University of Padova

Department of Land, Environment Agriculture and Forestry

MSc in Mediterranean Forestry and Natural Resources Management

The use of barcoding sequences for the construction of phylogenetic relationships in the Euphorbiaceae

Supervisor:
Alessandro Vannozzi
Co-supervisor:
Prof. Dr. Oliver Gailing

Submitted by:
Bikash Kharel
Matriculation No. 1177536

ACADEMIC YEAR 2017/2018

Acknowledgments

This dissertation has come to this positive end through the collective efforts of several people and organizations: from rural peasants to highly academic personnel and institutions around the world. Without their mental, physical and financial support this research would not have been possible. I would like to express my gratitude to all of them who were involved directly or indirectly in this endeavor. To all of them, I express my deep appreciation.

Firstly, I am thankful to Prof. Dr. Oliver Gailing for providing me the opportunity to conduct my thesis on this topic. I greatly appreciate my supervisor Alessandro Vannozzi for providing the vision regarding Forest Genetics and DNA barcoding. My cordial thanks and heartfelt gratitude goes to him whose encouragements, suggestions and comments made this research possible to shape in this form. I am also thankful to Prof. Dr. Konstantin V. Krutovsky for his guidance in each and every step of this research especially helping me with the CodonCode software and reviewing the thesis.

I also want to thank Erasmus Mundus Programme for providing me with a scholarship for pursuing Master's degree in Mediterranean Forestry and Natural Resources Management (MEDFOR) course. Besides this, I would like to thank all my professors who broadened my knowledge during the period of my study in University of Lisbon and University of Padova.

Additionally, I am grateful to the EFForTS project for providing me opportunity to use the DNA data for Barcoding and further analysis. Many thanks to Hardiando Mangopa, Katja Rembold and JJ Afraistin for helping in morphological identification of samples. Heartfelt gratitude to the members of working group ZO2 who provided me guidance in each step throughout from beginning to now.

Deep appreciation for my friends.

At last but not the least, I want to pay gratitude to my family members who constantly inspired me during the study. I owe all my success to them.

Contents

List of Figures

List of Tables

Appendixes

Abstract

DNA barcoding has proven to be one of the successful method for the rapid identification of species. Identification of species through the use of DNA barcoding has become a new trend in the world of modern biology sciences. This method has been widely used in the different fields including floral exploration. However, unavailability of universal genes for all plant species, identification have been difficult. In spite of debates for using suitable gene for plant species, *matK* and *rbcL* are selected as the core barcodes for plants. The availability of molecular data as well as modern technologies in DNA sequencing have made DNA barcoding a popular process in many taxonomic studies. DNA barcoding is not a replacement to the traditional taxonomic classification but seen 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 Sumatra's Euphorbiaceae. The two core barcodes for the plant (*rbcL* and *matK*) were used as DNA barcodes for Euphorbiaceae. These two barcodes were evaluated based on their performance in identifying species and verification of morphological identification of Euphorbiaceae samples.

The study sites were located in Sumatra, Indonesia. Sixty-seven Euphorbiaceae leaves samples were collected from thirty-two plots distributed in four different land use system of Bukit Duabelas and the Harapan Landscape. thirty-two study plots distributed in two landscapes, Bukit Duabelas National Park and Harapan Rainforest. Morphological identification was done by taxonomist by comparison with reference vouchers at Herbarium Bogoriensis and BIOTROP herbarium Bogor, Indonesia. Dried-leaf specimens were collected and dried in silica-gel for the DNA analysis. For the extraction of DNA barcodes, all laboratory procedures have been done in Department of Forest Genetics and Forest Tree Breeding , University of Goettingen, Goettingen, Germany.

Sequence editing was carefully done using Codoncode software to each 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. *Drypetes littoralis* from Putranjivaceae family was chosen as an outgroup to root the phylogenetic trees.

The result of this study showed *rbcL* was easy to amplify and sequence than *matK.* Out of 67 samples, 58 samples for *rbcL* was successfully amplified and 52 samples were successfully

sequenced. While for *matK* only 40 and 33 samples were successfully amplified and sequenced respectively. The amplification success rate of rbcL was 86% and sequencing success rate was 78%. While the amplification and sequencing success rate of *matK* was 60% and 49% respectively.

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.

Using *rbcL*, *matK* and combination of both markers, six phylogenetic trees were constructed in this study using two different methods (Neighbor Joining and Maximum Likelihood). For the construction of phylogenetic trees, sequences obtained from the leaves samples and sequences downloaded from NCBI were used. Trees constructed by both the methods showed similar topologies. The rbcL, *matK* sequences from collected samples correctly clustered together with the Genbank sequences representing the same genera. However, only very few of sequences clustered together with the Genbank sequences representing same species. Some of the samples sequences which were morphologically identified as Euphorbiaceae samples clustered together with species of the Moraceae family. This shows they have been morphologically misclassified as Euphorbiaceae samples.

In conclusion*,* two barcode regions i.e. *matK* and *rbcL* were not satisfying in all manners. As the core barcodes, these two markers were effective to be used in plant species identification at least up to genus level and higher than genus level. The combination of *matK* and *rbcL*, however, was proven to have a higher level of discriminatory power.

1. Introduction

1.1 DNA barcoding

A DNA barcode is defined as short genomic sequence extracted from a standardized portion of genome (Walker, 2009), whereas DNA barcoding refers to the process applied for the quick identification of a species which is 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). Apart from species identification, DNA barcodes improve or supplement traditional taxonomy based on morphological characters (Hebert & Gregory, 2005). An ideal barcode must fulfil 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). 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).

The DNA barcode, "Mitochondrial gene cytochrome oxidase c subunit 1 (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 it across distantly related species (Hollingsworth et al., 2011).

1.2 Importance of DNA barcoding

There are approximately 8.7 million species on earth (Mora et al., 2011), out of which 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. 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). In addition, morphological species identification is time-consuming and unreliable.

On the one hand, there is a lack of professional taxonomists and morphological difficulties in species identification, on the other hand 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. DNA sequencing technologies and barcoding can help us solve the problem of fast and efficient species identification.

1.3 Plant molecular systematics

Molecular systematics is the approach of classification of organisms into related groups based on the molecular genetic data and molecular genetic markers representing organelle and nuclear genomes. During the past decades, molecular systematics has been successfully used for the phylogenetic analysis of evolutionary patterns and processes. Phylogenetic studies help us understand the species evolutionary relationships and this understanding can be applied in other related fields such as ecology, biogeography and conservation (Kreft & Jetz, 2010).

Molecular systematics employs a number of methods that uses macromolecules and molecular data to infer phylogenetic relationships. Phylogenetic relationships can be inferred from analysis of DNA sequences, DNA restriction sites, microsatellites, -allozymes, RAPDs and AFLPs (Simpson, 2010). Among molecular data, DNA sequence data are the most informative and enables more accurate arrangement of closely related species. The selection of appropriate DNA regions is considered critical for resolving the phylogenetic relationships (Soltis et al., 1998).

1.4 DNA sequence data

DNA sequence data refers to the sequence of four nucleotides (adenine= A, cytosine= C, guanine=G and thymine=T) in a given DNA regions of a particular organism, sample or taxon. The phylogenetic analysis uses information contained in the nuclear and orgaelle genomes. There are three types of DNA sequence data obtained from the DNA source. They are nuclear DNA (nDNA), mitochondrial DNA (mtDNA) and also chloroplast DNA (cpDNA) in plants

1.4.1 Nuclear DNA

Nuclear DNA (nDNA) is commonly used in phylogenetic and evolutionary studies. It is transmitted from parent to offspring through asexual or sexual reproduction (Simpson, 2010). nDNA is diploid or polyploid genome and located in the nucleus of eukaryotic organisms. It is biparentally inherited, thus provides more genetic information than the other two organelle genomes. The Internal transcribed spacer region (ITS) represent one of the most useful type of nDNA marker. It represents a nuclear genome region between 18S and 26S nuclear ribosomal DNA (nrDNA) genes. There are multiple copies of this region in a nuclear genome.

1.4.2 Mitochondrial DNA

Mitochondrial DNA (mtDNA) represents a small portion of DNA in eukaryotic cell of animal species. Mitochondria converts chemical energy from the food into the form that cells can use. It is maternally inherited in most organisms. The cytochrome c oxidase I gene (*COI* or *COXI*) is commonly used in the phylogenetic analysis, evolution biology studies and DNA barcoding of animals, but it is not very informative for the phylogeny study of plant species because of slower evolutionary rate in vascular plants (Kress et al,. 2005).

1.4.3 Chloroplast DNA (cpDNA)

Unlike mitochondrial DNA (cpDNA) is also inherited paternally in most angiosperms. Chloroplast is responsible for the photosynthesis in plants. Its genome has a long working history in testing the relationship between biological and geological phenomena in angiosperms (Kelchner, 2000). Its sequences are mostly used to study evolutionary patterns of plants (Raubeson and Jansen, 2005).

The genome is divided into three functional categories: exons, introns and intergenic regions. Introns and intergenic regions do not encode protein and are referred as noncoding (Shaw et al., 2005). The coding regions i.e. exons, evolve more slowly than the non-coding regions. Non-coding regions are used in molecular systematics, population genetic and phylo-geographic studies of plants. DNA sequences of non-coding cpDNA were first used in the construction of plant phylogenies (Taberlet et al., 1991). Out of many cpDNA markers representing non-coding region, the region between the *psbA* (en- codes photosystem II protein D1) gene and the *trnH (tRNA^{His})* gene is widely used at species level (Hao et al., 2010) while at the higher taxonomical level the *rbcL*, *matK*, *ndhF*, *atpB* and *rps2* genes and their introns are used to resolve phylogenetic relationships (Kim et al., 1999). In this research only *rbcL* and *matK* have been used for phylogenetic analysis.

1.4.3.1 *rbcL* barcoding region

The ribulose-1, 5-bisphosphate carboxylase large subunit (*rbcL*) is the most abundant enzyme in nature and is responsible for the autotrophy(Sen et al., 2011). It is encoded by the chloroplast *rbcL* gene and largely used in phylogenetic analysis in plants. It is well known for its tracing ability of evolutionary history of plants. For most land plants, the barcode region of *rbcL* can be easily amplified, sequenced and aligned but has relatively low discriminatory power (Hollingsworth et al., 2011). According to the various phylogenetic studies, *rbcL* was proven to be most suitable gene for the construction of evolutionary history at generic level but not at lower taxonomic level i.e. species level within same genus level (Soltis et al., 1998). For making it more useful in phylogenetic studies, it is suggested to combine *rbcL* with markers representing other regions (Vijayan and Tsou, 2010).

1.4.3.2 *matK* barcoding region

Maturase K (*matK*) is also one of the most frequently used barcode regions in phylogenetic studies. *matK* helps to successfully solves generic and species-level relationships because of its high discriminatory power (Hollingsworth et al., 2011). It is one of the rapidly evolving genes (Hilu & Liang, 1997) and is relatively closest analogue of *CO1* in animal barcoding. It is located within the intron of the chloroplast gene *trnK* (Vijayan and Tsou, 2010). It is one of most informative loci for the determination of phylogenetic relationships (Hilu et al., 2003).

1.5 Biodiversity in Sumatra

Southeast Asia has 4 biodiversity hotspots among 25 global biodiversity hotspots around the world. Countries like Indonesia, Malaysia and Philippines are megadiverse countries of the region (von et al., 2017). Indonesia harbors 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)). After Brazil, Indonesia have the second largest rainforest in the world (Hansen et al., 2009) and has 3% of world's total forest area (UN FAO, 2015).

Sumatra is the largest island in Indonesia and world's sixth largest island. 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 area in the world (Whitten et al., 2000). The biological diversity of tree species is extremely high in the Sumatran lowland forest.

1.5.1 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). The main reason for the mass clearing of the forests in the country was the expansion of palm oil cultivation and forest fire. In 2001, the World Bank reported that the loss of forest areas in Sumatra was estimable in 7 million hectares between 1985 to 1997. Between 2000 to 2012, about 1.21 million ha of lowland forest in Sumatra have been lost due to deforestation (Margono et al., 2014).

1.6 Euphorbiaceae plant family

Euphorbiaceae, the spurge plant family, is the fifth largest family of flowering plant. It is among one of the most diverse, large and complex family of angiosperms. This family is composed of over 340 genera and 8,000 species (Mwine & Van Damme, 2011). It is mostly distributed in the tropics, with most of the species in the Indo-Malayan region and tropical America (Rahman & Gulshana, 2014). It also has species in other non-tropical areas such as the Mediterranean basin, the middle east, Central Europe, South Africa and the Southern United States (Davis, Latvis, Nickrent, Wurdack, & Baum, 2007).

The Euphorbiaceae family has a remarkable variety of the growth forms which likely equal or surpass other angiosperm families (Halle, 1971). The leaves are alternate, rarely opposite with stipules. The flowers are terminal or axillary located solitary or in glomerulus (Webster, 1994b). The lowers are not colorful. They are folded by bracts, actinomorphic, achalydeous, monochlamydeous and rare dichlamydeous. The androecium shows one or many free stamens or in connate. The anthers are rimose dehiscence, rare poricide and may have nectariferous disc. The gynoecium is tricarpellary having uniovulate locus (Webster 1994b, Radcliffe-Smith, 2001).

Euphorbiaceae contains a huge variety of phytotoxins including alkaloids, glycosides, diterpenes and ricin-like toxins (Davis et al., 2007). Subfamilies Euphorbioideae and Crotonoideae have milky latex which is the main characteristics feature of these subfamilies. Latex is poisonous in Euphorbioideae while innocuous in Crotonoideae (Davis et al., 2007).

1.6.1 Classification and use of DNA barcoding in Euphorbiaceae

The major milestone in the classification of Euphorbiaceae was the classification by Adrien Jussieu(1824), who identified major series of genera and Jean Mueller who first provided detailed classification of the family into subfamilies, tribes and sub-tribes (Webster, 1994). The classification of Euphorbiaceae in 1866 was the milestone for the Euphorbiaceae classification (Webster, 1975). According to Webster (1975), Mueller was the first person to use coherent phylogenetic characteristics. Making Mueller's classification as a reference, Webster in 1975 classified Euphorbiaceae into five subfamilies i.e. Acalyphoideae, Crotonoideae, Euphorbioideae, Phyllanthoideaae and Oldfieldoiideae. The first three subfamilies are uni-ovulate whereas the last

two are bi-ovulate. This classification was done based on the number of ovules per locule, the presence of lacticifers and pollen grain morphology (Webster, 1975).

There was a continuous pressure and proposals for the modification of the family boundaries of the large and diverse Euphorbiaceae family. Split of uniovulate from biovulate taxa based on characters of seed coat, was advocated by Corner (1976) and Huber (1991). Molecular phylogenetic and phytochemical evidences confirmed the non-monophyly of Euphorbiaceae (Seigler, 1944b, Tokuoka and Tobe, 1995). Based on these results, Euphorbiaceae were separated into five families, the uni-ovulate subfamilies were included in family Euphorbiaceae *sensu lato [s.l.]* and other subfamilies of bi-ovulate were added into other families or grouped into their own families (Webster, 1994). The bi-ovulate subfamilies Oldfieldoideae and Phyllanthoideae formed families Phyllanthaceae, Picrodendraceae and Putraanjivaceae. The oniovulate plants i.e. Acalyphoideae, Crotonoideae, Euphorbioideae was included in families Pandaceae and Euphorbiaceae sensu stricto [s.s]. This classification was based on new molecular studies and molecular plastids *rbcL*, atpB, *matK*, 18S rDNA, and *trnl-f* markers and the nuclear PHYC gene (Tokuoka, 2007; Wurdack et al., 2005).

According to the most recent molecular phylogeny based classification (APG IV, 2016), family Euphorbiaceae s.s. has 3 subfamilies (Euphorbioideae, Crotonoideae and Acalyphoideae), consisting of 6,300 species in 247 genera (Wurdack et al., 2005). *Euphorbia L.* (Euphorbioideae), *Croton L.* (Crotonoideae) and *Acalypha* L. (Acalyphoideae) are the largest genera of the Euphorbiaceae s.s. family, consisting of 2,100, 1200 and 600 species respectively (Webster, 1994; Radcliffe-Smith, 2001; Carneiro-Torres et al., 2011). The species occurs in diverse growth forms like trees, shrubs, subshrubs and herbs (Barroso et al., 1991, Sousa and Lorenzi, 2006).

1.7 The EFForTS-Project

1.7.1 Background

This master's dissertation was conducted under 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)

1.7.2 Project Objective

Based on three major lines of research (i) environmental processes, (ii) biota and ecosystem services, and (iii) human dimensions(Drescher et al., 2016)), this project has set its major objective as to facilitate in-depth understanding of the consequences of rainforest transformation to functional diversity of that area.

1.7.3 Plot design

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 to ecological dimension while socioeconomic survey design is 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).

2. Rationale

There is a considerable debate regarding the choice for DNA barcoding genetic markers for the land plants. Different studies propose different barcode regions suitable for plant species. In the study by Kress and Erickson (2007), various coding and non-coding regions in nuclear and plastid genomes were suggested for the potential barcodes of plants. In 2009, a group of researchers belonging to the "Consortium for the Barcode of Life (CBOL) Plant Working Group", recommended two chloroplast loci (*rbcL* and *matK*) as a standard barcode set for plant DNA (CBOL Plant Working Group 2009). However, the proposed genes had various amplification problems in some families (Hollingsworth et al., 2009). In our study research presented here, two most widely used DNA barcodes loci *rbcL* and *matK* were used for barcoding of the samples collected in the above described plots and morphologically classified as belonging to the Euphorbiaceae plant family.

In the long history of species identification and classification, taxonomy has been mainly based on the morphological structures and phenotypic traits of individual organisms (Hebert & Gregory, 2005). The knowledge from traditional taxonomy has been used in every aspect of the current studies of the Earth's biodiversity (Stepanović et al., 2016). species identification has been a fundamental problem in the modern biology. The use of the DNA barcoding has great potential for a better taxonomic resolution and understanding the evolutionary history of the species. Various studies have shown that traditional taxonomy and DNA barcoding have been used for resolving the phylogeny, taxonomy and nomenclature of living organisms. The second aim of this study is to identify the morphologically unidentified samples and also compare the identified samples of Euphorbiaceae collected in Sumatra using BLASTn algorithm.

Finally, phylogenetic trees have been constructed for the comparison between molecular and morphological identification. This research will make use of two barcode regions *rbcL* and *matK* for the construction of phylogenetic tree using two methods i.e. Neighbor joining and Maximum likelihood method.

3. Objectives

The general objective is to use DNA barcodes for Euphorbiaceae species identification and establish their phylogenetic relationships. To achieve this objective, the following tasks have been completed:

General Objectives

i. Assessment of barcode universality and resolution power for species identification in Sumatra's Euphorbiaceae samples,

ii. Evaluate DNA barcoding performance in species identification,

iii. Construct phylogenetic relationships to compare molecular and morphological identification.

4. Materials and methods

4.1 Study area

The study was conducted in two landscapes in the Jambi Province- Sumatra, Indonesia namely Bukit Duabelas National Park and the Harapan Rainforest. The Bukit Duabelas National Park lies in the center of Jambi province. It is a small national park with an area of 605 km^2. The area of the park is mainly covered by secondary forest while the northern part consists of primary rainforest. The topography of the park varies from flat land (164 meter in altitude) to slightly hilly area (438 meters in altitude). On the other hand, Harapan forest covers 98,555 ha of rainforest in Jambi Province. The forest is one of the most biodiverse 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)

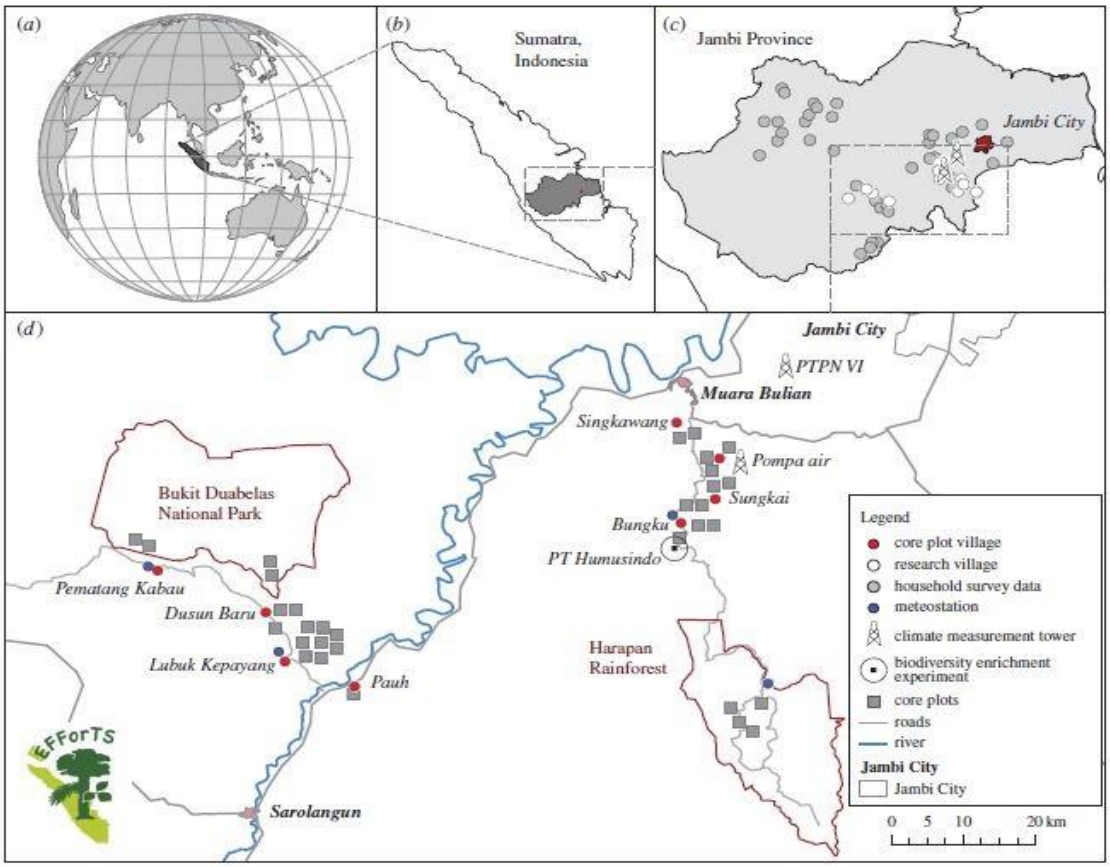


Figure 4.1: Map showing two study sites: Bukit Duabelas and Harapan respectively (Drescher et al., 2016)

4.2 Study plots

In each landscape, four core plots were established. The core plots represent the four different land use system (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-plots of 5*5 m.

4.3 Sample collection

From all the study plots, samples were collected. 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 in triplicate. After that, leaf tissues of approx. 2cm^2 size were collected and dried in silica-gel for DNA analysis. 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. The following table shows the samples collected from the various core plots and plots.

Table 4.1 List of samples collected

S.N	Sample ID	Core plots	Plots	Sub plots	Field name	Family
1	214	Oil palm	BO3	BO3a	Macaranga sp. 02	Euphorbiaceae
2	237	Oil palm	Straße vor BO3	-	Macaranga sp. 03	Euphorbiaceae
3	269	Oil palm	BO3	BO3c	Euphorbiaceae sp. 03	Euphorbiaceae
4	270	Oil palm	BO3	BO3c	Euphorbiaceae sp. 03	Euphorbiaceae
5	377	Oil palm	BO3	BO3e	Macaranga sp. 02	Euphorbiaceae
6	1265	Oil palm	BO5	-	cf. Mallotus sp. 02	Euphorbiaceae
7	1290	Oil palm	BO5	BO5c	Euphorbiaceae sp. 12	Euphorbiaceae
8	1303	Oil palm	BO5	BO5d	Euphorbiaceae sp. 13_1	Euphorbiaceae
9	1304	Oil palm	BO5	BO5d	Euphorbiaceae sp. 13_1	Euphorbiaceae
10	1328	Jungle rubber	BJ5	-	Macaranga "alba"	Euphorbiaceae
11	1335	Jungle rubber	BJ5	-	Macaranga cf. lowii	Euphorbiaceae
12	1336	Jungle rubber	BJ5	-	Macaranga cf. lowii	Euphorbiaceae
13	1341	near BJ5	near BJ5	-	Mallotus paniculatus	Euphorbiaceae
14	1369		near BJ5	-	Mallotus sp. 03	Euphorbiaceae
15	1370		near BJ5	-	Mallotus sp. 03	Euphorbiaceae
16	1394	Jungle rubber	BJ5	-	Macaranga "alba" 02	Euphorbiaceae
17	1408	Jungle rubber	BJ5	-	Macaranga "alba" 01	Euphorbiaceae
18	1608	Jungle rubber	BJ3	-	Macaranga sp. 07	Euphorbiaceae
19	1610	Jungle rubber	BJ3	-	Macaranga sp. 07	Euphorbiaceae
20	1611	Jungle rubber	BJ3	-	Macaranga "alba" 02	Euphorbiaceae
21	1635	Jungle rubber	BJ3	-	Macaranga sp. 08	Euphorbiaceae
22	1638	Jungle rubber	BJ3	-	Croton cf argyratus	Euphorbiaceae
23	1645	Jungle rubber	BJ3	-	Croton cf argyratus	Euphorbiaceae
24	1653	Jungle rubber	BJ3	-	Homalanthus sp. 01	Euphorbiaceae
25	1659	Jungle rubber	BJ3	-	Mallotus sp. 04	Euphorbiaceae

26	1697	Jungle rubber	BJ3	BJ3b	Euphorbiaceae sp. 17	Euphorbiaceae
27	1720	Jungle rubber	BJ3	BJ3c	Mallotus sp. 04	Euphorbiaceae
28	1728	Jungle rubber	BJ3	BJ3c	Croton sp. 2	Euphorbiaceae
29	1748	Jungle rubber	BJ3	BJ3d	Croton sp.2	Euphorbiaceae
30	1847	Jungle rubber	BJ4	-	Croton sp. 03	Euphorbiaceae
31	1905	Jungle rubber	BJ4	-	Macaranga triloba	Euphorbiaceae
32	1947	Jungle rubber	BJ4	BJ4a	Croton sp. 03	Euphorbiaceae
33	1961	Jungle rubber	BJ4	BJ4a	Mallotus sp. 05	Euphorbiaceae
34	1962	Jungle rubber	BJ4	BJ4a	Mallotus sp. 05	Euphorbiaceae
35	1978	Jungle rubber	BJ4	BJ4a	Euphorbiaceae sp. 20	Euphorbiaceae
36	2112	Jungle rubber	BJ4	BJ4c	Antidesma sp. 11	Euphorbiaceae
37	2113	Jungle rubber	BJ4	BJ4c	Antidesma sp. 11	Euphorbiaceae
38	2156	Forest	BF1	-	Macaranga sp. 09	Euphorbiaceae
39	2621	Jungle rubber	BJ6	-	Homolanthus sp. 01	Euphorbiaceae
40	2622	Jungle rubber	BJ6	-	Homolanthus sp. 01	Euphorbiaceae
41	2626	Jungle rubber	-	-	Macaranga sp. 10	Euphorbiaceae
42	2628				Croton sp. 04	Euphorbiaceae
43	2653	Jungle rubber	BJ6	BJ6a	Croton sp. 05	Euphorbiaceae
44	2678	Jungle rubber	BJ6	BJ6b	Macaranga sp. 11	Euphorbiaceae
45	2679	Jungle rubber	BJ6	BJ6b	Macaranga sp. 11	Euphorbiaceae
46	2693	Jungle rubber	BJ6	BJ6c	Croton sp. 05	Euphorbiaceae
47	2705	Jungle rubber	BJ6	BJ6c	Mallotus sp. 06	Euphorbiaceae
48	2709	Jungle rubber	BJ6	BJ6c	Croton sp. 05	Euphorbiaceae
49	2719		near BJ6	-	Croton sp. 05	Euphorbiaceae
50	2720		near BJ6	-	Mallotus sp. 06	Euphorbiaceae
51	2721		near BJ6	-	Mallotus sp. 06	Euphorbiaceae
52	3187	Rubber	HR4	HR4c	Euphorbiaceae sp. 25	Euphorbiaceae
53	3335	Jungle rubber	HJ4	-	Croton sp. 06	Euphorbiaceae

54	3600	Jungle rubber	HJ3	HJ3c	Macaranga sp. 12	Euphorbiaceae
55	3659	Jungle rubber	HJ2	-	Macaranga sp. 12	Euphorbiaceae
56	3761	Jungle rubber	HJ2		Endospermum cf. diademum	Euphorbiaceae
57	3873	Oil palm	HO1	HO1d	Euphorbiaceae sp. 26	Euphorbiaceae
58	4063	Forest	HF1	-	Euphorbiaceae sp. 28	Euphorbiaceae
59	4083	Forest	HF1	-	Pimelodendron zoanthogyne	Euphorbiaceae
60	4128	Forest	HF1	-	Tree 90	Euphorbiaceae
61	4324	Forest	HF1	HF1b	Tree 90	Euphorbiaceae
62	4421	Forest	HF1	HF1d	Croton sp. 07	Euphorbiaceae
63	4422	Forest	HF1	HF1d	Croton sp. 07	Euphorbiaceae
64	4661	Forest	HF2	-	Croton argyratus	Euphorbiaceae
65	4731	Forest	HF2	HF2c	Macaranga trichocarpa	Euphorbiaceae
66	4785	Forest	HF2	HF2e	Croton cascarilloides	Euphorbiaceae
67	5185	Forest	HF4	HF4b	Euphorbiaceae sp. 29	Euphorbiaceae

4.4 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.

4.5 Laboratory methods

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

4.5.1 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.

4.5.2 Polymerase Chain Reaction (PCR), DNA amplification

The *rbcL* and *matK* markers were amplified using polymerase Chain reaction (PCR) and the universal primers used for the amplification are listed in Table 4.2 below:

Table 4.2 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	*MatK*newF	GTTCAAACTCTTCGCTACTGG	(Kress et al., 2009), (Yu et al., 2011)
	*MatK*newR	GAGGATCCACTGTAATAATGAG	
	3FKim(*matK*)	CGTACAGTACTTTTGTGTTTACGAG	
	1Rkin(*matK*)	ACCCAGTCCATCTGGAAATCTTGGTTC	

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.

Table 4.3: Reaction mixture of PCR reagents

Reagents	Volume (15 ul)
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 below in Table 4.4:

Table 4.4: PCR protocol

Steps	Condition
Step 1	Denaturation at 95^{o}Cfor 15 minutes
Step 2	35 cycles of Denaturation at 94^{o}c for 1 minute Annealing at 50^{o}C for 1 min Extension at 72^{o}C for 1:30 minutes
Step 3	Final extension at 72^{o}C 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.

4.5.3 DNA sequencing

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

Sequencing reactions were performed using the ABI Prism™ 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 4.5 and 4.6 respectively:

Table 4.5: Sequencing reaction mixture

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

Table 4.6: Sequencing reaction protocol

Step	Condition
1	Initial denaturalization for 1 min at 96^{o}C
2	34 cycles of • Denaturation for 10 minutes at 96^{o}C • Annealing for 10 minutes at 45^{o}C • Elongation for 4 minutes at 60^{o}C
3	Final extension for 20 minutes at 72^{o}C

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 number of repetitions to obtain successful sequences were excluded.

4.6 DNA Sequence analysis

CodonCode Aligner™ software was used to align sequences and edit them by trimming the bad quality nucleotides at the ends of the forward and reverse sequences. The 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 consequences 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.

Sequences obtained from the collected samples and sequences downloaded from the NCBI Genbank were aligned for each markers. 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 make the equal length multiple sequence alignments. The aligned *rbcL* and *matK* sequences were concatenated for the same samples using SequenceMatrix software(Vaidya et al., 2011) and then the concatenated alignments were exported as NEXUS files.

4.7 Sequence from the NCBI GenBank

Sequence data from related species of the Euphorbiaceae 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 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 alignment (Edgar, 2004)

4.8 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 using NCBI website and the *rbcL* and *matK* sequences from collected the 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.

4.9 Phylogenetic trees

Phylogenetic trees were reconstructed, based on the aligned sequences from the laboratory (marked with X) 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. *Drypetes littoralis* from the Putranjivaceae family was chosen as an outgroup species to root the phylogenetic trees. To facilitate accurate alignment, species from the Putranjivaceae family was chosen as outgroup because of its relatively low sequence divergence from Euphorbiaceae (Tokuoka, 2007; Tokuoka & Tobe, 2006)

The Neighbor joining and maximum likelihood methods were used to genrate phylogenetic trees. 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 tree which the highest likelihood probability of explaining the sequence relationships (Felsenstein, 1981). For both the trees bootstrap support was computed using 1000 replicates.

5. Results

5.1 Morphological classification of the samples

The morphological classification of the collected samples was done by the professional taxonomists. Out of 67 Euphorbiaceae samples, it was not possible to morphologically identify 8 samples (see Appendix 1). The morphological classification assigned samples to 3 subfamilies (Acalyphoideae, Crotonoideae and Euphorbioideae) and 10 genera in the Euphorbiaceae family.

For further analysis, each sequence was named according to the morphological classification name with belief that identification was accurate since the herbarium vouchers were carefully matched and compared with reliable reference vouchers. For samples that were impossible to classify morphologically, field collection ID name was used.

Table 5.1: Composition of Euphorbiaceae plant family samples

Sub-family	Genus	Number of samples
Acalyphoideae	*Mallotus*	6
	Macaranga	19
	Alchornea	6
	Melanolepis	3
	Cephalomappa	2
	Neoscortechinia	3
Crotonoideae	*Croton*	15
	Endospermum	1
Euphorbioideae	*Balakata*	3
	Pimelodendron	1

5.2 DNA sequence characteristics

DNA was extracted relatively from the dried-leaf specimens. The PCR amplification and sequencing of the *matK* marker was more difficult than the *rbcL*. For the *rbcL* marker amplification success rate was 89% and sequencing success rate was 83%, while for the *matK* the amplification and sequencing success rate were only 62% and 52% respectively (Table 5.2)

A total of 83 *rbcL* and 65 *matK* sequences including 2 sequences of the outgroup were downloaded from NCBI. For Euphorbiaceae, *matK* sequences were also less available in the GenBank database than *rbcL* sequences. There was difference in alignment length for the both markers. The alignment length for *rbcL* was 510 bp while for *matK* 641 bp.

Table 5.2: Sequence data for the *rbcL* and *matK markers*

Parameters	*rbcL*	*matK*
Number of sequence downloaded from NCBI GenBank database	83	67
Number of samples used for amplification and sequencing	67	67
Number of successfully amplified samples	58	40
Amplification success rate	86%	60%
Number of successfully sequenced samples	52	33
Sequencing success rate	78%	49%
Aligned length	510 bp	613 bp

5.2.1 *rbcL* barcoding marker

The amplification of this region was successful for most of the leaf samples. In total, 67 leaf samples were used for PCR amplification and sequencing. Among them, 58 samples (86%) were successfully amplified and only 52 (78%) were successfully sequenced. In total, together with sequences downloaded from GenBank 135 rbcL sequences were used for further analysis. The final length of the multiple sequence alignment after the manual editing was 510 bp.

5.2.2 matK barcoding marker

The amplification and sequencing success of the *matK* marker was worse compared to the *rbcL* marker. Among 67 leaf samples, only 40 samples (60%) were successfully amplified and 33

samples (49%) were successfully sequenced. So, finally only 29 sequences were used for the further analysis. The alignment together with 67 sequences downloaded from the NCBI GenBank resulted in total of 96 samples in the 641 bp long multiple sequence alignment.

5.3 Identification and Barcode analysis using Nucleotide BLAST (BLASTn)

5.3.1 Analysis of unidentified samples

Out of 8 unidentified samples, only 4 and 5 samples were successfully amplified and sequenced for *rbcL* and *matK* respectively. The molecular identification these unidentified samples was attempted using *rbcL*, *matK* and combination of both markers (Table 5.3). The combined use of both markers was successful in finding the best homologous sequences with identity close to 100% and E-value equal to 0.0 using BLASTn. More in detail, sample corresponding to IDs 1728, 1748 and 2628) were identified as *Vernicia fordii* with E-value equal to 0.0 and identical match equal to 98%. Less clear was the molecular identification of sample ID 4421 using the combination of both rbcL and *matK*, with different species belonging to genus *Croton* showing the same identity and E-values.

The molecular identification based on *rbcL region* produced many ambiguous results with different species showing the same identity percentages and E-values (0.0). However, it's worth to mention that the barcode showed a high discriminatory power in the identification of sample ID 1748 as *Vernicia fordii.* All the other query results for *rbcL* were quite ambiguous, not only failing in species but even in genera discrimination. Most of the result showed variation in the genera for a sample. For example: the query of *rbcL* sequence of Croton sp. 2 (sample ID: 1728) and query of ID 2628, showed species from *Trigonostemon* genus.

MatK was also not reliable in finding the best hit for the species identification. It produced ambiguous results. The identical match percentage was around 100% and E-value was 0.0.

Table 5.3 shows the results based on the highest scoring hits of BLASTn. Multiple scoring hits of the same query sequence with different taxa represent ambiguous results. Sequence ID number 377 (Macaranga sp 2) was successfully sequenced only for the *rbcL* marker.

Table 5.3: Species identification and verification using rbcL and *matK* barcode marker

Sample ID	Field name	*rbcL*			*matK*			*matK+rbcL*		
		Best matches	E-value	Identity	Best matches	E-value	Identity	Best matches	E-value	Identity
377	Macaranga sp. 02				Coccoceras muticum	0.0	98%			
					Mallotus cumingii	0.0	98%			
					Mallotus sp. JH-2017	0.0	98%			
					Mallotus leucocalyx	0.0	98%			
1728	Croton sp. 2	Trigonostemon bonianus	0.0	99%	Trigonostemon sp. KYUM-2014	0.0	99%	Vernicia fordii	0.0	98%
		Trigonostemon thyrsoideus	0.0	99%	Trigonostemon thyrsoideus	0.0	99%			
		Trigonostemon verrucosus	0.0	99%	Trigonostemon bonianus	0.0	99%			
1748	Croton sp.2	Vernicia fordii	0.0	98%	Trigonostemon thyrsoideus	0.0	99%	Vernicia fordii	0.0	98%
					Trigonostemon bonianus	0.0	99%			
					Trigonostemon sp. KYUM-2014	0.0	99%			
2628	Croton sp. 04	Trigonostemon thyrsoideus	0.0	99%	Trigonostemon sp. KYUM-2014	0.0	99%	Vernicia fordii	0.0	98%
		Trigonostemon bonianus	0.0	99%	Trigonostemon bonianus	0.0	99%			
			0.0	99%	Trigonostemon thyrsoideus	0.0	98%			
4421	Croton sp. 07	Croton tiglium	0.0	100%	Croton sp. 2 XCH-2015		98%	Croton sp. 2 XCH-2015	0.0	99%
		Croton	0.0	10	Croton laevifolius	0.0	98%	Croton	0.0	99%

	megalobotrys		0%				laevifolius		
	Croton gratissimus	0.0	100%	Croton verreauxii	0.0	98%	Croton tiglium	0.0	99%
	Croton zambesicus	0.0	100%				Croton verreauxii	0.0	99%
							Croton sylvaticus	0.0	99%

5.3.2 Barcode analysis for *rbcL* and *matK* markers

5.3.2.1 *rbcL*

Based on the BLASTn result or 48 *rbcL* sequences, 3 sequences (6%) from *Alchornea tiliifolia* (sample ID 1961, 1962 and 1978) found best match with sequences of same species, thus confirming the morphological identification. Thirty-nine sequences (81%) found best match with the species of same genus.

Some of the queries (6 query sequences corresponding to 12.5%) found best match from species belonging to different genera. Those sequences belonged to *Balakata baccata* and *Melanolepis multiglandulosa* which found best match with the species belonging to the genera *Triadica* and *Croton* respectively.

Two query sequences (4.2%) of *Endospermum diadenum* (sample ID 3761) and *Macaranga conifera* (sample ID: 3659) found their best match with the species belonging to Moraceae family.

Most of the query sequences showed ambiguous results while E-value for all the sequences was 0.0. The results of the BLASTn of *rbcL* sequences are reported in Appendix 2.

5.3.2.2 matK

The BLASTn analysis of *matK* sequences from 23 samples excluding morphologically unidentified samples, found that 17 *matK* query sequences (73%) were correctly assigned to correct genus. However, 6 query sequences (26%) matched the sequences representing different genera.

Macaranga trichocarpa (sample ID 3187*)* was successful in finding best matches with the sequences representing same species. While none of the other query sequences found the best match with the same species.

Query sequences of *Croton oblongus (*sample ID 2693 and 2719) *and Croton leiophyllus* (sample ID 3335) obtained the same ambiguous match result with E-value 0.0 and identity 100%. The query sequence of *Macaranga javanica* (sample ID 237) had E-value 0.0 and 100% identical match with species *Macaranga sp. JH-2017.* All the query sequences showed low E-value (0.0) and high identity percentage varying from 98% to 100%.

Table 5.3.2: The homologous sequences best matching the *matK* sequences based on the BLASTn analysis

S.N	Sample ID	Morphologically classified name	NCBI GenBANK best match	E-value	identity
1	237	*Macaranga javanica*	*Macaranga sp. JH-2017*	0	99%
			Macaranga sampsonii	0	99%
			Macaranga griffithiana	0	99%
			Macaranga hosei	0	99%
			Macaranga trichocarpa	0	99%
			Macaranga kurzii	0	99%
2	269	*Croton hirtus*	*Croton jutiapensis*	0	98%
3	270	*Croton hirtus*	*Croton jutiapensis*	0	99%
4	1608	*Macaranga gigantea*	*Macaranga sp. JH-2017*	0	99%
			Macaranga sampsonii	0	99%
			Macaranga griffithiana	0	99%
			Macaranga hosei	0	99%
			Macaranga gigantea	0	99%
			Macaranga inamoena	0	99%
			Macaranga hypoleuca	0	99%
5	1610	*Macaranga gigantea*	*Macaranga sp. JH-2017*	0	100%
6	1611	*Macaranga hosei*	*Macaranga sp. JH-2017*	0	99%
			Macaranga sampsonii	0	99%
			Macaranga griffithiana	0	99%
			Macaranga hosei	0	99%
			Macaranga andamanica	0	99%
7	1635	*Macaranga conifera*	*Macaranga sp. JH-2017*	0	99%
			Macaranga griffithiana	0	99%
			Macaranga andamanica	0	99%

			Macaranga sampsonii	0	99%
			Macaranga hosei	0	99%
			Macaranga trichocarpa	0	99%
			Macaranga kurzii	0	99%
			Macaranga inamoena	0	99%
8	1638	*Croton cascarilloides*	*Croton sp. 2 XCH-2015*	0	99%
			Croton laevifolius	0	99%
			Croton tiglium	0	99%
			Croton sylvaticus	0	99%
9	1645	*Croton cascarilloides*	*Croton sp. 2 XCH-2015*	0	99%
			Croton laevifolius	0	99%
			Croton tiglium	0	99%
			Croton sylvaticus	0	99%
			Croton sp. PM5533	0	99%
10	1653	*Balakata baccata*	*Triadica cochinchinensis*	0	99%
			Triadica sebifera	0	99%
11	1697	*Croton caudatus*	*Croton sylvaticus*	0	99%
			Croton sp. PM5220	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton laevifolius	0	99%
			Croton tiglium	0	99%
			Croton sp. FU-2528	0	99%
			Croton sylvaticus	0	99%
			Croton megalobotrys	0	99%
			Croton yanhuii	0	99%
			Croton kongensis	0	99%
			Croton crassifolius	0	99%
12	1847	*Croton argyratus*	*Croton tiglium*	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton laevifolius	0	99%
			Croton sylvaticus	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99
			Croton kongensis	0	99%
			Croton verreauxii	0	99%
			Croton cascarilloides	0	99%
			Croton megalobotrys	0	99%
			Croton lachnocarpus	0	99%
			Croton gratissimus	0	99%
13	2621	*Balakata baccata*	*Triadica cochinchinensis*	0	99%
14	2622	*Balakata baccata*	*Triadica cochinchinensis*	0	98%
15	2679	*Mallotus peltatus*	*Mallotus brachythyrsus*	0	96%
16	2693	*Croton oblongus*	*Croton sp. 2 XCH-2015*	0	100%

			Croton laevifolius	0	100%
17	2705	*Melanolepis multiglandulosa*	*Croton sylvaticus*	0	99%
			Croton sp. PM5533	0	99%
			Croton tiglium	0	99%
			Croton sp. 2 XCH-2015	0	99%
18	2719	*Croton oblongus*	*Croton sp. 2 XCH-2015*	0	100%
			Croton laevifolius	0	100%
19	2720	*Melanolepis multiglandulosa*	*Croton sylvaticus*	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton tiglium	0	99%
			Croton sp. 2 XCH-2015	0	99%
20	2721	*Melanolepis multiglandulosa*	*Croton sylvaticus*	0	99%
			Croton sp. PM5533	0	99%
			Croton tiglium	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton sp. PM5220	0	99%
21	3187	*Macaranga trichocarpa*	*Macaranga trichocarpa*	0	99%
			Macaranga sampsonii	0	99%
			Macaranga sp. JH-2017	0	99%
22	3335	*Croton leiophyllus*	*Croton sp. 2 XCH-2015*	0	100%
			Croton laevifolius	0	100%
23	3659	*Macaranga conifera*	*Macaranga sp. JH-2017*	0	99%
			Macaranga sampsonii	0	99%
			Macaranga griffithiana	0	99%
			Macaranga hosei	0	99%
			Macaranga andamanica	0	99%
			Macaranga trichocarpa	0	99%
			Macaranga kurzii	0	99%
			Macaranga gigantea	0	99%
			Macaranga aleuritoides	0	99%

5.3.2.3 Combination of *rbcL* and *matK*

The best BLASTn matches for the joint query sequences representing 23 samples of morphologically classification samples, was not of the same species. For examples: the query sequences of *Balakata baccata* matched the best GenBank sequencing representing *Triadica cochinchinensis*, the query sequence of *Mallotus peltatus* matched the best *Macaranga sp. JH-2017*

and query sequence of *Croton hirtus* matched the best with *Croton jutiapensis*. The query sequences of *Macaranga conifera, Macaranga gigantea* and *Macaranga hosei* matched the best sequences from two species i.e. *Macaranga sp. JH-2017* and *Macaranga griffithiana* having the same E-value (0.0) and identical percentage of 100%. All the query sequences of Croton genus (except *Croton hirtus*) showed ambiguous match despite E-value equals to 0.0 and identical percentage of 99%. The query sequence of *Mallotus peltatus and Melanolepis multiglandulosa* showed best match with the species from the different genus i.e. *Macaranga* and *Croton* respectively with low E-value (0.0) and 98% to 99% identical match. Except 5 query sequences, all other query sequences showed ambiguous match with the sequences in the NCBI database. None of the query sequences found best match with the same species.

Table 5.3.2.2: The homologous sequences best matching the rbcL and *matK* sequences based on the BLASTn analysis

S.N	Sample ID	Morphological identified name	NCBI Genbank best match	E value	Identity
1	237	*Macaranga javanica*	*Macaranga sp. JH-2017*	0	99%
			Macaranga griffithiana	0	99%
			Macaranga andamanica	0	99%
			Macaranga sampsonii	0	99%
			Macaranga trichocarpa	0	99%
2	269	*Croton hirtus*	*Croton jutiapensis*	0	99%
3	270	*Croton hirtus*	*Croton jutiapensis*	0	99%
4	1608	*Macaranga gigantea*	*Macaranga sp. JH-2017*	0	100%
			Macaranga griffithiana	0	100%
5	1610	*Macaranga gigantea*	*Macaranga sp. JH-2017*	0	100%
			Macaranga griffithiana	0	100%
6	1611	*Macaranga hosei*	*Macaranga sp. JH-2017*	0	100%
			Macaranga griffithiana	0	100%
7	1635	*Macaranga conifera*	*Macaranga sp. JH-2017*	0	100%
			Macaranga griffithiana	0	100%
8	1638	*Croton cascarilloides*	*Croton tiglium*	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton laevifolius	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton kongensis	0	99%
9	1645	*Croton cascarilloides*	*Croton tiglium*	0	99%

			Croton sp. 2 XCH-2015	0	99%
			Croton laevifolius	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton kongensis	0	99%
10	1653	*Balakata baccata*	*Triadica cochinchinensis*	0	100%
11	1697	*Croton caudatus*	*Croton sylvaticus*	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton tiglium	0	99%
			Croton megalobotrys	0	99%
12	1847	*Croton argyratus*	*Croton tiglium*	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton laevifolius	0	99%
			Croton sylvaticus	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
13	1947	*Croton argyratus*	*Croton tiglium*	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton laevifolius	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton kongensis	0	99%
14	2621	*Balakata baccata*	*Triadica cochinchinensis*	0	99%
15	2622	*Balakata baccata*	*Triadica cochinchinensis*	0	98%
16	2679	*Mallotus peltatus*	*Macaranga sp. JH-2017*	0	98%
			Macaranga hosei	0	98%
17	2693	*Croton oblongus*	*Croton sp. 2 XCH-2015*	0	100%
18	2705	*Melanolepis multiglandulosa*	*Croton sylvaticus*	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton tiglium	0	99%
			Croton megalobotrys	0	99%
19	2719	*Croton oblongus*	*Croton sp. 2 XCH-2015*	0	100%
20	2720	*Melanolepis multiglandulosa*	*Croton sylvaticus*	0	99%
			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton tiglium	0	99%
			Croton megalobotrys	0	99%
21	2721	*Melanolepis multiglandulosa*	*Croton sylvaticus*	0	99%

			Croton sp. PM5533	0	99%
			Croton sp. PM5220	0	99%
			Croton tiglium	0	99%
			Croton megalobotrys	0	99%
22	3335	*Croton leiophyllus*	*Croton sp. 2 XCH-2015*	0	100%
23	3659	*Macaranga conifera*	*Macaranga sp. JH-2017*	0	100%
			Macaranga griffithiana	0	100%

5.4 Phylogenetic analysis

Six phylogenetic trees were constructed based on the multiple alignment of *rbcL, matK* and both *rbcL* and *matK namrkes.* The Neighbor joining (NJ) and Maximum likelihood (ML) methods were used for the construction of the phylogenetic trees. In all the phylogenetic trees, the laboratory sequences are named according to the morphologically identified name followed by sample ID and X in the parentheses while species name are given for the sequences downloaded from NCBI (see Appendix

5.4.1 *rbcL*

Neighbor Joining Method

Using Neighbor joining method and *Drypetes littoralis* as an outgroup, the *rbcL* sequences from collected samples correctly clustered together with the GenBank sequences representing the same genera. However, only a few of the sequences clustered together with the GenBank sequences representing the same species. The result showed various clusters belonging to different subfamilies.

The sequences from collected samples classified as *Macaranga* and *Mallotus* genera formed a paraphyletic- looking clade with the sequences of Acalyphoideae subfamily (clade I as shown in the figure 5.4.1) with bootstrap value of 65%. The *Balakata baccata* sequences (sample IDs 1653,2621, and 2622 as shown in the figure) formed clade with the species of Euphorbioideae subfamily (Clade II in the figure). Within that clade, two *Balakata baccata* sequences form a subclade with *Sapium baccatum* with a high bootstrap value of 82%. The five sequences of *Alchornea tiliifolia* (sample ID 1369, 1370, 1962, 1961, 1978) formed clade with Acalyphoideae subfamily with high bootstrap

value of 99% (represented by III in the figure 5.4.1). Only sequences of *Neoscortechinia kingii* (sample ID 4324) and *Pimelodendron griffithianum* (sample ID 4083) clustered together with the sequence of same species with bootstrap value 88% and 90% respectively. The morphologically unclassified samples (sample IDs 2628, 1728 and 1748) formed a paraphyletic- looking clade with a bootstrap value of 79% (clade IV in the figure). The sequences of *Melanolepis multiglandulosa* (sample ID 2705, 2720 and 2721) formed a paraphyletic clade with the species of Crotonoideae subfamily with very high bootstrap value of 99% (clade V in the figure 5.4.1).

Another interesting result in this phylogenetic tree is that sequences morphologically classified as *Endospermum cf diademum* (sample ID 3761) and *Macaranga conifera* (sample ID 3659) clustered together with species of the Moraceae family (but not with the sequences of Euphorbiaceae family) with strong 99% bootstrap value (clade VI in the figure). This shows that the samples have been could be morphologically misclassified as belonging to Euphorbiaceae.

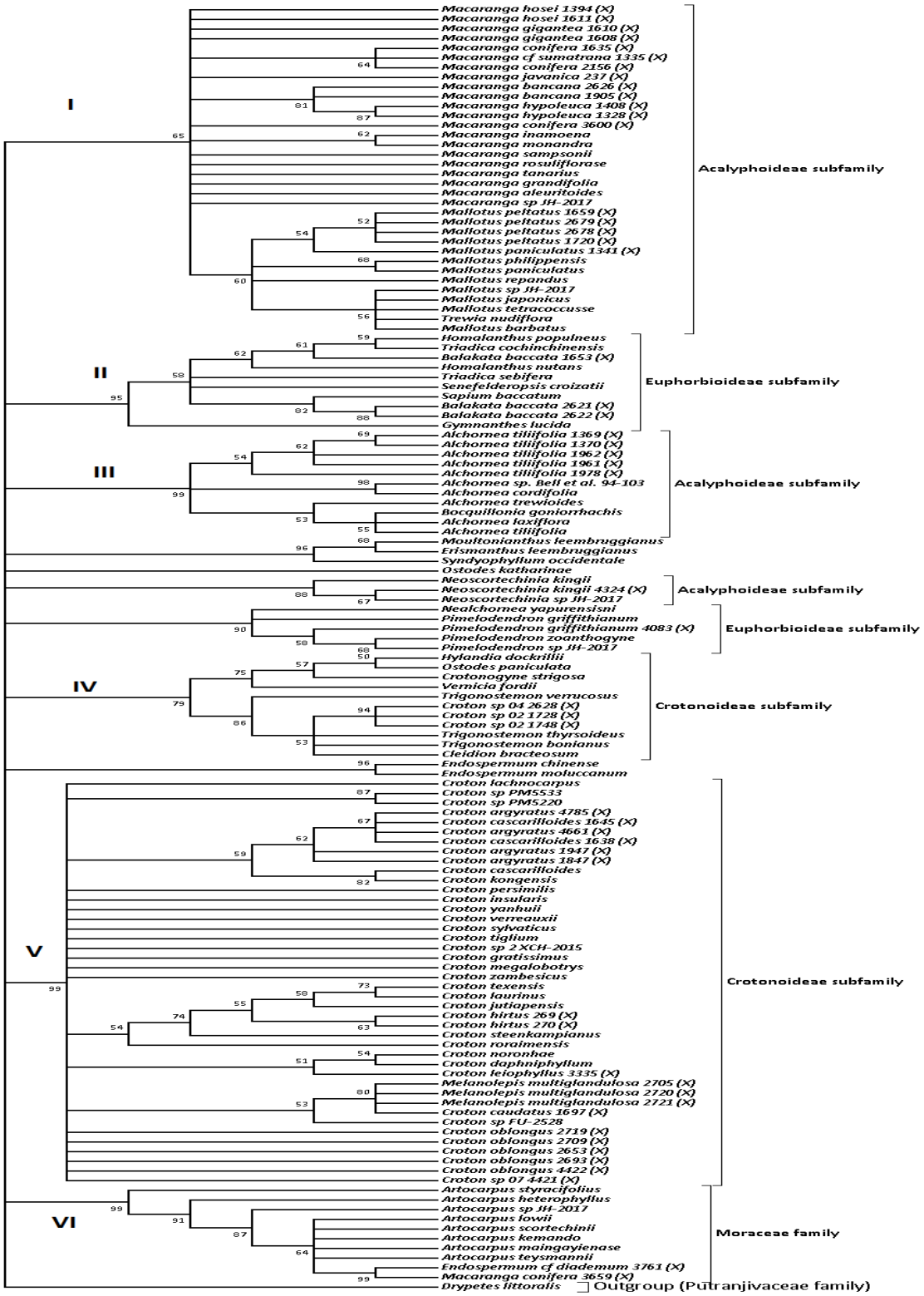


Figure 5.4.1: Neighbor joining phylogenetic tree of the samples representing the Euphorbiaceae plant family based on the *rbcL* gene sequences. The sequences representing collected samples are marked by X in the parentheses.

The phylogenetic relationships were inferred using the Neighbor-Joining tree(Nei, 1987).The optimal tree with the sum of branch length = 0.49604728 is presented in Fig 5.4.1. The bootstrap values (percentage of 1000 replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the clusters (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The genetic 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 133 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 510 positions in the final dataset. Evolutionary analyses were conducted in MEGA7(Kumar et al., 2016).

Maximum likelihood

The phylogenetic tree reconstructed by this method showed similar topology to the tree constructed by Neighbor joining method (Appendix 4). This method was also successful in correctly differentiating the sequences representing collected samples according to the species and genus level. This method also clustered morphologically misclassified samples i.e. *Melanolepis multiglandulosa* (sample ID 2705, 2720 and 2721) with supposedly correct genus *Crotonoideae* subfamily and with Endospermum cf diademum (sample ID 3761). Macaranga conifera (sample ID 3659) clustered with the sequences of Moraceae family.

The phylogenetic relationships were inferred by using the Maximum Likelihood method based on the Kimura 2-parameter model(Kimura, 1980). The tree with the highest log likelihood (-2421.42) is presented in Appendix 4. The bootstrap values (percentage of 1000 replicate trees in which the associated taxa clustered together in the bootstrap test) are shown next to the clusters. 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 133 nucleotide sequences. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

5.4.2 *matK*

The topology of the phylogenetic trees based on the *matK* marker and using neighbor joining and maximum likelihood was similar for both methods. Both methods distinguished different subfamily with high bootstrap support.

Neighbor joining

The phylogenetic tree reconstructed for *matK* sequences using Neighbor Joining method and *Drypetes littoralis* from Putranjivaceae family as an outgroup, showed a high bootstrap value of 96% (Figure 5.4.2). The sequences of *Macaranga and Mallotus* representing collected samples formed a paraphyletic clade with the sequences of Acalyphoideae subfamily available in NCBI with a high bootstrap value of 99% (denoted by I in figure). This clade included sequences representing collected samples belonging to M. *hosei, M. gigantea, M. conifer, M. javanica, M. tritrocarpa and Mallotos peltatus*. The morphologically misclassified sample of Macaranga sp (sample ID 377) formed a subclade (denoted by II in figure) with another morphologically misclassified sample of *Mallotus peltatus* (sample ID 2679) with bootstrap value of 68%. *Balakata baccata* sequences (sample ID 2621, 2622 and 1653) were placed in the clade with a very strong bootstrap of 100%(represented by III in figure 5.4.2. The *Croton* samples with IDs 2628, 1748 and 1726 formed a clade having a bootstrap value of 100% (represented by IV). Sequences from *Croton* genus and *Melanolepis multiglandulosa* were placed together in the subfamily Crotonoideae clade with bootstrap value of 99% (depicted as V in figure 5.4.2).

Species from Moraceae family were included to demonstrate the effectiveness of *matK* in differentiating the species among families (phylogenetic tree reconstructed by using *rbcL* successfully differentiated species of different families). The species of Moraceae family formed a clade with bootstrap 100%. Moraceae was not rooted (0% bootstrap value) with the Euphorbiaceae species. This shows *that matK* have very high discriminatory power at family level.

The phylogenetic relationships were inferred using the Neighbor-Joining method (Nei, 1987). The optimal tree with the sum of branch length = 0.86533472 is presented in figure 5.4.2. The bootstrap values (percentage of 1000 replicate trees in which the associated taxa clustered together in the bootstrap test) are shown next to the clusters (Felsenstein, 1985). The genetic distances were computed using the Maximum Composite Likelihood method (Kumar et al., 2016) and are in the units of the number of base substitutions per site. The analysis involved 94 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 613 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

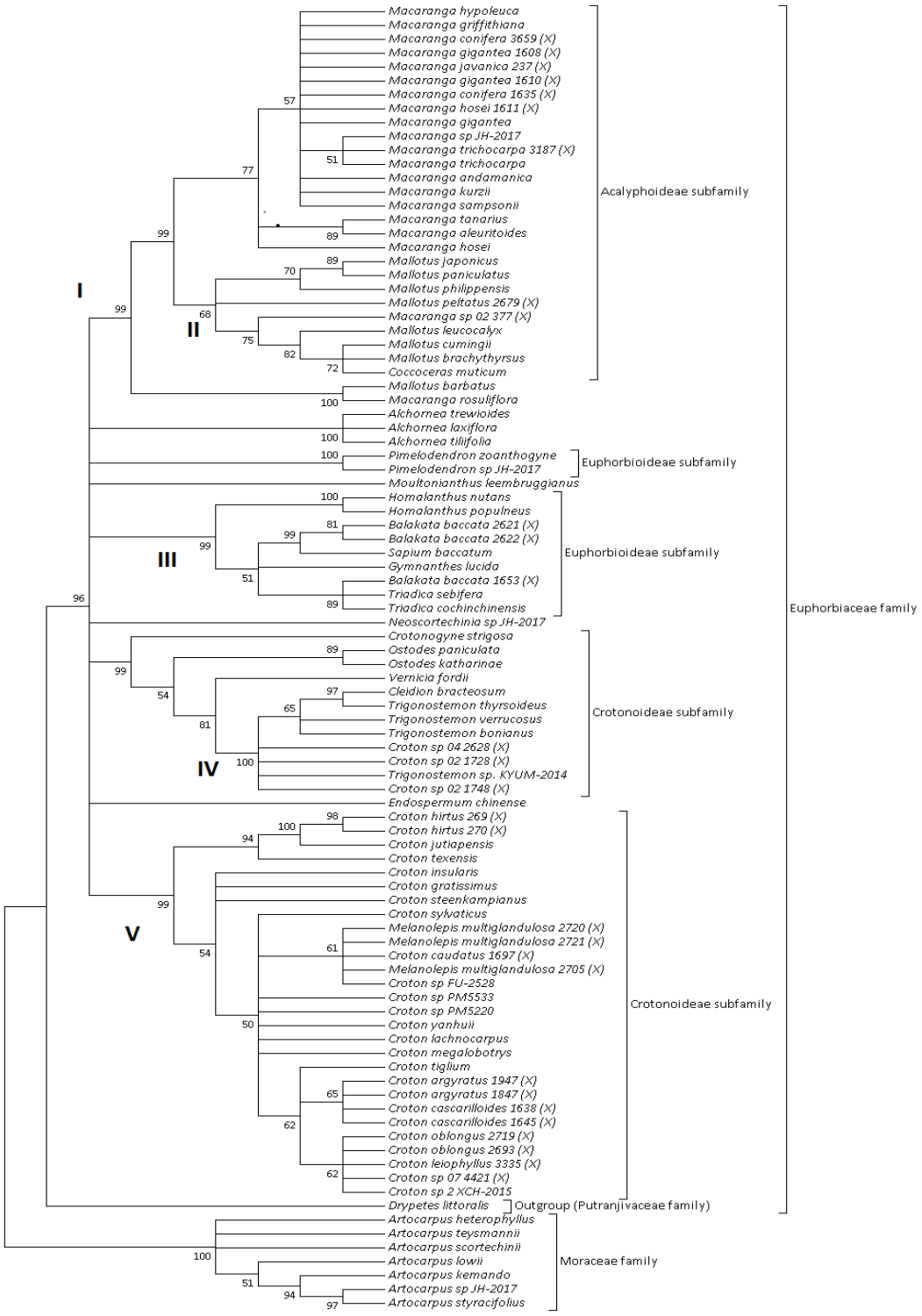


Figure 5.4.2: Neighbor joining phylogenetic tree of the samples representing the Euphorbiaceae plant family based on the *matK* gene sequences The laboratory sequences representing collected samples are marked with X in the parentheses.

Maximum likelihood

The positioning of the laboratory *matK* sequences were almost same like phylogenetic tree of neighbor joining tree. Outgroup was rooted with 91% bootstrap value. There were differences in the bootstrap value of the reconstructed clades and the sequences of *Mallotus peltatus (*sample ID 2679*)* and *Macaranga sp* (sample ID 377) got separated from each other and formed individual sub-clade. The phylogenetic tree constructed under this method is shown in Appendix 5

The phylogenetic relationships were inferred by using the Maximum Likelihood method based on the Kimura 2-parameter model(Kimura, 1980). The tree with the highest log likelihood (-3659.74) is shown in Appendix 5. The bootstrap values (percentage of 1000 replicate trees in which the associated taxa clustered together in the bootstrap test) are shown next to the clusters. 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 94 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 566 positions in the final dataset. Evolutionary analyses were conducted in MEGA7(Kumar et al., 2016).

5.4.3 Combination of *matK* and *rbcL*

Neighbor joining method

Drypetes littoralis as an outgroup and combination of both DNA barcodes i.e. *rbcL* and *matK* were used for construction of phylogenetic tree. Outgroup was rooted with bootstrap value of 64%. The combination of these gene regions resulted phylogenetic tree having major four paraphyletic clades (Shown in the figure 5.4.3)

Clade I contains the species belonging to the *Crotonoidea* subfamily with bootstrap value of 100%. Sequences from collected samples of *Croton* genus successfully clustered in this clade. The *Melanolepis multiglandulosa* sequences (sample ID 2705, 2720 and 2721) morphologically identified under *Acalyphoideae* subfamily also clustered in this clade. This shows that the samples could have been misidentified and placed into *Crotonoidea* subfamily.

The second clade (clade II) contained 3 misidentified samples that were clustered with genus *Trigonostemon* with high a bootstrap value of 100% while *Balakata baccata* samples successfully

clustered with species of Euphorbioideae subfamily with 100% bootstrap value. Clade IV contained sequences from collected samples of *Macaranga* and *Mallotus* clustered with other *Macaranga* and *Mallotus* sequences downloaded from NCBI.

Sequence from collected samples of *Macaranga conifera* (sample ID 3659) clustered with the species belonging to the Moraceae family. This shows that this samples have been misclassified.

The phylogenetic relationships were inferred using the Neighbor-Joining method(Nei, 1987). The optimal tree with the sum of branch length equal to 0.65147188 is shown in figure 5.4.3. The bootstrap values (percentage of 1000 replicate trees in which the associated taxa clustered together in the bootstrap test) are shown next to the clusters (Felsenstein, 1985). The genetic 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 79 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 1123 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

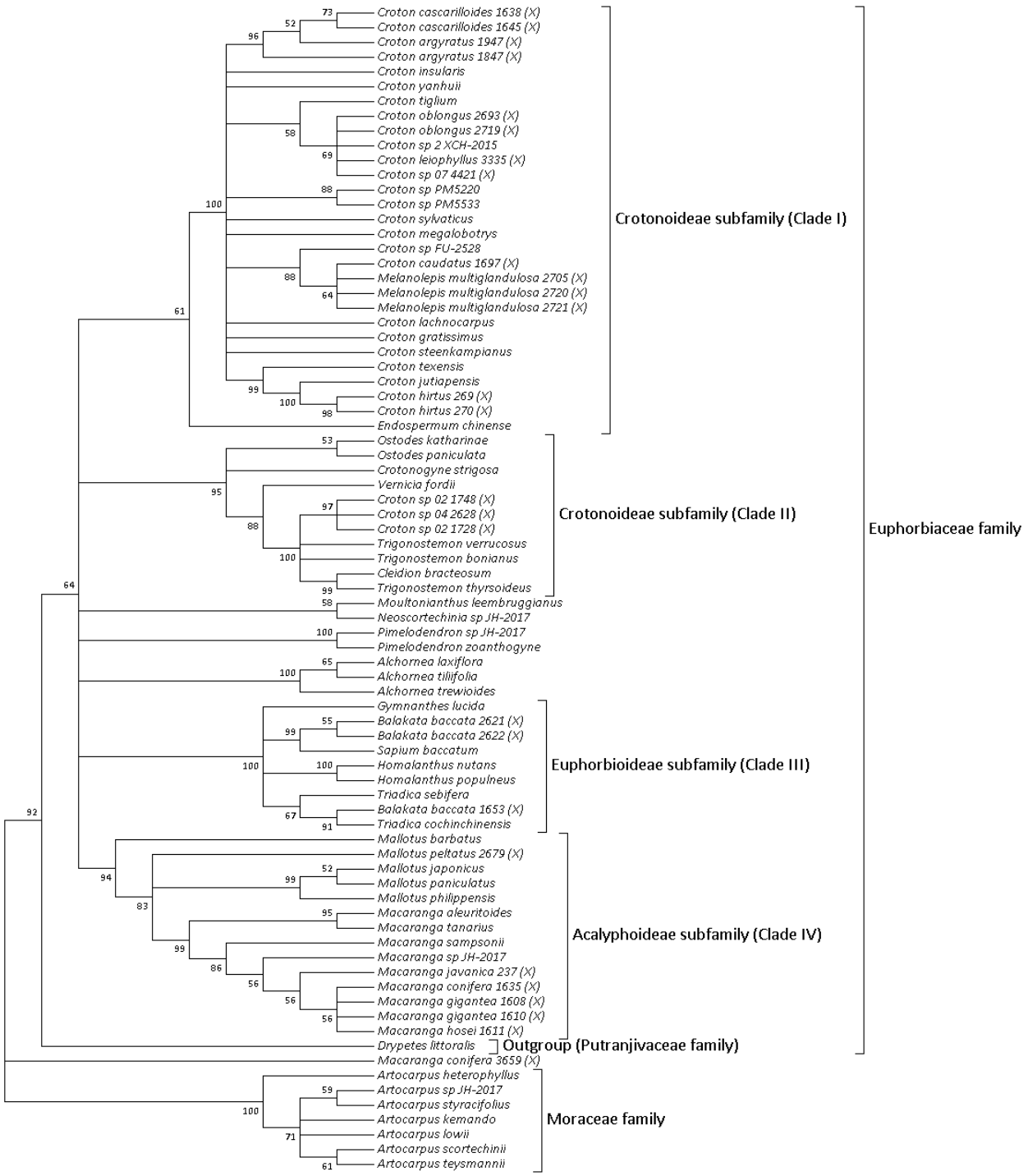


Figure 5.4.3: Neighbor joining phylogenetic tree of the samples representing the Euphorbiaceae plant family based on the *rbcL* and *matK* gene sequences

Maximum Likelihood

The positioning of the *matK* and *rbcL* sequences representing the collected samples were almost similar to phylogenetic tree constructed by neighbor joining tree method **(Appendix 6)**

The phylogenetic relationships were inferred by using the Maximum Likelihood method based on the Kimura 2-parameter model (Kimura, 1980). The tree with the highest log likelihood (-6032.74) is presented in Appendix 6. The bootstrap values (percentage of 1000 replicate trees in which the associated taxa clustered together in the bootstrap test) are shown next to the clusters. 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 79 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1084 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016)

6. Discussion

6.1 Barcode regions for Euphorbiaceae

It is well accepted that 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). Thus, the calculation of success rate of amplification and sequencing of different sequence can give a better evaluation of DNA barcoding in plant identification. This study focused on the calculation of amplification and sequencing success rate for two main primers pairs for *rbcL* and *matK* to fulfill its objectives.

The result of this study concluded that *rbcL* has relatively higher amplification (86%) and sequencing (78%) success rate compared to those for *matK* (60% and 49% 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*.

Many studies showed that it was almost impossible to calculate the amplification success rate of *matK* even by using the universal primers. Many other studies showed that *matK* was the most difficult sequence to amplify regardless the application of various conditions and dilutions. Again in case of tropical flora amplification seemed very difficult using *matK* as shown in study of(Gonzalez et al., 2009) and compared to temperate flora (de Vere et al., 2012), (Bruni et al., 2010). This is explained by (Gillman, Keeling, Gardner, & Wright, 2010) by the higher rate of evolution in the tropical flora compared to the temperate flora.

It is a great achievement that in our study could calculate the amplification success rate of the *matK* sequence as 60%.

It is also a notable result that while many studies concluded the aligned length of *matK* sequence appeared short, our study brought very positive result in terms of relatively long aligned length of *matK* equaled 613 bp. Furthermore, the results from this study showed that the problem in *matK* sequence is more with getting a higher number of sequences to be sequenced than to get a higher number of samples to be amplified.

Many other studies showed that *matK* marker can be useful as DNA barcode when they are used in certain taxa such as spices (De Mattia et al., 2011), tea plants ((Stoeckle et al., 2011)) and palm ((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.

For a more conclusive and strong results, higher number of representations is recommended for further studies. Again, with the tropical taxa and taxa with limited dispersal it is very challenging to use successfully use the Barcoding because of substantial phylogeographic structure. In this case use of taxonomically broad analysis may bring the better result. Additionally, analysis should extend beyond the focal geographic region so that evaluation and discrimination of potential sister taxa can be carried out. Examination of groups with frequent (possibly cryptic) hybridization, recent radiations, and high rates of gene transfer from mtDNA to the nucleus is also necessary(Moritz & Cicero, 2004)

Uncertainty is an inseparable part of any scientific research. Especially in biodiversity, accuracy is a relative term and never absolute because there are difficulties linked to plant species definition. Thus, there is a challenge for plant DNA barcoding to find the most suitable markers to tackle these problems. With this challenge lies an opportunity to advance the DNA sequencing technology and bioinformatics tools.

6.2 Identification using BLASTn

Our result comprises of many unknown sequences which needs to be compared with the reference sequence. 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).

Among the 67 samples, 59 samples were morphologically classified (Annex 1). Among 8 unclassified samples only 5 samples could be amplified and sequenced. Before moving to the discussions about those reasons let us take some theoretical assumptions in considerations.

Although there are no statistical methods that can measure the accuracy of identifications of 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. In simple way, 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.

Considering this and looking at our table 5.3. The sample ID 1728 showed that it matches different genera of Euphorbiaceae based on *rcbL* and matched two different genera based on *matK*. For both markers and all matching genera the E-value was 0.0 and identity % was very almost 100%. The sample ID 2628 more or less shows the same result. Other problems could be due to the complete sampling of all groups/genera in this study and incomplete representation of sampled taxa in the GenBank database. The sample ID 1748 matched sequences from the same genus based on *rbcL marker but* two genera based on *matK*.

A unique case is the sample ID 4421, where both markers showed that the sample was from same genus but from the different species. This clearly shows that the barcode database may lacks species level information. Whether, use of different markers would make us able to accomplish the species level identification for questionable samples in a question for further studies. The sample ID 377 was only analyzed using *rbcL* marker because it could not be sequenced successfully for the *matK* primer.

6.3 Identification success according to the best-close hit match analysis

Our results showed that *rbcL, matK* and combination of these barcode markers can lead to a correct identification of the species at the genus level. However, very few sequences found best match with the sequences of same species. This shows that BLASTn matches alone cannot be used for sample identification at species level. This can happen due to several reasons: i) not having enough sequences of our concerned species in NCBI GenBank database ii) not having enough sequences variations in our barcode regions. Thus increase of nucleotide databases, might be help us in the reliable identification. Also the use of other barcode regions for Euphorbiaceae may help to improve identification at the species level.

6.4 Phylogenetic analysis and comparison of molecular identification and morphological identification

The phylogenetic analysis conducted in this study to see if *matK* and *rbcL* barcodes resolve the investigated species into appropriate taxonomical grouping. The topologies of the nine phylogenetic trees (discussed above) reconstructed in this study based on two methods were more or less in consensus with each other. Only some differences in the clade position and bootstrap values was seen. The phylogenetic tree reconstructed helped us in the correct identification of the misclassified samples. Some of the misclassified samples are discussed below:

Family level

a. One of the *Macaranga* sample (field name Macaranga sp. 12, sample ID 3659) was morphologically identifed as *Macaranga conifera* (Appendix 1). But when seen in the phylogenetic trees, the sample placed itself in the clade belonging to the Moraceae family with a very high bootstrap values (Figure 5.4.1 and Appendix 4). This was also confirmed by the BLASTn result of the sequence of that species (Appendix 2).

b. Sample that was morphologically classified as *Endospermum cf diademum* also formed a clade with species of the Moraceae family (Figure 5.4.1 and Appendix 4) phylogenetic trees constructed by both Neighbor joining method and Maximum Likelihood method based on the *rbcL marker*. The BLASTn search also found high identical matches percentage of the sample with *Artocarpus* species of the Moraceae family.

Subfamily level

a. Three samples (sample IDs 2705, 2720 and 2721) were morphologically classified as *Melanolepis multiglandulosa* (that taxonomically belong to Acalyphoideae subfamily in Euphorbiaceae), but in the phylogenetic trees they clustered together in the clade with species from the Crotonoideae subfamily with 97% to 100% bootstrap values.

b.

Genus level

a. Three sample (Sample ID 1653, 2621 and 2622) were identified as Balakata baccata species. But when seen in the phylogenetic trees, it was seen together with species of Homalanthus genus.

7. Conclusion

Two DNA barcodes regions namely *rbcL* and *matK* were used to 1) verify the original morphological classification for samples from the tropical region as belonging to Euphorbiaceae family 2) understand the phylogenetic relationship of species in the Euphorbiaceae family and 3) compare the differences and similarities in morphological and molecular based identification of the species of Euphorbiaceae.

The *matK* and *rbcL* markers can be used as DNA barcodes of the sampled tropical plants. Although it is easier to amplify and sequence the *rbcL* marker, it did not provided enough information to discriminate the samples at the species level. The *matK* marker was more difficult to amplify and to sequence and it also did not perform well in discriminating the samples at the species level. However, it had a higher discrimination power than *rbcL* region. The combination of both the barcode markers showed better results in the species identification. For the better identification at species level, the use of other barcode regions like *cpDNA* markers and *psbA* should be explored.

The nucleotide BLAST in Genbank (NCBI), did not give clear results for the both species identification and barcode analysis of two regions. Many ambiguous results made difficult in proper species identification and barcode analysis. The expansion of the nucleotide database might help with better species identification and barcode analysis.

Finally, comparison between molecular and morphological identification was done using six phylogenetic trees. Trees were reconstructed based on *matK, rbcL,* and both markers using two different phylogenetic tree construction methods (Neighbor Joining and Maximum Likelihood). This analysis showed that neither *matK and rbcL* alone nor the combined marker were able to give a better identification at species level. However, the phylogenetic trees were successful in discriminating samples at the genus and other higher taxonomic levels.

8. 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.

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 Selection of candidate coding DNA barcoding regions for use on land plants, (March 2016), 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*,

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 plant systematics. *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).

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 subtribeSonchinae (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*,

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.

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.).

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, 1–19.

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.

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.

Tokuoka, T., & Tobe, H. (2006). Phylogenetic analyses of Malpighiales using plastid and nuclear DNA sequences, with particular reference to the embryology of Euphorbiaceae sens. str. *Journal of Plant Research*, *119*(6), 599–616.

UN FAO. (2015). *Global Forest Resources Assessment 2015 - Desk reference*. Retrieved from

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 ™ Series Editor*. *Life Sciences* (Vol. 531).

Webster, G. L. (1994). Classification of the, (December 1973), 3–32.

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.

9. Appendixes

Appendix 1: Details of the morphological classification of collected samples

S.N	Sample ID	Field name	Morphological name Species
1	214	Macaranga sp. 02	*Mallotus peltatus*
2	237	Macaranga sp. 03	*Macaranga javanica*
3	269	Euphorbiaceae sp. 03	*Croton hirtus*
4	270	Euphorbiaceae sp. 03	*Croton hirtus*
5	377	Macaranga sp. 02	*Not identified*
6	1265	cf. Mallotus sp. 02	*Alchornea tiliifolia*
7	1290	Euphorbiaceae sp. 12	*Not identified*
8	1303	Euphorbiaceae sp. 13_1	*Macaranga cf. trichocarpa*
9	1304	Euphorbiaceae sp. 13_1	*Macaranga trichocarpa*
10	1328	Macaranga "alba"	*Macaranga hypoleuca*
11	1335	Macaranga cf. lowii	*Macaranga cf. sumatrana*
12	1336	Macaranga cf. lowii	*Macaranga cf. sumatrana*
13	1341	Mallotus paniculatus	*Mallotus paniculatus*
14	1369	Mallotus sp. 03	*Alchornea tiliifolia*
15	1370	Mallotus sp. 03	*Alchornea tiliifolia*
16	1394	Macaranga "alba" 02	*Macaranga hosei*
17	1408	Macaranga "alba" 01	*Macaranga hypoleuca*
18	1608	Macaranga sp. 07	*Macaranga gigantea*
19	1610	Macaranga sp. 07	*Macaranga gigantea*
20	1611	Macaranga "alba" 02	*Macaranga hosei*
21	1635	Macaranga sp. 08	*Macaranga conifera*
22	1638	Croton cf argyratus	*Croton cascarilloides*
23	1645	Croton cf argyratus	*Croton cascarilloides*
24	1653	Homalanthus sp. 01	*Balakata baccata*
25	1659	Mallotus sp. 04	*Mallotus peltatus*
26	1697	Euphorbiaceae sp. 17	*Croton caudatus*
27	1720	Mallotus sp. 04	*Mallotus peltatus*
28	1728	Croton sp. 2	*Not identified*
29	1748	Croton sp.2	*Not identified*
30	1847	Croton sp. 03	*Croton argyratus*
31	1905	Macaranga triloba	*Macaranga bancana*
32	1947	Croton sp. 03	*Croton argyratus*
33	1961	Mallotus sp. 05	*Alchornea tiliifolia*
34	1962	Mallotus sp. 05	*Alchornea tiliifolia*

35	1978	Euphorbiaceae sp. 20	*Alchornea tiliifolia*
36	2112	Antidesma sp. 11	*Cephalomappa malloticarpa*
37	2113	Antidesma sp. 11	*Cephalomappa malloticarpa*
38	2156	Macaranga sp. 09	*Macaranga conifera*
39	2621	Homolanthus sp. 01	*Balakata baccata*
40	2622	Homolanthus sp. 01	*Balakata baccata*
41	2626	Macaranga sp. 10	*Macaranga bancana*
42	2628	Croton sp. 04	*Not identified*
43	2653	Croton sp. 05	*Croton oblongus*
44	2678	Macaranga sp. 11	*Mallotus peltatus*
45	2679	Macaranga sp. 11	*Mallotus peltatus*
46	2693	Croton sp. 05	*Croton oblongus*
47	2705	Mallotus sp. 06	*Melanolepis multiglandulosa*
48	2709	Croton sp. 05	*Croton oblongus*
49	2719	Croton sp. 05	*Croton oblongus*
50	2720	Mallotus sp. 06	*Melanolepis multiglandulosa*
51	2721	Mallotus sp. 06	*Melanolepis multiglandulosa*
52	3187	Euphorbiaceae sp. 25	*Macaranga trichocarpa*
53	3335	Croton sp. 06	*Croton leiophyllus*
54	3600	Macaranga sp. 12	*Macaranga conifera*
55	3659	Macaranga sp. 12	*Macaranga conifera*
56	3761	Endospermum cf. diademum	*Endospermum diadenum*
57	3873	Euphorbiaceae sp. 26	*Not identified*
58	4063	Euphorbiaceae sp. 28	*Neoscortechinia kingii*
59	4083	Pimelodendron zoanthogyne	*Pimelodendron griffithianum*
60	4128	Tree 90	*Neoscortechinia kingii*
61	4324	Tree 90	*Neoscortechinia kingii*
62	4421	Croton sp. 07	*Not identified*
63	4422	Croton sp. 07	*Croton oblongus*
64	4661	Croton argyratus	*Croton argyratus*
65	4731	Macaranga trichocarpa	*Macaranga trichocarpa*
66	4785	Croton cascarilloides	*Croton argyratus*
67	5185	Euphorbiaceae sp. 29	*Not identified*

Appendix 2: The best match hits of *rbcL* sequences using BLASTn

S.N	Sample ID	Herbarium name	NCBI best match result	E-value	identity
1	237	*Macaranga javanica*	*Macaranga tanarius*	0	99%
			Macaranga aleuritoides	0	99%
2	269	*Croton hirtus*	*Croton texensis*	0	99%
			Croton persimilis	0	99%
3	270	*Croton hirtus*	*Croton persimilis*	0	99%
			Croton texensis	0.0	99%
			Croton gratissimus	0.0	99%
4	1328	*Macaranga hypoleuca*	*Macaranga sp. JH-2017*	0	99%
			Macaranga tanarius	0	99%
			Macaranga aleuritoides	0	99%
			Macaranga sampsonii	0	99%
5	1335	*Macaranga sumatrana*	*Macaranga tanarius*	0	98%
			Macaranga aleuritoides	0	98%
			Macaranga monandra	0	98%
6	1341	*Mallotus paniculatus*	*Mallotus repandus*	0	99%
			Mallotus japonicus	0	99%
			Mallotus philippensis	0	99%
7	1369	*Alchornea tiliifolia*	*Alchornea tiliifolia*	0	99%
			Alchornea trewioides	0	99%
8	1370	*Alchornea tiliifolia*	*Alchornea tiliifolia*	0	99%
			Alchornea trewioides	0.0	99%
			Alchornea laxiflora	0.0	99%
9	1394	*Macaranga hosei*	*Macaranga*	0	99%

			tanarius		
			Macaranga aleuritoides	0	99%
10	1408	*Macaranga hypoleuca*	*Macaranga sp. JH-2017*	0	99%
			Macaranga tanarius	0	99%
			Macaranga grandifolia	0	99%
			Macaranga sampsonii	0	99%
11	1608	*Macaranga gigantea*	*Macaranga tanarius*	0	99%
			Macaranga grandifolia	0	99%
			Macaranga sp. JH-2017	0	99%
			Macaranga sampsonii	0	99%
			Macaranga aleuritoides	0	99%
12	1610	*Macaranga gigantea*	*Macaranga tanarius*	0	99%
			Macaranga grandifolia	0	99%
			Macaranga sp. JH-2017	0	99%
			Macaranga sampsonii	0	99%
13	1611	*Macaranga hosei*	*Macaranga tanarius*	0	99%
			Macaranga grandifolia	0	99%
			Macaranga sp. JH-2017	0	99%
			Macaranga sampsonii	0	99%
			Macaranga aleuritoides	0	99%
14	1635	*Macaranga conifera*	*Macaranga sampsonii*	0	99%
			Macaranga tanarius	0	99%
			Macaranga grandifolia	0	99%
			Macaranga sp. JH-2017	0	99%

			Macaranga rosuliflora	0	99%
15	1638	*Croton cascarilloides*	*Croton cascarilloides*	0	99%
			Croton kongensis	0.0	99%
			Croton persimilis	0.0	99%
			Croton persimilis	0.0	99%
16	1645	*Croton cascarilloides*	*Croton cascarilloides*	0	99%
			Croton persimilis	0.0	99%
			Croton insularis	0.0	99%
17	1653	*Balakata baccata*	*Triadica cochinchinensis*	0	99%
			Triadica sebifera	0	99%
18	1659	*Mallotus peltatus*	*Mallotus sp. JH-2017*	0	99%
			Mallotus barbatus	0	99%
			Mallotus sp. SH-2010	0	99%
			Mallotus japonicus	0	99%
			Mallotus repandus	0	99%
			Trewia nudiflora	0	99%
19	1697	*Croton caudatus*	*Croton tiglium*	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton roraimensis	0	99%
			Croton daphniphyllum	0	99%
20	1720	*Mallotus peltatus*	*Mallotus sp. JH-2017*	0	99%

			Mallotus barbatus	0	99%
			Mallotus sp. SH-2010	0	99%
			Mallotus japonicus	0	99%
			Mallotus tetracoccus	0	99%
			Mallotus paniculatus	0	99%
21	1847	*Croton argyratus*	*Croton persimilis*	0	99%
			Croton insularis	0	99%
			Croton tiglium	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%
22	1905	*Macaranga bancana*	*Macaranga tanarius*	0	98%
			Macaranga sampsonii	0	98%
23	1947	*Croton argyratus*	*Croton kongensis*	0	99%
			Croton cascarilloides	0	99%
			Croton persimilis	0	99%
			Croton insularis	0	99%
			Croton tiglium	0	99%
24	1961	*Alchornea tiliifolia*	*Alchornea tiliifolia*	0	99%
25	1962	*Alchornea tiliifolia*	*Alchornea tiliifolia*	0	99%
26	1978	*Alchornea tiliifolia*	*Alchornea tiliifolia*	0	99%
27	2156	*Macaranga conifera*	*Macaranga tanarius*	0	99%
			Macaranga sampsonii	0	99%
			Macaranga aleuritoides	0	99%
			Macaranga monandra	0	99%

28	2621	*Balakata baccata*	*Triadica sebifera*	0	99%
29	2622	*Balakata baccata*	*Triadica sebifera*	0	99%
			Triadica cochinchinensis	0	99%
30	2626	*Macaranga bancana*	*Macaranga sp. JH-2017*	0	99%
			Macaranga tanarius	0	99%
			Macaranga grandifolia	0	99%
			Macaranga sampsonii	0	99%
31	2653	*Croton oblongus*	*Croton tiglium*	0	100%
			Croton megalobotrys	0	100%
			Croton gratissimus	0	100%
			Croton zambesicus	0	100%
32	2678	*Mallotus peltatus*	*Mallotus sp. JH-2017*	0	99%
			Mallotus japonicus	0	99%
			Mallotus repandus	0	99%
			Mallotus philippensis	0	99%
33	2679	*Mallotus peltatus*	*Mallotus sp. JH-2017*	0	99%
			Mallotus japonicus	0	99%
			Mallotus repandus	0	99%
			Mallotus philippensis	0	99%
34	2693	*Croton oblongus*	*Croton tiglium*	0	100%
			Croton megalobotrys	0	100%
			Croton gratissimus	0	100%
			Croton zambesicus	0	100%
			Croton noronhae	0	99%

35	2705	*Melanolepis multiglandulosa*	*Croton tiglium*	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%
			Croton roraimensis	0	99%
36	2709	*Croton oblongus*	*Croton tiglium*	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%
			Croton sp. 2 XCH-2015	0	99%
			Croton roraimensis	0	99%
37	2719	*Croton oblongus*	*Croton tiglium*	0	100%
			Croton megalobotrys	0	100%
			Croton gratissimus	0	100%
			Croton zambesicus	0	100%
38	2720	*Melanolepis multiglandulosa*	*Croton tiglium*	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%
			Croton roraimensis	0	99%
39	2721	*Melanolepis multiglandulosa*	*Croton tiglium*	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%

			Croton roraimensis	0	99%
40	3335	*Croton leiophyllus*	*Croton noronhae*	0	99%
			Croton daphniphyllum	0	99%
			Croton tiglium	0	99%
			Croton megalobotrys	0	99%
			Croton gratissimus	0	99%
			Croton zambesicus	0	99%
			Croton roraimensis	0	99%
41	3600	*Macaranga conifera*	*Macaranga tanarius*	0	96%
			Macaranga aleuritoides	0	96%
			Macaranga sampsonii	0	96%
			Macaranga sp. JH-2017	0	96%
			Macaranga monandra	0	96%
			Macaranga grandifolia	0	96%
42	3659	*Macaranga conifera*	*Artocarpus teysmannii*	0	99%
			Artocarpus scortechinii	0	99%
			Artocarpus maingayi	0	99%
			Artocarpus lowii	0	99%
			Artocarpus kemando	0	99%
43	3761	*Endospermum diadenum*	*Artocarpus sp. JH-2017*	0	99%

Appendix 3: List of plant species and corresponding GenBank accession numbers retrieved from the database for *rbcL* and *matK*

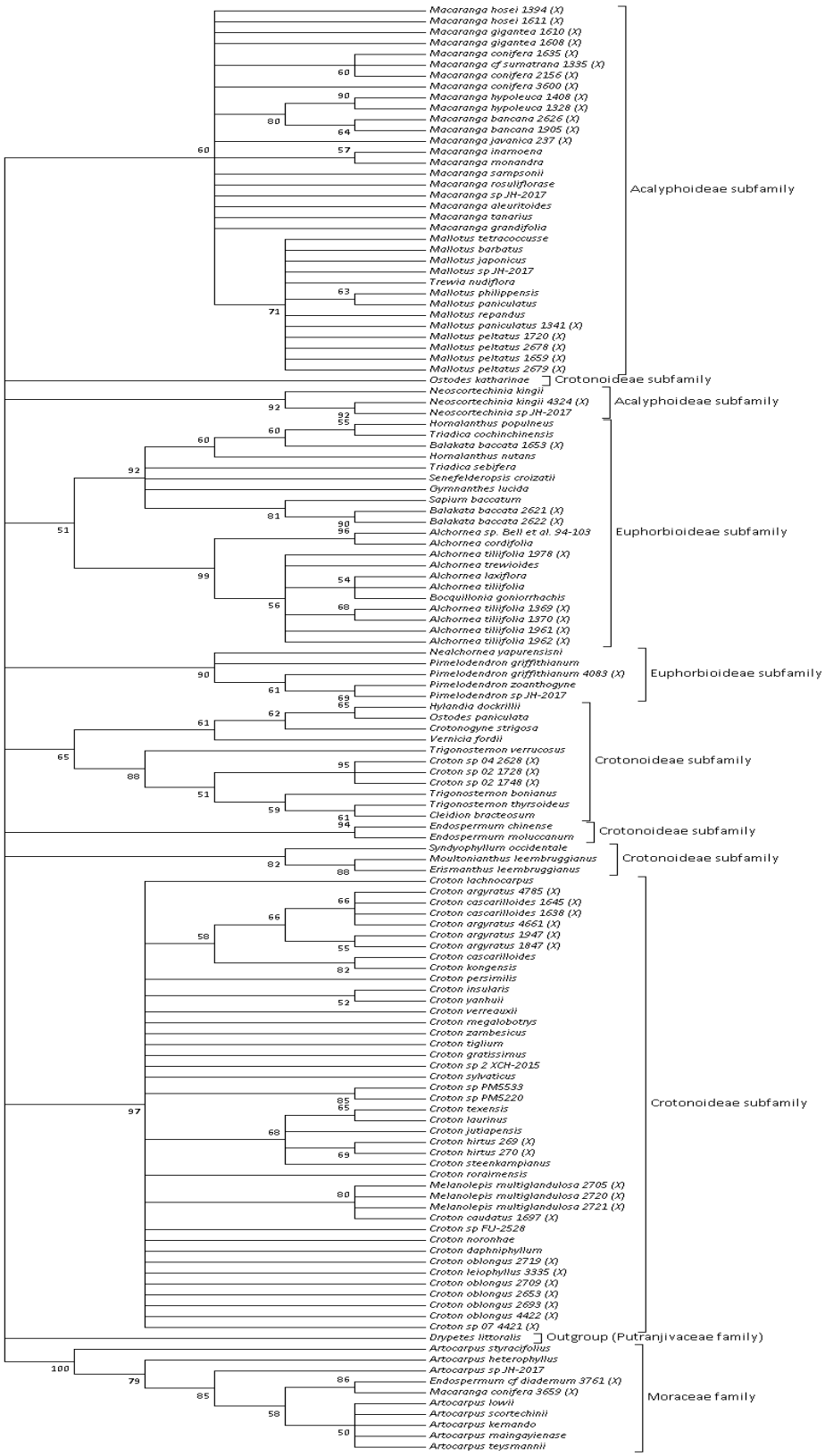
S.N	Species Name	*rbcL*		*matK*	
		Accession number	GI	Accession number	GI
1	*Alchornea cordifolia*	AY794959.1	62861330		
2	*Alchornea laxiflora*	AY794957.1	62861326	JX517659.1	407910101
3	*Alchornea sp. Bell et al. 94-103*	AY794956.1	62861324		
4	*Alchornea tiliifolia*	KR528671.1	874511900	KR530319.1	874515171
5	*Alchornea trewioides*	GU441783.1	289190238	GU441801.1	289190274
6	*Artocarpus heterophyllus*	KF724291.1	597518532	KU856361.1	1151021571
7	*Artocarpus kemando*	KU856248.1	1151020659	KU856368.1	1151021585
8	*Artocarpus lowii*	KU856259.1	1151020681	KU856380.1	1151021609
9	*Artocarpus maingayi*	KU856261.1	1151020685		
10	*Artocarpus scortechinii*	KU856293.1	1151020749	KU856414.1	1151021677
11	*Artocarpus sp. JH-2017*	MF435679.1	1269807013	MF419031.1	1270118720
12	*Artocarpus styracifolius*	KJ440018.1	657171891	HQ415243.1	331704491
13	*Artocarpus teysmannii*	KU856300.1	1151020763	KU856421.1	1151021691
14	*Bocquillonia goniorrhachis*	AY794958.1	62861328		
15	*Cleidion bracteosum*	KR529013.1	874512568	KR530607.1	874515744
16	*Coccoceras muticum*			EF582665.1	157613783
17	*Croton cascarilloides*	KR529034.1	874512610		
18	*Croton daphniphyllum*	EF405836.1	126166013		
19	*Croton gratissimus*	EU213460.1	167891329	JX517905.1	407910593
20	*Croton insularis*	AB233877.1	119368069	AB233773.1	118917503
21	*Croton jutiapensis*	JQ591437.1	384590418	JQ587444.1	384582432
22	*Croton kongensis*	KR529037.1	874512616		
23	*Croton lachnocarpus*	KP094558.1	756776227	HQ415239.1	331704483
24	*Croton laurinus*	EF405842.1	126166025		
25	*Croton megalobotrys*	EU213463.1	167891335	EU214234.1	167890104
26	*Croton noronhae*	EF405848.1	126166037		

27	*Croton persimilis*	KF523366.1	545290159		
28	*Croton roraimensis*	EF405852.1	126166045		
29	*Croton sp. 2 XCH-2015*	KR529049.1	874512640	KR530635.1	874515800
30	*Croton sp. FU-2528*	AB936042.1	641309156	AB936043.1	641309159
31	*Croton sp. PM5220*	KC628320.1	480308882	KC627680.1	480307602
32	*Croton sp. PM5533*	KC628460.1	480309162	KC627789.1	480307820
33	*Croton steenkampianus*	JF265379.1	326394123	JX517563.1	407909909
34	*Croton sylvaticus*	JX572488.1	409976019	JX517596.1	407909975
35	*Croton texensis*	KT458041.1	984913133	KT456905.1	984910904
36	*Croton tiglium*	GQ436320.1	290585649	KP093548.1	756774207
37	*Croton verreauxii*	KM895498.1	770589811		
38	*Croton yanhuii*	KR529052.1	874512646	KR530639.1	874515808
39	*Croton zambesicus*	EF405856.1	126166053		
40	*Crotonogyne strigosa*	KC628194.1	480308630	KC627586.1	480307414
41	*Drypetes littoralis*	AB233926.1	118917733	AB233822.1	118917601
42	*Endospermum chinense*	KJ440013.1	657171881	KJ510913.1	657172069
43	*Endospermum moluccanum*	AJ402950.1	9909639		
44	*Erismanthus leembruggianus*	MF435457.1	1269806569		
45	*Gymnanthes lucida*	AY794858.1	62861136	KJ012630.1	588283225
46	*Homalanthus nutans*	AB267957.1	155029278	AB268061.1	155029470
47	*Homalanthus populneus*	AY380350.1	39653953	EF135548.1	149213086
48	*Hylandia dockrillii*	AY794882.1	62861179		
49	*Macaranga aleuritoides*	AB267922.1	155573926	AB268026.1	155029400
50	*Macaranga andamanica*			HQ415380.1	331704759
51	*Macaranga gigantea*			EF582626.1	157613698
52	*Macaranga griffithiana*			AB925046.1	619326262
53	*Macaranga grandifolia*	AY794935.1	62861283		
54	*Macaranga hosei*			KU519674.1	1016572596
55	*Macaranga hypoleuca*			EF582627.1	157613700
56	*Macaranga kurzii*			EF582629.1	157613704
57	*Macaranga inamoena*	KF496314.1	530444015		
58	*Macaranga monandra*	KC628123.1	480308488		
59	*Macaranga rosuliflora*	KR529617.1	874513769	KR531108.1	874516742
60	*Macaranga sampsonii*	KJ440011.1	657171877	HQ415381.1	331704761
61	*Macaranga sp. JH-2017*	MF435467.1	1269806589	MF418964.1	1270118586
62	*Macaranga tanarius*	AB233866.1	118917665	AB233762.1	118917481
63	*Macaranga trichocarpa*			EF582631.1	157613708
64	*Mallotus barbatus*	KR529656.1	874513847	KR531145.1	874516816

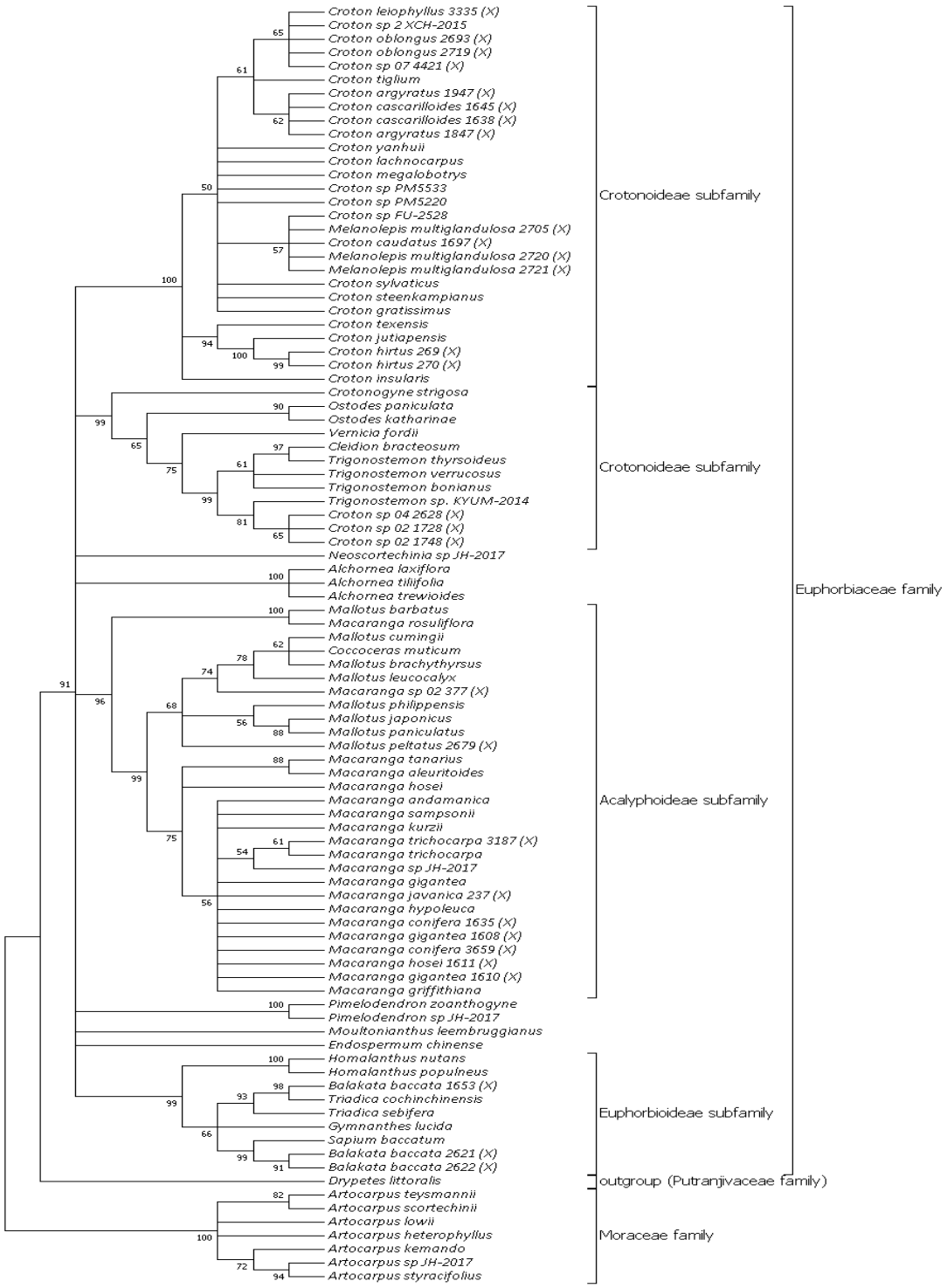
65	*Mallotus brachythyrsus*			EF582634.1	157613714
66	*Mallotus cumingii*			EF582642.1	157613730
67	*Mallotus japonicus*	AY794934.1	62861282	AB268027.1	155029402
68	*Mallotus leucocalyx*			EF582654.1	157613757
69	*Mallotus paniculatus*	KP094350.1	756775811	AB924949.1	619326072
70	*Mallotus philippensis*	GU441775.1	289190222	HQ415385.1	331704769
71	*Mallotus repandus*	GU441787.1	289190246		
72	*Mallotus sp. JH-2017*	MF435469.1	1269806593		
73	*Mallotus tetracoccus*	KR529669.1	874513873		
74	*Moultonianthus leembruggianus*	AY794982.1	62861373	FJ670015.1	261873585
75	*Nealchornea yapurensis*	AY794865.1	62861149		
76	*Neoscortechinia kingii*	AJ402977.1	9909842		
77	*Neoscortechinia sp. JH-2017*	MF435474.1	1269806603	MF418957.1	1270118572
78	*Ostodes katharinae*	KR529861.1	874514256	KR531324.1	874517173
79	*Ostodes paniculata*	AB267948.	155029262	AB268052.1	155029452
80	*Pimelodendron griffithianum*	AB233887.1	119368081		
81	*Pimelodendron sp. JH-2017*	MF435478.1	1269806611	MF418981.1	1270118620
82	*Pimelodendron zoanthogyne*	AJ418812.1	17066144	EF135582.1	149213147
83	*Sapium baccatum*	KR529984.1	874514502	KR531442.1	874517399
84	*Senefelderopsis croizatii*	AY794860.1	62861140		
85	*Syndyophyllum occidentale*	AY794967.1	62861344		
86	*Trewia nudiflora*	AY663648.1	55792804		
87	*Triadica cochinchinensis*	KY501149.1	1324898049	AB925031.1	619326232
88	*Triadica sebifera*	AY794859.1	62861138	AB268065.1	155029478
89	*Trigonostemon bonianus*	KR530166.1	874514866	KR531598.1	874517710
90	*Trigonostemon thyrsoideus*	KR530171.1	874514876	KR531604.1	874517722
91	*Trigonostemon verrucosus*	AY788192.1	62003646	FJ670020.1	261873595
92	*Vernicia fordii*	KF022509.1	573017057	KF022442.1	573016923

Note: The blank spaces represent there were no sequences available for that particular marker. The highlighted row represents species used for outgroup.

Appendix 4: Phylogenetic relationships *rbcL* sequences based on Maximum Likelihood



Appendix 5: Phylogenetic relationships of *matK* sequences based on Maximum Likelihood



Appendix 6: Phylogenetic relationships *of rbcL and matK* sequences based on Maximum Likelihood