	*Cryptocarya crassinervia*	Lauraceae
44	4618	Forest	HF2	HF2a	Litsea sp. 33	*Litsea cf. machilifolia*	Lauraceae
45	4649	Forest	HF2	-	Litsea sp. 31	*Litsea machilifolia*	Lauraceae
46	4655	Forest	HF2	-	Neolitsea sp. 01	*Cryptocarya densiflora*	Lauraceae
47	4658	Forest	HF2	-	Endiandra sp. 06	*Litsea machilifolia*	Lauraceae
48	4687	Forest	HF2	HF2c	Knema sp. 07	*Knema laurina*	Myristicaceae
49	4703	Forest	HF2	HF2c	Actinodaphne sp. 04	*Actinodaphne oleifolia*	Lauraceae
50	4704	Forest	HF2	-	Actinodaphne sp. 04	*Actinodaphne oleifolia*	Lauraceae
51	4804	Forest	HF3	-	Lauraceae sp. 39	*Cryptocarya ferrea*	Lauraceae
52	4835	Forest	HF3	-	Lauraceae sp. 40	*Dehaasia incrassata*	Lauraceae
53	4835	Forest	HF3	-	Lauraceaesp.40	*Dehaasia incrassata*	Lauraceae
54	4842	Forest	HF3	-	Litsea sp. 35	*Litsea lanceolata*	Lauraceae
55	4881	Forest	HF3	HF3a	Litsea sp. 36	*Knema laurina*	Myristicaceae
56	4911	Forest	HF3	HF3b	Beilschmiedia sp. 02	*Endiandra rubescens*	Lauraceae
57	4933	Forest	HF3	HF3c	Litsea sp. 37	*Cryptocarya ferrea*	Lauraceae
58	4948	Forest	HF3	-	Lauraceae sp. 38	*Beilschmiedia cf. madang*	Lauraceae
59	4992	Forest	HF3	HF3d	Lauraceae sp. 41	*Beilschmiedia madang*	Lauraceae
60	5044	Forest	HF3	HF3e	Litsea sp. 38	*Litsea forstenii*	Lauraceae
61	5060	Forest	HF3	HF3e	Litsea sp. 38	*Litsea forstenii*	Lauraceae

Annex 2: Misclassified sample

s.no	Sample ID	Field name	Species name	Family
1	1481	*Cf.Actinodaphne sp 3*	*Lithocarpus blumeanus*	Fagaceae
2	1899	*Persea sp 03*	*Terminalia subspathulata*	Combretaceae
3	3461	*Endiandra sp 4*	*Lithocarpus blumeanus*	Fagaceae
4	3462	*Endiandra sp 4*	*Lithocarpus blumeanus*	Fagaceae
5	4509	*Laureceae sp. 37*	*Xanthophyllum rufum*	Polygalaceae
6	4532	*Laureceae sp.35.3*	*Xanthophyllum rufum*	Polygalaceae
7	4547	*Laureceae sp.35.3*	*Xanthophyllum rufum*	Polygalaceae
8	4687	*Knema sp 7*	*Knema laurina*	Myristicaceae
9	4881	*Litsea sp 36*	*Knema laurina*	Myristicaceae

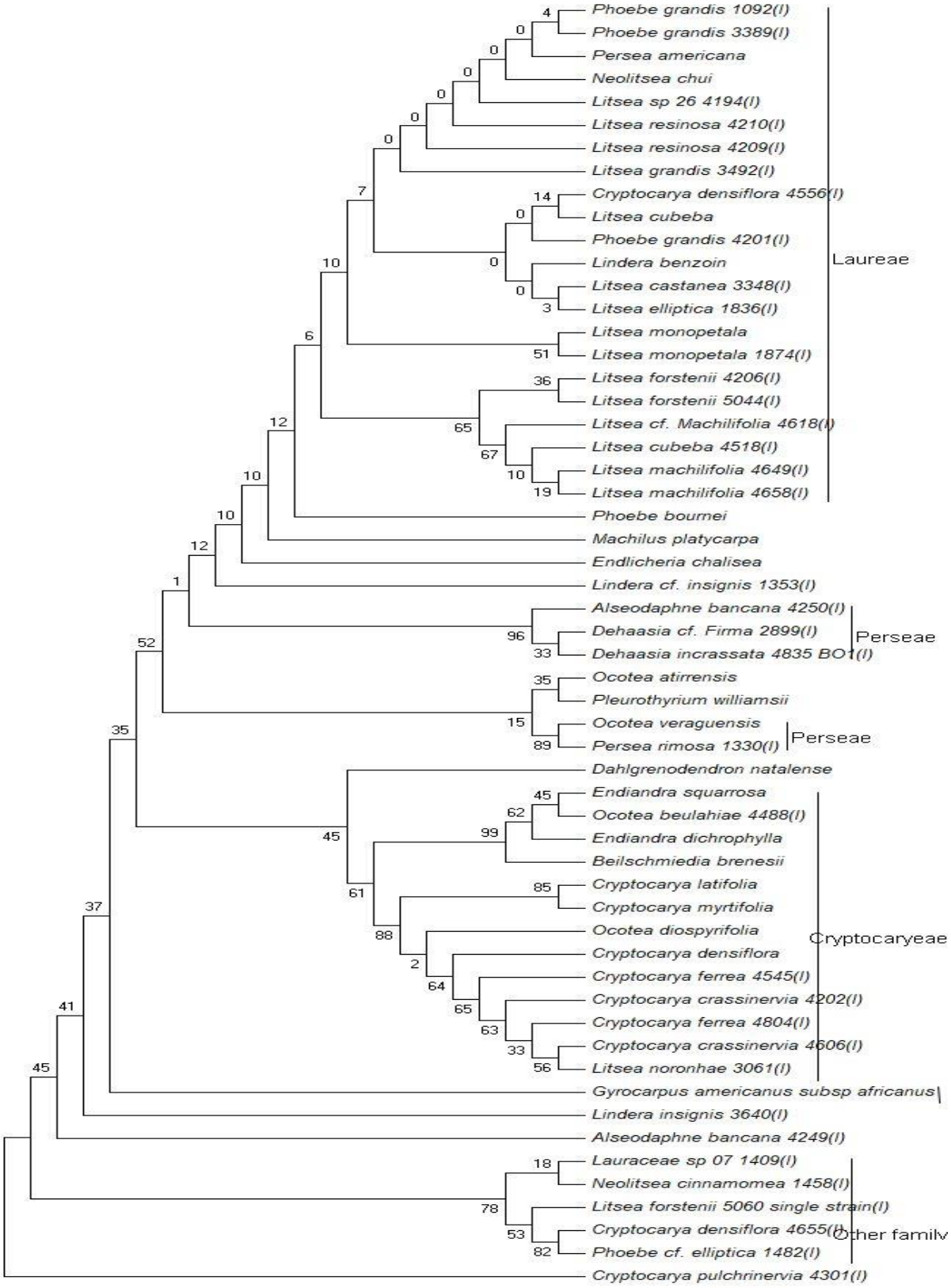
Annex 3: The homologous sequences best matching the *rbcL* sequences based on the BLASTn analysis

s.no	sample ID	name of species	best hit from NCBI	e value	identi	Accession
1	1092	*Phoebe grandis*	*Laurus nobilis*	0	99%	KY085912.1
			Phoebe omeiensis	0	99%	KX437772.1
2	1330	*Persea rimosa*	*Litsea monopetala*	0	99%	KF912876.1
			Cinnamomum camphora	0	99%	MF156716.1
3	1353	*Lindera cf. insignis*	*Litsea martabanica*	0	99%	KX546957.1
			Lindera communis	0	99%	KX546879.1
4	1409		*Pouteria campechiana*	0	99%	KX426215.1
			Manilkara subsericea	0	99%	KF981288.1
5	1458	*Neolitsea cinnamomea*	*Boswellia sacra*	0	99%	KY085915.1
			Canarium pimela	0	99%	MF166608.1
6	1482	*Phoebe cf. elliptica*	*Lithocarpus henryi*	0	100%	AY147097.1
			Quercus acutissima	0	99%	MH607377.1
			Quercus djiringensis	0	99%	LC318798.1
			Lithocarpus nitidinux	0	99%	LC318966.1
7	1836	*Litsea elliptica*	*Neolitsea zeylanica*	0	100%	KX909581.1
			Litsea rubicunda	0	100%	MF435345.1
8	1874	*Litsea monopetala*	*Litsea salicifolia*	0	99%	KX546976.1
			Lindera metcalfiana	0	99%	KR529537.1
9	2899	*Dehaasia cf. Firma*	*Alseodaphne semecarpifolia*	0	99%	NC_037491.1
			Lindera thomsonii	0	99%	KX546915.1
10	3061	*Litsea noronhae*	*Cryptocarya calcicola*	0	99%	KX546867.1
			Alseodaphne andersonii	0	99%	KR528686.1
11	3348	*Litsea castanea*	*Litsea chunii*	0	99%	MH116240.1
			Phoebe tavoyana	0	99%	KX547163.1
12	3389	*Phoebe grandis*	*Laurus nobilis*	0	100%	KY085912.1
			Phoebe omeiensis	0	100%	KX437772.1
13	3437	*Litsea umbellate*	*Syzygium cumini*	0	99%	GQ870669.3
			Luma apiculata	0	99%	KX162972.1
			Acca sellowiana	0	99%	KX289887.1
14	3492	*Litsea grandis*	*Neolitsea zeylanica*	0	99%	KX909581.1
			Phoebe rufescens	0	99%	KX547159.1
			Machilus robusta	0	99%	KX547046.1
15	3640	*Lindera insignis*	*Acer truncatum*	0	96%	NC_037211.1

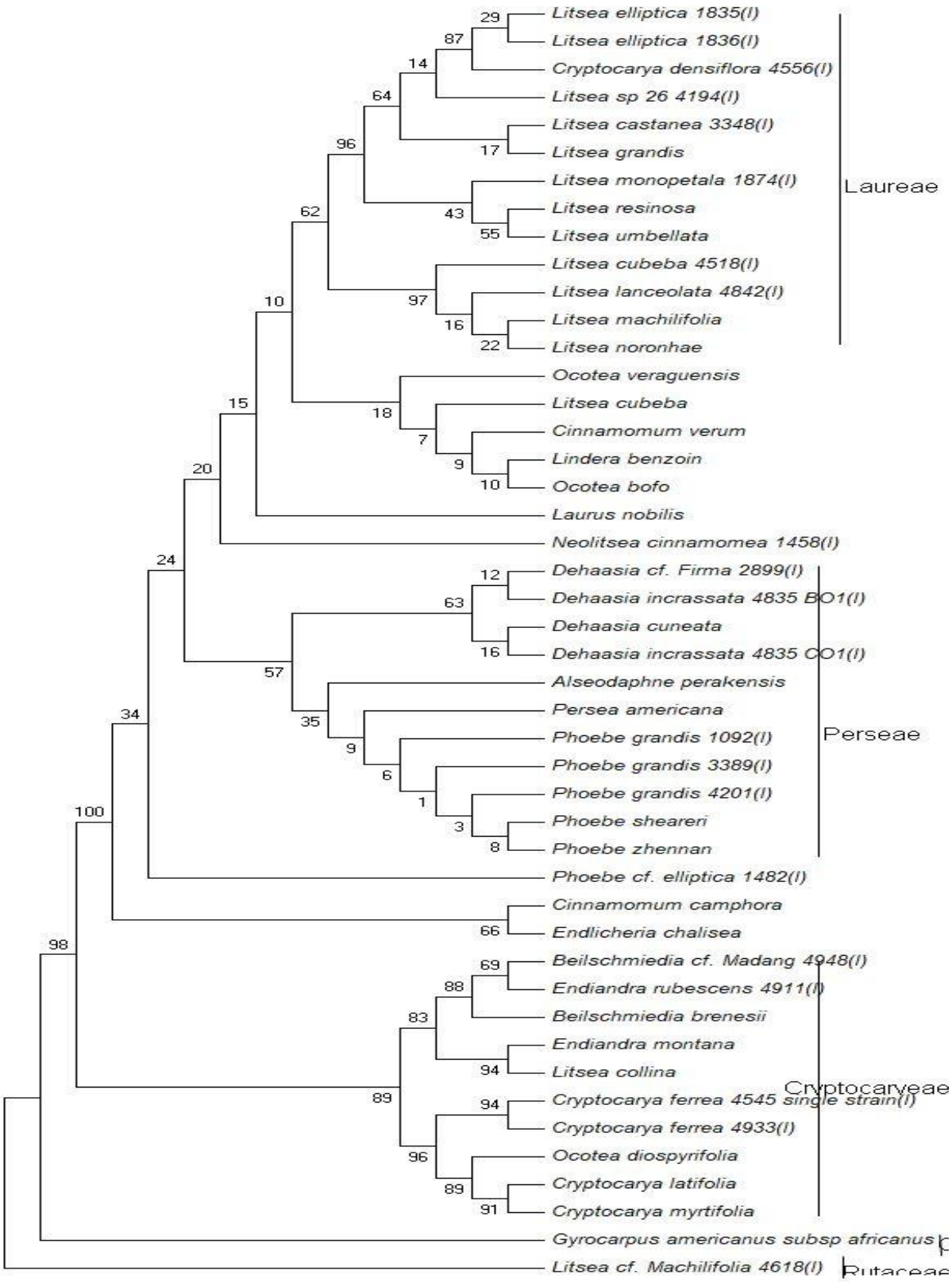
			Dipteronia sinensis	0	96%	KT878501.1
16	4194		*Laurus nobilis*	0	99%	KY085912.1
			Litsea verticillata	0	99%	KJ439988.1
			Machilus thunbergii	0	99%	NC_038204.1
17	4201	*Phoebe grandis*	*Phoebe omeiensis*	0	100%	KX437772.1
			Phoebe neurantha	0	99%	MH394355.1
			Machilus yunnanensis	0	99%	KT348516.1
18	4202	*Cryptocarya crassinervia*	*Cryptocarya chinensis*	0	99%	LC212965.1
			Alseodaphne andersonii	0	99%	KR528686.1
19	4206	*Litsea forstenii*	*Litsea glutinosa*	0	99%	KU382356.1
			Laurus nobilis	0	99%	AF197593.1
20	4209	*Litsea resinosa*	*Phoebe omeiensis*	0	99%	KX437772.1
			Machilus balansae	0	99%	KT348517.1
21	4210	*Litsea resinosa*	*Phoebe omeiensis*	0	99%	KX437772.1
			Machilus pauhoi	0	99%	NC_038203.1
22	4249	*Alseodaphne bancana*	*Mauloutchia chapelieri*	0	99%	AF197594.1
			Coelocaryon preussii	0	99%	AY743437.1
			Staudtia kamerunensis	0	99%	KC628429.1
			Virola michelii	0	99%	FJ038130.1
23	4250	*Alseodaphne bancana*	*Alseodaphne semecarpifolia*	1.00E -127	82%	NC_037491.1
			Litsea martabanica	5.00E -126	82%	KX546957.1
24	4301	*Cryptocarya pulchrinervia*	*Cryptocarya chinensis*	0	94%	LC212965.1
			Phoebe neurantha	0	93%	MH394355.1
25	4360	*Cryptocarya crassinervia*	*Cryptocarya calcicola*	1.00E -161	98%	KX546866.1
			Cryptocarya concinna	1.00E -161	98%	KJ439989.1
26	4488	*Ocotea beulahiae*	*Endiandra discolor*	0	99%	KT588615.1
			Beilschmiedia yunnanensis	0	100%	KX546815.1
			Cinnamomum chartophyllum	0	100%	KR528997.1

27	4518	*Litsea cubeba*	*Laurus nobilis*	0	99%	KY085912.1
			Litsea glutinosa	0	99%	KU382356.1
28	4545	*Cryptocarya ferrea*	*Cryptocarya chinensis*	0	99%	LC212965.1
			Cryptocarya chinensis	0	99%	KJ439983.1
29	4556	*Cryptocarya densiflora*	*Laurus nobilis*	0	99%	KY085912.1
			Machilus balansae	0	99%	KT348517.1
			Litsea monopetala	0	99%	KF912876.1
30	4606	*Cryptocarya crassinervia*	*Cryptocarya chinensis*	0	99%	LC212965.1
			Cryptocarya putida	0	99%	JN564212.1
			Cryptocarya acutifolia	0	100%	KX546862.1
31	4618	*Litsea cf. Machilifolia*	*Phoebe omeiensis*	0	99%	KX437772.1
			Litsea glutinosa	0	99%	KU382356.1
			Persea Americana	0	99%	L14620.1
32	4649	*Litsea machilifolia*	*Laurus nobilis*	0	99%	KY085912.1
			Machilus balansae	0	99%	KT348517.1
			Phoebe neurantha	0	99%	MH394355.1
33	4655	*Cryptocarya densiflora*	*Callerya vasta*	0	99%	AY308806.1
			Tibetia liangshanensis	0	97%	MF193597.1
			Caragana gerardiana	0	96%	FJ537196.1
34	4658	*Litsea machilifolia*	*Litsea glutinosa*	0	99%	KU382356.1
			Phoebe neurantha	0	99%	MH394355.1
35	4804	*Cryptocarya ferrea*	*Cryptocarya calcicola*	0	98%	KX546867.1
			Cryptocarya chinensis	0	98%	LC212965.1
			Alseodaphne andersonii	0	98%	KR528686.1
36	5044	*Litsea forstenii*	*Litsea baviensis*	0	98%	KJ439993.1
			Phoebe omeiensis	0	98%	KX437772.1
			Persea parvifolia	0	98%	JF966616.1

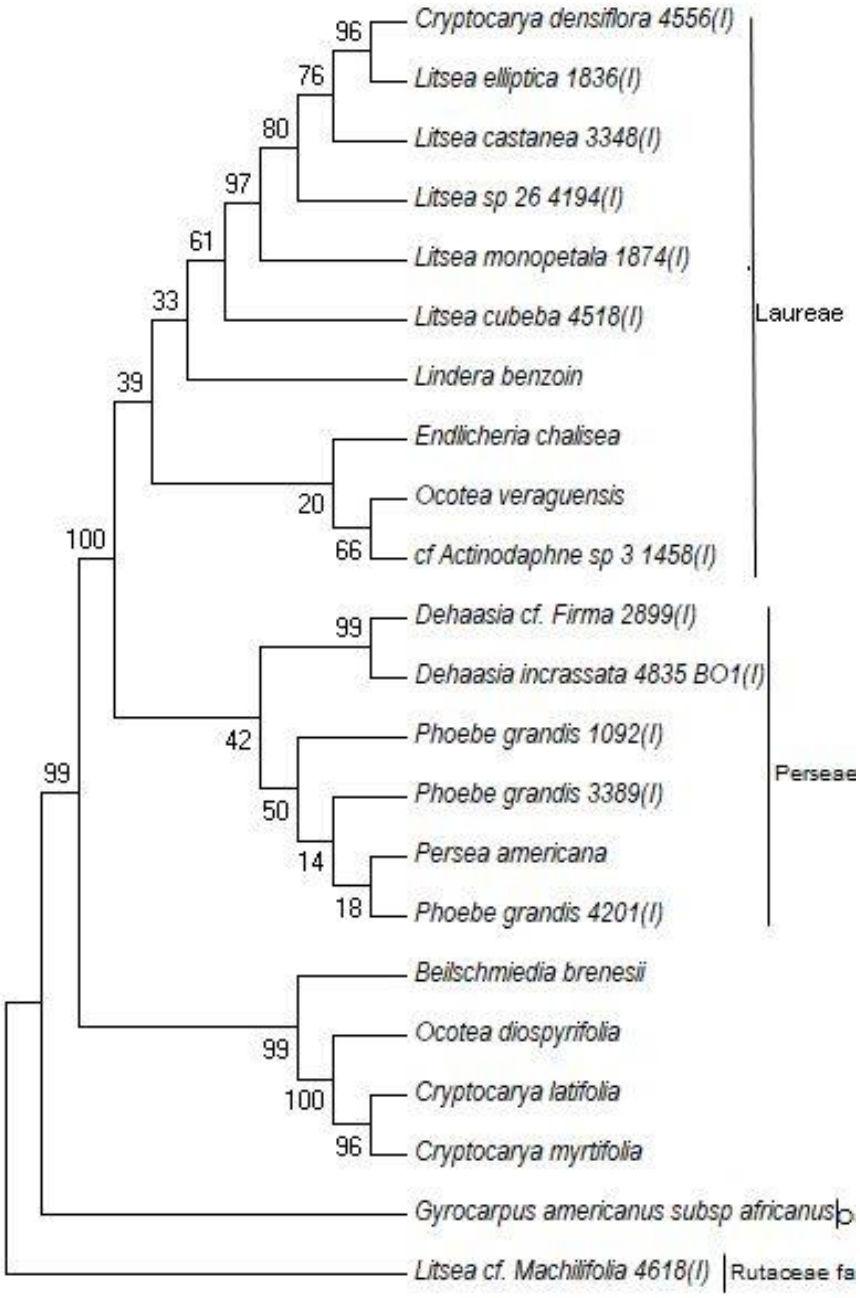
Annex 4: Phylogenetic tree constructed by Maximum likelihood method of the samples representing the Laureceae plant family based on *rbcL* gene sequences.



Annex 5: Phylogenetic tree constructed by Maximum likelihood method of the samples representing the Laureceae plant family based on *matK* gene sequences.



Annex 6: Phylogenetic tree constructed by Maximum likelihood method of the samples representing the Laureceae plant family based on *rbcL* and *matK* gene sequences.



STATUTORY DECLARATION

I hereby assure that this thesis is the result of my own work and investigations, except where otherwise stated .This work has not been submitted before to any other university for any kind of degree.

Signed…….(Lalit Kumar Dangol)
Date: