PHYLOGENETIC STUDIES OF NEW WORLD SPECIES IN THE PLANT GENUS *PSYCHOTRIA* (RUBIACEAE)

by

LING DONG

(Under the Direction of JAMES H. LEEBENS-MACK)

ABSTRACT

This dissertation aimed to investigate genetic variation and modes of speciation in the tropical plant genus *Psychotria* (Rubiaceae), an extremely diverse group comprising an important component of the neo-tropical understory vegetation. Sequence comparison of *P. marginata* and *Coffea arabica* plastomes and available transcriptome data were used to develop non-coding sequence markers and perform multi-locus phylogenetic analyses across 19 species of *Psychotria* and related taxa (*Rudgea*, *Palicourea* and *Coffea)*. In the first study, the complete sequence of the *P. marginata* chloroplast genome was characterized, and non-coding plastid markers were characterized and assessed for level of polymorphism. The second study investigated the species-level relationships of the 19 species using 5 nuclear (ITS, NL-57800, NL-217, NL-103, NL-A04) and 4 plastid (*psbE-petL*, *trnT-psbD*, and *trnK-rps16* and *trnL-trnF*) non-coding sequence markers. Maximum-likelihood gene trees for each locus and coalescent species-tree analyses showed significant improvement of bootstrap support for relationships within *Psychotria*, although relationships towards the tips of the tree were not well supported, presumably due to rapid radiation of neo-tropical *Psychotria* species. Subgenus-level relationships, as well as the paraphyly of subgenus *Heteropsychotria,* were well supported in

both the gene trees and the species tree, and were also consistent with previous studies. In addition, polyphyletic patterns of *P. pilosa* and *P. allenii* haplotypes on the gene trees indicated incomplete lineage sorting and the possibility of a cryptic species complex as a result of rapid evolution. Thus coalescent simulations were performed in the third study to further assess whether the relationships inferred in the gene tree and species tree analyses could be caused solely through random sorting processes under a regionally sympatric speciation model with gene flow between populations. As expected, random sorting after the speciation event could cause polyphyletic haplotypes on gene trees, which made species delimitation difficult, particularly when effective population sizes are large, gene flow is frequent, and divergence time is short. The simulated data in turn supported the hypothesis that *P. pilosa* is at a stage in the speciation process where sorting of ancestral lineages is not complete, causing haplotypes from different populations to form paraphyletic groups with its sister species, *P. cyanococca,* in some of the gene trees.

INDEX WORDS: chloroplast genome, phylogeny, speciation, gene tree, species tree, coalescent, *Psychotria*, Rubiaceae

# PHYLOGENETIC STUDIES OF NEW WORLD SPECIES IN THE PLANT GENUS *PSYCHOTRIA* (RUBIACEAE)

by

LING DONG

BS, Sichuan University, China, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

PHYLOGENETIC STUDIES OF NEW WORLD SPECIES IN THE PLANT GENUS

*PSYCHOTRIA* (RUBIACEAE)

by

LING DONG

| | |
|---|---|
| Major Professor: | James H. Leebens-Mack |
| Committee: | James L. Hamrick |
| | Wendy B. Zomlefer |
| | Kelly A. Dyer |
| | John P. Wares |

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2013

# DEDICATION

To my parents, Xianghong Dong and Xiaowei Ma; and my husband, Zhao Li.

# ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor, Jim Leebens-Mack, for offering me the opportunity to start this journey in the United States when I was first considering applying to the Ph.D. program in the Plant Biology at UGA, and for the mentorship and guidance he has given throughout my graduate studies. I would also like to thank the rest of my committee, Jim Hamrick, Wendy Zomlefer, Kelly Dyer and John Wares, for their insightful comments, thought-provoking questions as well as encouragement to help me get through this journey. I also wish to give special thanks to Dr. Liang Liu in the Department of Statistics for many helpful discussions and suggestions for the last chapter of my dissertation. It has been a nice experience to share a lot of good times with the Leebens-Mack lab group past and present. I enjoyed working with our post-docs, Joe McNeal, Hongyan Shan, and Alexa Telgmann; my fellow graduate students, Michael McKain, Jeremy Rentsch, Alex Harkess, Jason Comer, Karolina Heyduk, and Lauren Eserman; and our lab techs, Chang Liu, Charlotte Carrigan Quigley and Dat Trong Hoang, all of whom have been great friends and colleagues.

My dissertation work would not have been completed without assistance from staff members of La Selva and Las Cruces Biological Stations in Costa Rica. I thank Orlando Vargas, José González and Federico Oviedo. Their amazing knowledge on the diverse tropical plant species as well as their generous help made my sampling process in the field much easier and smoother.

My sincere thanks also go to the great people in the Plant Biology Department and UGA. Dr. Brigitte Bruns, the lab coordinator in our department, and Dr. Kristen Miller, the lab

coordinator in Biological Sciences, have been very helpful and supportive for my work as a graduate lab assistant. Particularly, Brigitte has always been caring and offered lots of suggestions on my teaching all the way from when I was a teaching apprentice for PBIO1210 when I just started graduate school in 2007. I would also like to thank Susan Watkins, our Graduate Admissions Counselor, who has always been very friendly, efficient and helpful on various issues that students would go through in graduate school.

Finally, I appreciate the emotional support from my family – my parents Xianghong Dong and Xiaowei Ma; as well as my husband Zhao Li – throughout my life.

TABLE OF CONTENTS

Page

# LIST OF TABLES

Page

# LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION AND LITERATURE REVIEW

## 1.1 Gene Trees vs. Species Trees - Assessing Species Diversification Using Phylogenetic Tools

The evolutionary history of species diversification includes a dramatic increase of complexity within and among organisms over millions of years. Understanding current biodiversity and the processes that promote speciation is a major goal for ecologists and evolutionary biologists. The process of speciation includes restriction of gene flow between diverging populations and genetic drift and/or selection acting on ancestral variation within the diverging populations. Understanding the phylogenetic relationships between different groups of taxa is crucial for in inferring patterns of species diversification. Whereas much of molecular phylogenetics thus far has relied on one or a few single-locus gene trees to represent species-level relationships (Gielly and Taberlet 1994, Fitch 1995, Alvarez and Wendel 2003), even in the absence of interspecific gene flow, single-locus gene histories (including those for non-recombining plastid genomes) may not coincide with an actual species tree (Pamilo and Nei 1988, Doyle 1992, Maddison 1997, Page and Charleston 1997, Rosenberg and Tao 2008). There are usually three major sources of gene tree discordance: deep coalescence due to incomplete lineage sorting; interspecific gene flow including introgressive hybridization and horizontal gene transfer; and misspecification of orthology due to reciprocal loss of duplicated gene copies after duplication events (Doyle 1992, Brower et al. 1996, Maddison 1997, Degnan and Rosenberg 2009).

Incomplete lineage sorting (ILS) is an outcome of the retention of ancestral polymorphism through one or more speciation events. During the process of speciation, it is possible that ancestral alleles, which were shared by populations of an ancestral taxon due to either large population size or balancing selection, failed to go to fixation at the time of lineage split (Brower et al. 1996). The ancestral polymorphism would thus be kept in the newly formed species, leading to a pattern of gene genealogy that does not reflect the actual species divergence process. This pattern has been identified in various groups of both plants and animals (Comes and Abbott 2001, Jakob and Blattner 2006, Pollard et al. 2006, Carstens and Knowles 2007, Edwards et al. 2008).

Hybridization/introgression and horizontal gene transfer can both cause inter-specific gene flow, leading to the breaking of lineage confines by certain genes that move across different genomes. Hybridization and introgression occur extensively in plants but are also found in animals (Cullingham et al. 2012, Toews and Brelsford 2012, Kane et al. 2013, Mao et al. 2013). The formation of new hybrid taxa, as well as the opportunity of adaptive divergence provided by introgression of certain alleles, could both facilitate speciation depending on the interactions between genetic drift and selection (Abbott et al. 2013). However, the transferred alleles may be lost, retained or fixed, yielding reticulate evolutionary processes that violate the assumption in traditional cladistics that relationships between taxa are strictly bifurcating, and thus making phylogenetic reconstructions more complex (Mallet 2005, Arnold 2006).

Gene duplication events produce paralogous gene copies that evolve independently of each other. In fact, gene and genome duplications are sources of functional diversity that are believed to play an important role in evolution (Taylor and Raes 2004, Van de Peer et al. 2009). Whereas orthologs (gene copies produced by speciation events) are useful for phylogenetic

analyses of species relationships, mixed analyses of orthologs and paralogs can confound estimation of species relationships. Sampling of paralogous genes from species included in a phylogenetic analysis could produce a gene tree that disagrees with the species tree. This has been particularly problematic in plant phylogenetics for analyses of nuclear genes.

In the face of the processes that generate gene tree-species tree discordance, numerous approaches have been developed to combine information from more than one gene through analyses of sequence data from multiple loci across the genome. One of them is based on total-evidence when concatenated data from multiple genes ('supermatrix') are used to estimate a tree as the 'species tree' (de Queiroz and Gatesy 2007). Concatenated data can yield a very well supported phylogeny, but this approach is based on a simple assumption that all genes share the same evolutionary history and variation in trees estimated from individual loci are just phylogenetic noise. Since this approach treats each gene equally, longer genes may have more effect on the results than shorter genes. When the sampled loci do share the same history, concatenated data can yield a well-resolved and well-supported tree. But when there is incongruence in gene histories, a supermatirx analysis would ignore this variation and interpreting the inferred tree may be complicated (e.g. which gene histories are recovered and which are not). For example, with an ancestral polymorphism, a well-resolved, concatenated plastid gene tree can result, which is not actually reflective of the true species history. A consensus-based super-tree approach is an alternative to the super-matrix approach. Individual gene trees are estimated and a consensus of those estimates is then calculated to represent the species tree (Sanderson et al. 1998). For example, gene tree parsimony analysis selects a species phylogeny that minimizes the incongruence between the gene trees and the species tree (Slowinski et al. 1997). This approach treats each gene equally but independently, and one can

have an idea of gene tree conflicts and the frequency of alternative relationships. Though branch support can be obtained by bootstrapping both lineages (resampling sequences within species) and characters within genes (Joly and Bruneau 2009), super-tree approaches as with super-matrix approaches, assume that genes share a common history. Again, interpreting the results of a super tree can become complicated when this not the case.

Coalescent theory, which was originally developed to model population genealogies, has been adapted for phylogenetic/phylogenomic investigations (Rannala and Yang 2003). The original coalescent model attempts to trace all the alleles of a gene in a population back in time until reaching a single ancestral copy, providing the gene genealogy (the coalescent) of those alleles within the population (Hudson 1990, Kingman 2000). A multi-species coalescent model has been derived from this model and applied to multiple populations connected by the species phylogeny. It generalizes the Wright-Fisher model of genetic drift to describe a probability distribution of random gene trees that evolve along the branches of a species tree (Degnan and Rosenberg 2009). Coalescences of gene lineages from different species can only occur earlier than the splitting of the species to which they belong. Because the Wright-Fisher model is based on non-recombining loci with constant effective population size (Ne), random mating and neutral evolution (no selection), the multi-species coalescent has the same assumptions.

Phylogenetic methods that use multispecies coalescent often use a likelihood or Bayesian framework to estimate the species tree based on the probability distributions of gene trees, and most of these methods assume gene tree discordance is only explained by the effect of random sorting of sampled alleles. Some methods take the full dataset with parameter-rich algorithms under a likelihood framework, including maximum-likelihood and Bayesian methods. For example, BEST (Liu 2008) implements a Bayesian hierarchical model to jointly estimate gene

trees and the species tree from multi-locus sequence data, and can also estimate divergence times and population sizes at the same time. A similar program *BEAST (Heled and Drummond 2010) differs from BEST in that it co-estimates the species tree and all gene trees in one Bayesian MCMC analysis, assuming the estimation to be better when using information from all gene trees simultaneously. Other methods use summary statistics of gene trees instead of full data. Examples of summary-statistic based methods include GLASS, STEM, STAR/STAEC, and MP-EST. The Global LAteSt Split (GLASS) method (Mossel and Roch 2010) clusters species using minimum coalescence based on the assumption that gene coalescences always occur earlier than species coalescence. Given a collection of gene trees, it calculates the minimum gene coalescence times for all pairs of species across genes and constructs a species tree using those minimum coalescence times. Assuming a constant mutation rate across genes within a species, GLASS can also be extended to use molecular distances to construct species phylogenies (Mossel and Roch 2010). STEM [Species Tree Estimation using Maximum Likelihood (Kubatko et al. 2009)] takes the same principle of species tree construction by minimum coalescence and searches the tree space using a likelihood algorithm. However, the program works well only when the input gene trees satisfy the molecular clock and constant rate across genes. STAR (Species Tree estimation using Average Ranks of coalescences) and STEAC (Species Tree Estimation using Average Coalescence times) both take a distance method in estimating a species tree (Liu et al. 2009b). STAR defines a new parameter called the ‘average ranks of coalescence’. The rank of coalescence at the root of the tree is defined as the total number of taxa in the tree, and the rank decreases by one as each node goes from the root to the tips. Under the coalescent model, sequences within the ancestral population of the species tree are equally likely to coalesce with each other, and the order of expected ranks of coalescence among sequences is

consistent with the order of populations in the species tree. Thus a Neighbor Joining (NJ) tree (Saitou and Nei 1987) can be reconstructed from a distance matrix of the average rank across all genes and all pairs of alleles between two species as an estimation of the species tree. STEAC differs with STAR in using a distance matrix of average coalescence times instead of average ranks across the gene trees, and it will output branch lengths while STAR only yields a species tree topology (Liu et al. 2009b). Recently, Liu et al. (2010) have proposed a new likelihood approach for estimating species trees under the coalescent model. Although the likelihood of a species tree under the multispecies coalescent model has been derived by Rannala and Yang (2003), Liu et al. (2010) showed that the maximum likelihood estimate (MLE) of the species tree (topology, branch lengths, and population sizes) from gene trees under this formula does not exist. They developed a pseudo-likelihood function to obtain maximum pseudo-likelihood estimates (MPE) of species trees, with branch lengths in coalescent units (Liu et al. 2010).

Although likelihood methods taking full dataset (such as BEST) have been shown in simulations to be robust in estimating species trees under high levels of gene tree incongruence and very low levels of horizontal gene transfer (Chung and Ane 2011, Leache and Rannala 2011), a major issue with these methods is the computational time and restricted utility with small datasets (Liu et al. 2009a), and both BEST and *BEAST are sensitive to missing data (Liu 2008, Heled and Drummond 2010). None of these methods are actually modeling gene flow, so even low rates of horizontal gene transfer (i.e. gene flow) could mislead analyses under some conditions. Given the complexity of the problem, only a small fraction of “tree/parameter space” may be actually sampled in the MCMC search without achieving the global optimum. On the other hand, summary statistics methods can be much more time-efficient with the capability to handle large datasets, particularly those in phylogenomic studies, although a larger sample size

may be required to achieve optimal confidence levels (Liu et al. 2009a). However, the accuracy of species tree inferences is not only a problem of which phylogenetic methods to choose, but also a matter of the species divergence history and sampling strategy (McCormack et al. 2009). Various studies using both simulations and empirical data have shown that the shape of the species tree, which is determined by the time of divergence and effective population sizes, as well as the number of genes and number of individuals sampled per species, could affect the performance of species tree estimations (Maddison and Knowles 2006, Liu et al. 2008, Liu et al. 2009b, McCormack et al. 2009, Huang et al. 2010, Leache and Rannala 2011). In fact, the so-called "anomaly zone" is defined by a rapid succession of speciation events that results in a bias in the ILS process causing the most common gene history to conflict with the species tree(Degnan and Rosenberg 2006, Kubatko and Degnan 2007, Degnan and Rosenberg 2009, Huang and Knowles 2009, Rosenberg 2013). In general, shallower divergence among large populations (short species tree with fat branches) is more likely to have ILS and thus causing more problems in species tree estimation; while sampling more individuals per species and/or more loci in the coalescent analyses can help increase informative phylogenetic signals in the dataset.

## 1.2 Genus *Psychotria* (Rubiaceae) as a Model Group to Study Species Diversification and Evolution

In plants, species-rich angiosperm families often provide good model taxa to study evolutionary diversification, as species-richness often depends on both the innate traits of the species and the diversity of environmental conditions. Rubiaceae is the fourth largest angiosperm family in the world with approximately 611 genera and 13000 species in three subfamilies, and

thirty genera have more than 100 species (Davis et al. 2009, WCSP 2013). Members of this family occur in nearly every major region of the world except continental Antarctica, the high arctic, and part of central Africa and Asia, with the highest diversity in the tropics, a pattern that is very similar to the global distribution of plant diversity overall (Davis et al. 2009). This family also comprises many economically valuable species, such as *Coffea arabica,* which ranks as one of the world's most valuable and widely traded commodities; *Cinchona officinalis* (quinine); *Psychotria spp.* [emetine (methylcephaeline) and cephaeline from *P. ipecacuanha* (Openshaw 1970)*;* dimethyltryptamine from *P.viridis* (Der Marderosian et al. 1969)]; *Rubia tinctoria* (madder); and various ornamental species in *Ixora, Gardenia, Mussaenda, Portlandia* and *Serissa*.

*Psychotria* (tribe Psychotrieae, subfamily Rubioideae) is the largest genus within Rubiaceae, also considered as the third largest angiosperm genus (Frodin 2004). *Psychotria* contains over 1800 accepted species, about 750 of which occur in the Americas, 350 in Africa, and 750 in Asian-Pacific regions (WCSP 2013). *Psychotria* species are mainly woody shrubs to small trees with small, insect-pollinated white flowers, and often comprise a large proportion of the understory vegetation in the wet tropical low-land forests over the world (Sohmer 1988). It has been considered as a model system for the study of speciation and evolutionary ecology in the tropics (Hamilton 1989a), especially for understanding factors driving speciation including niche specialization (Nepokroeff et al. 1999, Valladares et al. 2000, Sakai and Wright 2008) and neutral evolutionary processes (Hubbell 2001, Sedio et al. 2012, Sedio et al. 2013, Sterck et al. 2013).

On the other hand, the phylogeny within Rubiaceae is quite complex and poorly resolved at a variety of taxonomic levels. Over 50 phylogenetic studies have been published for this

family over the past two decades, but many relationships remain ambiguous (Davis et al. 2007, Razafimandimbison et al. 2008, Bremer 2009, Paul et al. 2009). Subfamily Rubioideae is the largest subfamily in Rubiaceae (ca. 7475 species), including the Psychotrieae alliance (Bremer and Manen 2000) with ca. 3000 species (Govaerts 2006). However, the internal relationships for this subfamily are problematic mainly due to the poor understanding of the relationships within the Psychotrieae alliance. Tribe Psychotrieae is a monophyletic group comprising 12 genera, but relationships within the tribe are still unclear. Robbrecht and Manen (2006) proposed to split Psychotrieae into two tribes, but results from Bremer and Eriksson (2009) did not support this division and they suggested the tribe be maintained as one. Within Psychotrieae, the genus *Psychotria* has been shown to be highly paraphyletic (Nepokroeff et al. 1999, Bremer and Manen 2000, Andersson 2002). Specifically, subgenus *Heteropsychotria* is paraphyletic with some species more closely related to *Palicourea* species than other *Psychotria* species, and the clade *Psychotria sensu stricto* is paraphyletic as species formerly assigned to subtribe Hydnophytinae are nested within this clade. In addition, the resolution and support values for the internal branches are usually low or not provided (Nepokroeff et al. 1999, Andersson 2002, Paul et al. 2009).

A major challenge in resolving the genus-wide phylogeny is the paucity of diagnostic characters for many groups within the family (Nepokroeff et al. 1999). This might be due partly to the relatively sparse sampling of this species-rich group, and probably also high intra-specific variation for many traits in terms of both morphology and DNA sequence polymorphism. Many morphological characters have been explored to compensate for the lack of variation in floral morphologies within the Psychotrieae complex, including the anatomy of pollen grain (Johansson 1993), stipules(Andersson 2002), petioles (Martínez-Cabrera et al. 2009), as well as a

number of leaf micromorphological characters such as domatia, vascular tissue arrangement of the petiole, epidermal characters, and presence/absence of styloid crystals (Moraes et al. 2011). However, these characters either partially contradicted with the relationships based on floral characters, or could only be appropriate for delimiting inter-generic relationships instead of differentiating between species. At the molecular level, at least 15 loci (Robbrecht and Manen 2006, Davis et al. 2007, Bremer 2009, Paul et al. 2009, Borhidi 2011, Barrabe et al. 2012, Sedio et al. 2012) have been used for the phylogenetic analysis in family Rubiaceae to date, either from the plastid genome (atpB-rbcL, ndhF, matK, rbcL, rps16 intron, trn(T)L-F, trnSG, accD-psaI psbA-trnH) or nuclear genome (ETS, ITS, nontranscribed spacer [NTS], pep-C large, pep-V small, Tpi, *nepGS*, *RPB2*). However, most phylogenetic studies of *Psychotria* and other Psychotrieae included analyses of fewer than three loci, with ITS and one or two plastid loci being most commonly used. The chloroplast genome evolves slowly relative to the plant nuclear genome (Wolfe et al. 1987), and therefore plastid markers may not provide enough phylogenetic signal for resolution of relationships within a rapidly and recently diversifying group such as *Psychotria.* Thus, more data are needed to help understand the story behind the diversity we see in this large taxon – the evolutionary history and mechanisms of species diversification in *Psychotria*.

### 1.3 Overview of Dissertation

In this study, my overall goal is to help answer a fundamental question: what are the evolutionary driving forces for the diversification of *Psychotria*? Specifically, is neutral evolution (random drift) sufficient to explain their patterns of diversity, or is any selective force necessary to be involved to result in their current relationships? Using a phylogenetics-based

comparative approach, I investigate the diversification on a sample group within two clades (subg. *Psychotria* and subg. *Heteropsychotria*) of the New World *Psychotria*.

In the second chapter of this dissertation, I report a number of non-coding sequence markers from both the nuclear and chloroplast genomes that were developed for *Psychotria*, and show their utility by assessing the sequence variations across sampled species and comparing those with the commonly used ITS and *trnL-trnF* regions. These sequence markers were developed from ETS data for both *Psychotria* and *Coffea*, as well as through chloroplast sequencing of *P. marginata*. I also report the whole chloroplast genome sequence of *P. marginata*, characterize the genome organization, and compare the sequence to the chloroplast genome sequence of *Coffea arabica*. The third chapter focuses on the multi-locus phylogenetic analysis to resolve the species-level relationships among a group of Central American *Psychotria*. With the non-coding markers described in the second chapter, single-gene phylogenies are reconstructed for each nuclear locus, and a chloroplast phylogeny is also estimated for the concatenated cp loci. I also conduct a multi-species coalescent analysis utilizing all markers to estimate a species phylogeny for the sampled group, and compare gene trees with the species tree to assess the level and causes of topological concordance. In the fourth chapter, I perform coalescence simulations to test hypotheses on incomplete lineage sorting and gene flow during sympatric versus allopatric species split.

# CHAPTER 2

# IDENTIFICATION OF VARIABLE NON-CODING LOCI FOR PHYLOGENETIC STUDIES IN *PSYCHOTRIA* (RUBIACEAE) WITH WHOLE CHLOROPLAST SEQUENCING[1]

[1] DONG, L. AND J.H. LEEBENS-MACK. To be submitted to *Molecular Ecology Resources*.
[2] DONG, L. AND J.H. LEEBENS-MACK. To be submitted to *Systematic Botany*.

## ABSTRACT

As whole chloroplast genome sequencing has become more routine, the identification of variable plastid sequence markers for phylogenetic studies as well as DNA barcoding have also become more straightforward by direct comparison of cp genomes across species from related taxa. Here we report the chloroplast genome sequence of *Psychotria marginata,* the second species within the fourth largest angiosperm family Rubiaceae, to have its cp genome sequenced. The *P. marginata* chloroplast genome is completely collinear with that of *C. arabica* in terms of gene content and gene order. However, three genes, *infA*, *accD* and *rpl33* were found to contain stop codons in the middle of their coding sequences, and thus may have lost their function in *Psychotria*. Pairwise comparison of *P. marginata* and *C. arabica* plastomes also allowed identification and testing of variable non-coding markers, and we developed 10 primer pairs flanking non-coding inter-genic regions or introns of the *P. marginata* cp genome. Sequence variations were assessed in 19 species of *Psychotria* and its sister taxa (*Rudgea*, *Palicourea* and *Coffea)*, showing much higher inter-specific polymorphisms than the widely used trnL-trnF region. The maximum likelihood phylogeny constructed from three of the 10 loci and *trnL-trnF* showed significant improvement in statistical support of internal branches compared with previous genus-wide phylogenies inferred from *rbcL* and ITS sequences. In all, these markers revealed significantly higher variation suitable for lower-level phylogenetic studies as well as potential utility as genus-specific barcoding marker candidates in family Rubiaceae.


## 2.1 INTRODUCTION

Chloroplast DNA sequences have been utilized as markers to resolve plant phylogenies since the early 1990's (Taberlet et al. 1991, Chase et al. 1993). Researchers have been actively

looking for chloroplast regions to use for various levels of phylogenetic/taxonomic investigations. Both coding sequences such as *rbcL*, *ndhF* and *matK* (Olmstead and Palmer 1994), and non-coding regions such as *trnL-trnF* and *atpB-rbcL* intergenic spacers (Taberlet et al. 1991, Golenberg et al. 1993, Gielly and Taberlet 1994) were among the first markers found to be suitable for assessing genetic diversity, and were subsequently used extensively in plant phylogenetic studies. Currently, cp sequence markers are still very commonly used in a wide spectrum of studies from population genetics, phylogeography to phylogenetics (Cruzan 1998, Newton et al. 1999, Jansen et al. 2007, Parks et al. 2009, Barniske et al. 2012). Specifically, noncoding regions of the chloroplast genome have often been used for population-level studies as well as lower-level phylogenies because they are assumed to have less functional constraint and are free of selective pressure, thus providing greater levels of variation than coding regions (Gielly and Taberlet 1994). However, despite the fact that over 30 non-coding chloroplast loci have been proposed over more than 20 years, generalizing the use of particular loci across a broad phylogenetic range is still difficult, mainly because different non-coding regions offer different levels of useful phylogenetic signal across a given group of taxa (Shaw et al. 2005, Shaw et al. 2007). Even the same loci can be informative for some taxa but not for others, or only show appropriate variations for higher-level comparisons but not among closely related taxa. As a result, most of studies on plant molecular phylogenetics have relied on a limited number of the most popular loci.

In addition to the use of non-coding sequence markers in molecular phylogenetics, population genetics and phylogeography, DNA barcoding, which takes advantage of the nucleotide diversity of certain short DNA segments in species identification processes, has become another field of application for non-coding cp markers and been much discussed over the

past decade (Hebert et al. 2003, Blaxter 2004, Chase et al. 2005, Kress et al. 2005, Rubinoff et al. 2006, Vijayan and Tsou 2010). Although the mitochondrial cytochrome C oxidase subunit I (COI) has been facing some challenges as a universal barcode locus in animals, it has shown sufficient polymorphisms to discriminate some groups of closely related species and is still relatively robust compared to the barcoding system in plants (Taylor and Harris 2012). Because plant mtDNA has a very slow substitution rate (Wolfe et al. 1987), COI cannot be used as a proper region for barcoding in plants. Researchers have been testing various sequences from both plant nuclear and chloroplast genomes in order to find suitable candidates for barcoding. Chase et al. (2007) initially proposed a two-option, three-region system with either 1). *rpoCl*, *rpoB* and *matK* or 2). *rpoCl*, *malK* and *psbA-trnH* as viable markers for land plant barcoding; and two years later they updated the system to the combination of rbcL and matK as a universal barcode system of land plants (Hollingsworth et al. 2009). Over the years, more than 10 different chloroplast loci in total (*rbcL, accD, matK, rpoB, rpoC1, ycf5, trnH-psbA, atpF-atpH, psbK-psbI,* and *trnL-trnF*) have been proposed as potential barcoding markers for plants by various additional groups of researchers (Vijayan and Tsou 2010). However, a major focus of barcoding studies in plants is still on assessing the relative efficiency of different markers, and empirical studies have shown the limitations of these markers both in their ability to be universally amplified and applicability in higher resolution circumscriptions of taxa (Vijayan and Tsou 2010). Therefore, it seems like the search for a specific barcoding system for plants will remain active in the near future (Taylor and Harris 2012).

The chloroplast genome generally ranges from 120-170kb in size (Palmer 1985, Sugiura 1992, Raubeson 2005) with limited variation among different plant taxa in its structure, gene content and gene order. Sequencing the whole chloroplast genome has greatly facilitated plant

phylogenetic studies, and with the innovation of next-gen high-throughput sequencing technologies in the late 1990's, sequencing chloroplast genomes is becoming increasingly common. To date, over 100 angiosperm chloroplast genomes have been sequenced with most orders represented (Cai et al. 2006, Jansen et al. 2007, Guisinger et al. 2010, Moore et al. 2010, Delannoy et al. 2011, Jo et al. 2011, Nie et al. 2012, Hand et al. 2013, Ku et al. 2013). These data have been, and will continue to serve as great resources for the generation of variable non-coding sequence markers for relevant studies. Given these recent technical advances and the rapid accumulation of data, we propose that it is probably easier and more straightforward nowadays to just search for non-coding cp markers that are more variable for specific groups of taxa rather than those that can be more universally applied, both for molecular phylogenetic studies and as potential candidates of DNA barcoding. With the convenience of cp genome sequencing and the increasing availability of cp genome sequence data from various taxa, genome-wide sequence comparisons between species from relatively closely-related clades can be a feasible and effective approach for simultaneously identifying multiple non-coding regions with relatively high level of variation, and those regions are also likely to be successfully applied to all the species within that particular group. A well-resolved phylogeny often serves as the basis for testing other biological hypotheses, such as evolutionary history, hybridization, character evolution, and biogeography. From this standpoint, it is probably more crucial to incorporate more polymorphic markers to have better resolutions within specific taxa so that relevant hypotheses can be critically addressed, rather than to search for more universal markers that might yield imperfect phylogenetic hypotheses that could bias further extrapolations. The same argument also applies to DNA barcoding, where species-specific markers designed for a certain group can be more useful in successfully differentiating species within genera, while those

generalist markers might only be good for family-level or inter-generic level resolution. In this study, we show that comparisons between *Psychotria marginata* and *Coffea arabica* chloroplast genomes yielded a number of highly variable non-coding sequence markers that can be useful for phylogenetic studies on closely related taxa at the genus level. Thus comparing the chloroplast genomes of *P. marginata* and *C. arabica* (Samson et al. 2007) allows identification and testing of non-coding regions with the highest levels of polymorphism as markers at the intra-familial level in family Rubiaceae.

In plants, species-rich angiosperm families often provide good model taxa to study evolutionary diversification, as species-richness often depends on both the innate traits of the species and the diversity of environmental conditions. Rubiaceae is the fourth largest angiosperm family in the world with approximately 13000 species in 611 genera, of which thirty genera have more than 100 species (Davis et al. 2009). This family comprises many economically valuable genera as well as popular ornamentals, such as *Coffea* which ranks as one of the world's most valuable and widely traded commodity; *Cinchona* with quinine in *Cinchona officinalis*; ipecacuanha produced in *Psychotria*, an expectorant; *Rubia tinctoria* which produces madder as a dye; and various ornamental species in *Ixora, Gardenia, Mussaenda, Portlandia* and *Serissa*. *Psychotria* is the largest genus within Rubiaceae, and is also considered to be the third largest angiosperm genus (Frodin 2004). It contains over 1800 accepted species, about 750 of which are in the Americas, 350 in Africa, and 750 in Asian-Pacific regions (WCSP 2013). *Psychotria* species are mainly woody shrubs to small trees with small, insect-pollinated white flowers, and some species (E.g., *Psychotria ipecacuanha* and *Psychotria viridis*) produce important alkaloids. With its remarkable species diversity and richness, *Psychotria* often comprises a large proportion of the understory vegetation in the wet tropical low-land forests over the world (Sohmer 1988),

and has been considered as a model system for the study of speciation and evolutionary ecology in the tropics (Hamilton 1989a).

Despite a long-time interest in the diversity and evolution of this huge group (Sohmer 1978, Sohmer 1988, Hamilton 1989c, b, a, Taylor 1996, Nepokroeff et al. 1999, Nepokroeff et al. 2003, Razafimandimbison et al. 2008), the inter-specific relationships within *Psychotria* have not been fully resolved due to a paucity of phylogenetically informative molecular variation among closely-related species. To date, at least nine different chloroplast loci (Davis et al. 2007, Bremer 2009, Paul et al. 2009) have been used for phylogenetic analyses (atpB-rbcL, ndhF, matK, rbcL, rps16, trn(T)L-F, trnSG, accD-psaI psbA-trnH) in family Rubiaceae. However, most of those studies on *Psychotria* and other species within tribe Psychotrieae only included analyses of fewer than 3 loci, with *rbcL* being the most commonly used plastid marker. Since the chloroplast genome, especially the coding regions, evolves very slowly relative to the plant nuclear genome (Wolfe et al. 1987), such plastid markers may not give enough phylogenetic signal for resolution of relationships within a rapidly and recently diversifying group like *Psychotria.* We sequenced the whole chloroplast genome of *P. marginata* and developed 10 markers from different regions of intron/inter-genic spacers. Here, we report the complete sequence of the *Psychotria* chloroplast genome, comparison of genome organization with *Coffea arabica,* and characterization of sequence variations for the 10 non-coding markers.

## 2.2 METHODS

### *DNA extraction and assessment of plastid DNA abundance*

Leaf material of a *Psychotria marginata* individual was collected from the UGA Plant Biology Greenhouses (Athens, GA) and kept at -80°C until isolations were made. Genomic DNA

was extracted using a CTAB protocol modified from Doyle & Doyle (1987). The percentage of plastid DNA was assessed using quantitative PCR (qPCR) with genomic DNA template. The rbcL gene specific primers were used (RbcL_120F: 5'TGGCAGCATTYCGAGTAACT3' RbcL_230R: 5'ACGATCAAGRCTGGTAAGTC3', 130 bp amplicon). qPCR was performed using SYBR® GreenERTM qPCR SuperMix Universal (Invitrogen) in a Mastercycler® ep Realplex (Eppendorf) with 3 technical replicates of *P. marginata* sample. The qPCR cycling program was set as: 50°C 2 min, 95°C 10 min, (95°C 15s, 58°C 15 s, 68°C 20 s) × 45 followed by a melting curve analysis.

*DNA sequencing, genome assembly and annotation*

The *P. marginata* chloroplast genome was sequenced using Illumina short-read sequencing technology, which produced a total of 7576049 single-end reads of 71bp in length. The short reads were assembled using both reference-based program Yasra (Ratan 2009), with *Coffea arabica* as a reference genome, and a *de novo* assembly program Velvet (Zerbino and Birney 2008). About 1% of the reads were put into the assembly, and gap regions were filled by re-sequencing PCR fragments generated from primers designed in the flanking regions. The resulting contigs from both programs were merged in SEQUENCHER v4.2 (GeneCodes, Ann Arbor, MI, USA) and finalized into a single linear sequence of the complete chloroplast genome. The genome sequence was annotated using the Dual Organellar GenoMe Annotator (DOGMA) pipeline (Wyman et al. 2004). Initial BLASTX and BLASTN searches against a custom database of previously published chloroplast genomes identified putative protein-coding genes, tRNAs and rRNAs. For genes with low sequence identity, BLAST searches were performed to identify the positions of the start/stop codons and the translated amino acid sequences for manual annotations.

*Identification of non-coding sequence markers*

The cp genome of *P. marginata* and *C. arabica* (NC_008535) were pairwise aligned using MULAN (Ovcharenko et al. 2005) to identify regions suitable for primer design. Regions of introns and/or inter-genic spacers with high levels of variation as visualized in the alignment were identified manually, and forward and reverse primers were designed in the flanking coding regions using Prima-clade Primer Visualization (Gadberry et al. 2005) and Primer3 (Rozen and Skaletsky 2000) for 10 loci which displayed polymorphism in comparisons between the two genomes. By doing this, we assume that there is a positive correlation between sequence variations at the inter-generic level (between *Coffea* and *Psychotria*) and that at the genus level (between species within *Psychotria*).

Sixteen *Psychotria* species and three other species from related genera (*Palicourea*, *Rudgea* and *Coffea*) collected from Costa Rica and UGA Plant Biology greenhouses, ranging from one to five individuals per species, were used to test the primer sets (Table 2.1). Genomic DNA was extracted from each individual using the modified CTAB protocol. PCR was performed for each primer set in a 25 µl reaction mixture with 2 µl of 10× PCR buffer, 25 mM $MgCl_2$, 30 - 100 ng genomic DNA, 10 µM of each primer, 10 mM dNTPs, and 1 U *Taq* DNA polymerase (lab-made). The PCR program consisted of an initial denaturing step at 94°C for 5 min, followed by 30 cycles of amplification (94°C for 1 min, 54 °C for 30sec, 72°C for 45 sec), and a final elongation step at 72°C for 5 min. Amplification products were resolved by electrophoresis in 1.5% agarose gels and visualized by EtBr staining or GelRed™ Nucleic Acid Gel Stain (Biotium). Primer sequences, GenBank accession numbers, and allele size ranges are shown in Table 2. All 10 primer sets were first tested in *P. marginata, P. horizontalis* and *C. arabica.* Three of the ten loci were selected to perform further phylogenetic analysis across all

sampled taxa in Table 2.1. PCR products were cleaned up by adding 0.75 µl SAP (shrimp-alkaline phosphatase; New England Biolabs Inc. USA) and 0.25 µl EXO (Exonuclease I; New England Biolabs Inc. USA) per 8 µl PCR product and digest at 37 °C for 60 min, followed by 80 °C for 25 min in a thermocycler. Sanger sequencing reactions were then performed on these cleaned-up PCR products using their corresponding forward and reverse primers in separate reactions. The reaction mixture contained 1 µl PCR product, 0.6 µl forward or reverse primer, 0.5 µl BigDye® Terminator v3.1 (Applied Biosystems, Life Technologies Corporation, USA) and 2.0 µl 5x sequencing buffer (Applied Biosystems, Life Technologies Corporation, USA) in 10 µl reaction volume. The cycle-sequencing program consisted of 94 °C for 2 min, followed by 35 cycles of 94 °C for 20 sec, 50 °C for 10 sec and 60 °C for 1 min 30 sec, with a final extension at 60 °C for 8 min. The products were sequenced using ABI 3730 DNA analyser at Georgia Genomics Facility (GGF) at the University of Georgia, Athens, GA.

*Phylogenetic Analysis*

The sequences obtained were visualized and edited using SEQUENCHER v4.2 (GeneCodes, Ann Arbor, MI, USA). Sequences were aligned using MUSCLE (Edgar 2004), and imported into Geneious R6 (Biomatters, available from http://www.geneious.com/) to visualize levels of variation (as percent of sequence identity). To test the assumption that sequence variation between *Coffea* and *Psychotria* reflects the variation among species within *Psychotria*, correlations between the inter-generic sequence variation and the intra-generic sequence variation for each locus were calculated using MS Excel 2011 (Mac) for pairwise comparisons between *P. marginata* and *P. horizontalis*, as well as between *P. marginata* and *C. arabica*. The *trnK-rps16*, *trnT-psbD* and *psbE-petL* IGS were able to be amplified in the largest proportions of species sampled while showing high level of variation among *P. marginata, P. horizontalis* and

*C. arabica*. Those three regions were chosen to conduct phylogenetic analysis for all sampled species. The widely used *trnL-trnF* region was also added as a comparison. Maximum likelihood trees were generated using RAxML (Stamatakis 2006) for the concatenated dataset of trnK-rps16, trnT-psbD, psbE-petL and trnL-trnF, with GTR + Γ substitution model (Lanave et al. 1984, Yang 1994). Statistical support of internal branches was estimated by bootstrap analysis of 500 replicates. FigTree (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the output tree.

## 2.3 RESULTS

*Genome Organization*

The circular chloroplast genome of *Psychotria marginata* is 153430 bp in total length, and has the typical structure of land plant plastomes (Figure 2.1), including a large (LSC, 84251 bp) and a small (SSC, 17607 bp) single-copy region separated by two inverted repeats (IRs, 25779 bp each). The overall GC content of the chloroplast genome is about 37.6%, very similar to coffee (Samson et al. 2007), tomato (Kahlau et al. 2006), tobacco (Shinozaki et al. 1986), and even rice and maize (Hiratsuka et al. 1989, Maier et al. 1995, Tang et al. 2004). The genome contains a total of the 128 genes, of which 90 are present in the single copy regions and 19 are duplicated in the IRs. The coding region includes 76 protein genes, 29 tRNAs and 4 rRNAs, making up 55% of the total genome sequence. There are 18 intron-containing genes, including 12 protein-coding genes and 6 tRNAs. Fifteen of them have one intron and three of them have two introns, and these intron regions comprise about 26.1% of the total non-coding sequences. Inter-genic spacers (IGS) make up about 29.2% of the genome, ranging from 1 to 1787 bp. Only five IGS regions are longer than 1000 bp [*rps16-trnQ (UUG)*, *atpH-atpI*, *rpoB-trnC* (GCA),

*petN-petM*, *trnT(GGU)-psbD*, *ndhC-trnV(UAC)*, *psbE-petL* and *rps12-trnV (GAC)]*, and all but *rps12-trnV(GAC)* are located in LSC.

Whole-genome sequence comparisons (Figure 2.2) between *Psychotria* and *Coffea* (Samson et al. 2007) showed that these two chloroplast genomes were totally co-linear in terms of gene content, and as expected, coding regions have higher sequence similarity compared to IGS and introns, which indicates that non-coding regions contain higher sequence variation that could provide informative signal than coding regions in assessing diversity among species. In addition, *P. marginata* has part of the *rps19* gene duplicated at the IRa–LSC boundary as a result of expansion of the IR, the same pattern as in *Coffea* (Samson et al. 2007) and also members of the related family Solanaceae (Chung et al. 2006). Interestingly, although *infA* was shown to be functional in *Coffea* (Samson et al. 2007), we observed in *Psychotria* that *infA* is a pseudogene (6 internal stop codons), which is consistent with members of Solanaceae. Furthermore, two more genes, *accD* and *rpl33* also contain stop codons (one and four, respectively) in the middle of the ORFs so they may have become pseudogenes as well. This contrasts to most angiosperm chloroplast genomes in which these two genes are usually functional.

*Assessment of the 10 non-coding loci*

Primer sequences, fragment sizes and GenBank accession numbers for the 10 primer pairs are described in Table 2.2. Of the 10 primer pairs, six were amplified in both the two *Psychotria* species and *C. arabica*; three were amplified within *Psychotria*. The *ndhF-rpl32* sequence only amplified in *C. arabica* and thus was excluded from further analysis. These primers amplified a total of 9809 bp sequences in non-coding regions, constituting 14.2% of the total non-coding sequences as well as 6.4% of the *P. marginata* chloroplast genome size. Sequence variation in the pairwise alignments between *P. marginata* and *P. horizontalis* for intra-

generic comparison; and between *P. marginata* and *C. arabica* for inter-generic comparison are shown in Table 2.3. Intra-generic sequence identity ranged between 85% and 98.5%, while inter-generic sequence identity ranged from 42.8% to 79.9%. Compared to 95.6% intra-generic identity and 75.8% inter-generic identity for *trnL-trnF*, our data showed that all nine loci with sequences in the three species had a higher or at least similar level of polymorphism both within the genus and between genera. Moreover, the sequence length for each locus (Table 2.2) is significantly longer than that of the popular *trnL-trnF* (~500bp), providing a larger absolute number of variable sites that could potentially enhance the resolution of phylogenetic analyses. The Pearson's correlation coefficient between the inter-generic sequence variation and the intra-generic sequence variation was 0.7558, with a p value of 0.0085, showing a very significant positive correlation between the two. These results indicate that sequence polymorphisms between higher taxnomic level species can be used to predict that of lower taxonomic level species for the purpose of sequence marker development.

The *psbE-petL, trnK-rps16* and *trnT-psbD* regions were used to perform further sequencing and phylogenetic analyses across all the sampled species. When comparing with the widely used *trnL-trnF* region, all three loci showed higher variations especially for *trnK-rps16* and *trnT-psbD*; and the sequences were all at least twice the length of the *trnL-trnF* region (Fig 2.3). Sequence alignments of the four loci were concatenated to construct a phylogeny for the 19 species of Rubiaceae, as shown in Figure 2.4. The majority of branches have a bootstrap support of 90% - 100% except those that are near the tips of the tree, indicating very close relationships among these species. Although the two major clades representing subgenera *Psychotria* and subgenus *Heteropsychotria* had 100% bootstrap support, clear pattern of rapid radiation was shown within in the *Heteropsychotria* clade. In addition, paraphyly of both the *Heteropsychotria*

clade as well as the genus *Psychotria* was supported with high bootstrap values. These topologies were all consistent with the genus level phylogeny of Nepokroeff et al. (1999), which was constructed with ITS and *rbcL* data，but with significant increase of bootstrap support values for the major clades as well as some internal branches. In particular, 100% bootstrap support values for all 5 species sampled from subgenus *Psychotria* clade were obtained, whereas the relationships among the New World species of the same clade mainly had zero bootstrap support in the phylogeny of Nepokroeff et al. (1999).

## 2.4 DISCUSSION

*Genome Organization*

The gene content of plastid genomes is highly conserved among land plants, although gene and intron losses have been identified in some lineages (Raubeson 2005, Jansen et al. 2007). The basal lineages of angiosperms have a repertoire of 129 genes in the plastid genome, a conserved number across major angiosperm clades (Jansen et al. 2007). However, more than 60 gene or intron losses have been found along more derived lineages of monocots and eudicots (Jansen et al. 2007).

The *P. marginata* plastid genome falls within the normal range of genome size among angiosperms, and also shows a consistent gene order with that of *C. arabica*. On the other hand, three genes (*accD*, *rpl33* and *infA*) were shown to have been lost in the genome. The *infA* (translation initiation factor 1) gene has been shown to be highly unstable across angiosperms and is the most common gene to be lost in the plastid genome, with 11 independent losses in a 64-taxon sample of angiosperms according to Jansen et al. (2007) and 24 losses among a 309-taxon sample by Millen et al. (2001). Here we show that *infA* in *P. marginata* is a pseudogene as

opposed to the intact gene in *C. arabica* where (Samson et al. 2007). This result is consistent with previous studies where variation was found within Rubiaceae in terms of the presence or absence of the *infA* gene (Millen et al. 2001). In contrast, *accD* (acetyl-CoA carboxylase beta subunit) is apparently relatively more stable, and Jansen et al. (2007) mapped only six independent losses of this gene in monocots and eudicots, with the one in *Jasminum* (Oleaceae) being the most closely related lineage to *Psychotria* (Rubiaceae). There has been evidence suggesting that *accD* is an essential gene for leaf development in tobacco (Kode et al. 2005) as well as embryo development in *Arabidopsis* (Bryant et al. 2011). We observed a stop codon at AA position 14 of the gene in *P. marginata* (where it is a Q in *Coffea*), rendering it not functional. This result suggests that there may also be variation in the presence/absence of the *accD* gene within family Rubiaceae. Studies have shown potential functional replacement of *accD* by recent transfers of this plastid gene to the nucleus in Fabaceae (Magee et al. 2010) and Campanulaceae (Rousseau-Gueutin et al. 2013). Therefore, further research is needed to test if an alternative version of *accD* exists in the nuclear genome of *Psychotria*, although detailed investigation of this gene is beyond the scope of this particular study. Finally, the *rpl33* (ribosomal protein L33) gene is also identified as a pseudogene in *P. marginata,* although it has only been reported to be absent in common bean (*Phaseolus vulgaris*) (Guo et al. 2007) among angiosperms to date. We found four internal stop codons in the translated protein sequence for this gene in *P. marginata*. Although it seems to be a very conserved gene across angiosperm lineages, knockout studies in tobacco (Rogalski et al. 2008) showed that it is a non-essential ribosomal protein under normal growth conditions, but is required to sustain sufficient translation capacity under cold stress. It is possible that this gene is degraded in *Psychotria* because the possibility of the species undergoing cold stress is quite low given that they are

grown in tropical environments. But again, further researches on this particular gene across multiple species in *Psychotria* as well as other tropical taxa are needed to test this hypothesis.

*Utility of the Non-coding Markers*

By comparing the chloroplast genomes between *Coffea* and *Psychotria*, we demonstrated that sequence variation at higher taxnonomic levels can be used to successfully predict sequence variation at lower taxonomic levels. A significant correlation coefficient at $P<0.01$ suggested a strong positive correlation between the inter-generic sequence variation and intra-generic sequence variation. This is particularly useful for the development of non-coding cp sequence markers to be applied to non-model organisims, because with the strong predicting power by the genome-wide sequence comparisons between two relatively closely-related taxa, it is very straightforward to look for regions with high polymorphisms to be used as markers within a more specific taxonomic range. The fast accumulating chloroplast genome sequence data on NCBI makes it even easier to have immediately available sequences for the taxa of interest. Before the cp genome of *Psychotria* was sequenced, *Coffea arabica* was the only species in Rubiaceae with a complete chloroplast sequence. Therefore, the plastome data of *Psychotria* is an important additional contribution to the cp genome data pool of Rubiaceae, with the potential to be useful for comparative studies and marker development of other groups within this family.

With the same thermo-cycling parameters across primer sets, all polymorphic loci developed from this study were easy to amplify, and thus multiplex analyses are feasible in future studies with larger sample sizes. The phylogenetic tree from the maximum-likelihood analysis suggests that these sequence markers can be useful to resolve inter-specific relationships within genus *Psychotria* and other closely related taxa (Figure 2.4). With longer sequences and higher levels of variation offered by these loci, a better resolved tree with general improvements

on bootstrap support values was obtained compared to the Nepokroeff et al. phylogeny (1999), indicating that these markers can be expected to serve as valuable resources for other intra-family level studies of Rubiaceae, and also have the potential to be candidates of genus-specific barcoding markers for this family as well.

**Table 2.1** List of 19 species used in this study. Numbers of individuals sampled per species, location and voucher information for each species are included.

| Species | # of Indv. | Source & Voucher Info |
|---|---|---|
| *Psychotria chiapensis* Standl. | 2 | Dong, L. #001, Costa Rica (LSCR) |
| *Psychotria cyanococca* Dombrain | 2 | Dong, L. #002, Costa Rica (LSCR) |
| *Psychotria. brachiata* Sw. | 2 | Dong, L. #003, Costa Rica (LSCR) |
| *Psychotria surrensis* Donn.Sm. | 1 | Dong, L. #004, Costa Rica (LSCR) |
| *Psychotria calidicola* C.M.Taylor | 2 | Dong, L. #005, Costa Rica (LSCR) |
| *Psychotria pilosa* Ruiz & Pav. | 5 | Dong, L. OTS-LCHERB-002235, Costa Rica (HLDG) |
| *Psychotria chagrensis* Standl. | 2 | Dong, L. #007, Costa Rica (LSCR) |
| *Psychotria. panamensis* Standl | 1 | Dong, L. #008, Costa Rica (LSCR) |
| *Psychotria. allenii* Standl. | 2 | Dong, L. #009, Costa Rica (HLDG) |
| *Psychotria.steyermarkii* Standl. | 3 | Dong, L. OTS-LCHERB-002236 , Costa Rica (HLDG) |
| *Psychotria graciliflora* Benth. | 4 | Dong, L. OTS-LCHERB-002232, Costa Rica (HLDG) |
| *Palicourea padifolia* (Willd. ex Schult.) C.M.Taylor & Lorence | 3 | Dong, L. OTS-LCHERB-002231, Costa Rica (HLDG) |
| *Psychotria macrophylla* Ruiz & Pav. | 2 | Dong, L. OTS-LCHERB-002230, Costa Rica (HLDG) |
| *Rudgea skutchii* Standl. | 3 | Dong, L. OTS-LCHERB-002237, Costa Rica (HLDG) |
| *Psychotria officinalis* (Aubl.) Raeusch. ex Sandwith | 2 | Dong, L. OTS-LCHERB-002234, Costa Rica (HLDG) |
| *Psychotria. mortoniana* Standl | 2 | Dong, L. OTS-LCHERB-002233, Costa Rica (HLDG) |
| *Psychotria marginata* Sw. | 2 | Dong, L. #018, University of Georgia (GA) |
| *Psychotria horizontalis* Sw. | 1 | Dong, L. USCH-113890, University of South Carolina (USCH) |
| *Coffea arabica* L. | 1 | Dong, L. #019, University of Georgia (GA) |

**Table 2.2** Primer pairs developed for 10 polymorphic loci in the *P. marginata* plastome. Shown are: forward (F) and reverse (R) primer sequences, size of the PCR product and GenBank accession numbers. All primer pairs amplify at 54 °C annealing temperature. Primer pair 1 and 2 amplify from the small single copy region, 3-10 from the large single copy region. * = loci used in the phylogenetic analysis.

| Locus | Primer | Size Range (bp) | GenBank |
|---|---|---|---|
| ndhF-rpl32 | F: 5'-CGTYGGAAAAAGAAGAAGTCC-3' | 905 (*C. arabica*) | |
| | R: 5'-AAGACTGTCCAATATCCSTTC-3' | | |
| rpl32-trnL | F: 5'-GATCGAACATCAATTGCAAC-3' | 720-939 | |
| | R: 5'-GAAACGATGYGGTAGAAAGC-3' | | |
| trnK-rps16* | F: 5'-AAAGCCGAGTACTCTACCGTTG-3' | 999-1021 | |
| | R: 5'-ATTGATGTTCGATCCCGAAG-3' | | |
| atpF intron | F: 5'-TGCTTCCRTTTCCACTTTCC-3' | 784-789 | |
| | R: 5'-ATTTCAAGAATAGGCTGGATC-3' | | |
| rpoB-trnC | F: 5'-CAAACCCTGATCAATAAACC-3' | 1115-1147 | |
| | R: 5'-ATCAGGCGACACCCGGATTTG-3' | | |
| trnY-trnT | F: 5'-CTGGATTTGAACCAGCGTAG-3' | 1058-1182 | |
| | R: 5'-CGGATTTGAACCGATGACTT-3' | | |
| accD-psaI | F: 5'-CCAGAAGGGTTTAKTCGACCT-3' | 813-844 | |
| | R: 5'-TTGCAATTGCCGGAAATACT-3' | | |
| psbE-petL* | F: 5'-GGTGCTGACGAATAGCCAAC-3' | 1165-1223 | |
| | R: 5'-GAGGTTATAGTTAAAGCTGC-3' | | |
| trnT-psbD* | F: 5'-TCGGTTCAAATCCGATAAGG-3' | 1261-1407 | |
| | R: 5'-GTCCCTACGTAACCAGTCAT-3' | | |
| psaJ-rps18 | F: 5'-ATATCTCTCCGTGGCACCAG-3' | 899-939 | |
| | R: 5'-TCAATTMGATCCCCCGATTG-3' | | |

**Table 2.3** Sequence variation from the 9 plastid loci of *P. marginata, P. horizontalis* and *C. arabica*. Pairwise alignments were made between *P. marginata* and *P. horizontalis* for the intra-generic comparison, and between *P. marginata* and *C. arabica* for inter-generic comparison. The last locus is trnL-trnF sequence as a reference.

| CP locus | Inter-generic Identity | Intra-generic Identity |
|---|---|---|
| trnK-rps16* | 42.80% | 89.20% |
| atpF intron | 79.90% | 98.50% |
| rpoB-trnC | 69.30% | 92.50% |
| trnY-trnT | 43.20% | 87.20% |
| rpl32-trnL | 47.90% | 88.10% |
| accD-psaI | 69.50% | 96.80% |
| psbE-petL* | 75.50% | 95.70% |
| trnT-psbD* | 66.20% | 85% |
| psaJ-rps18 | 71.70% | 96.70% |
| trnL-trnF | 75.80% | 95.60% |

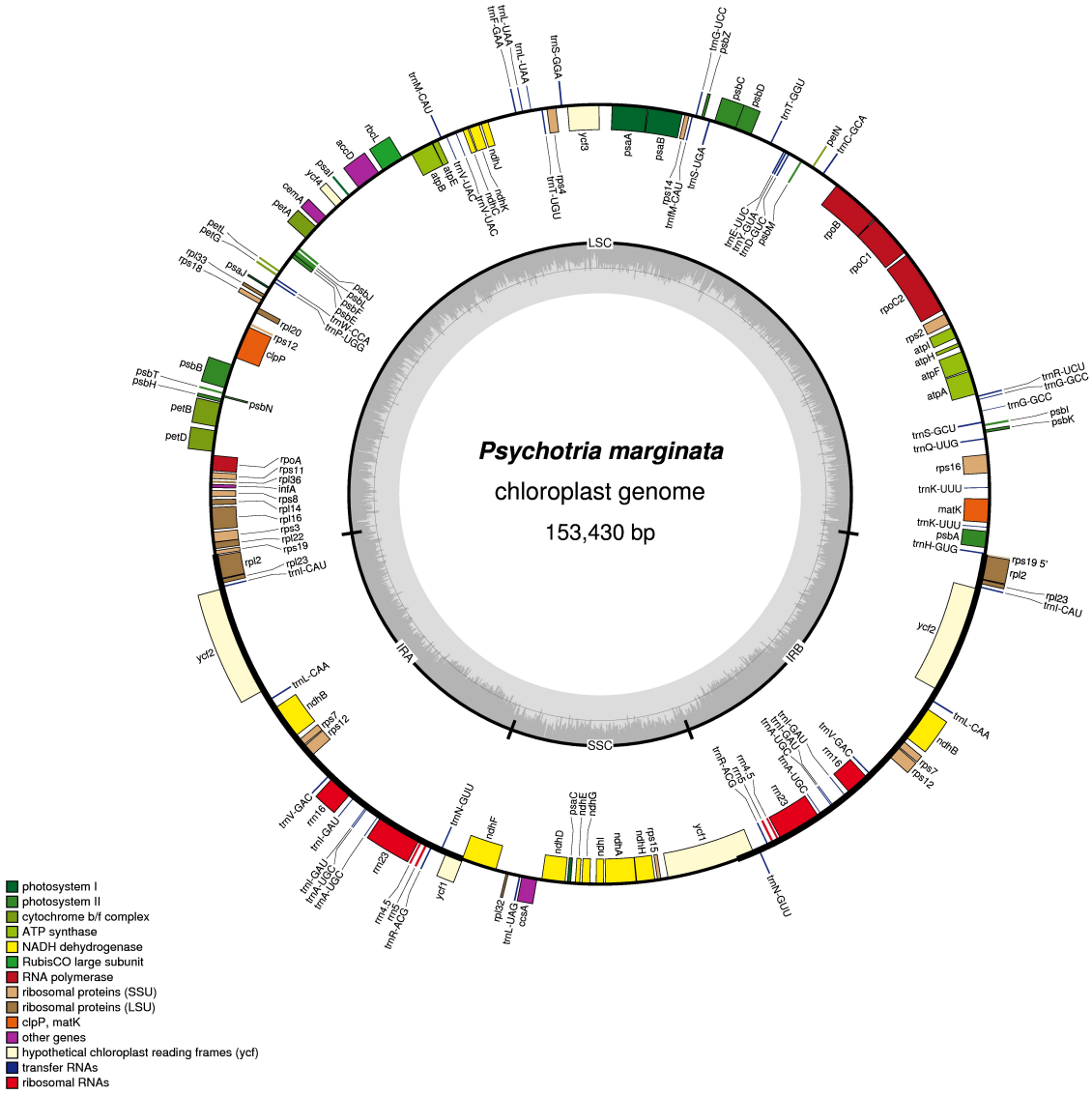


**Figure 2.1** Circular gene map of the *Psychotria marginata* chloroplast genome generated in this study. The 25779 bp inverted repeat regions (IRA and IRB) are shown in thick lines, separating the small (SSC, 17607 bp) and large (LSC, 84251 bp) single-copy regions. Genes on the inside of the map are transcribed in the clockwise direction and genes on the outside of the map are transcribed in the counterclockwise direction. The rps19 gene locates entirely in the IRA region and partly in the IRB region, while the ycf1 gene locates entirely in IRB but is truncated in IRA. The gray bars in the inner-most circle indicate levels of GC contents across the genome.

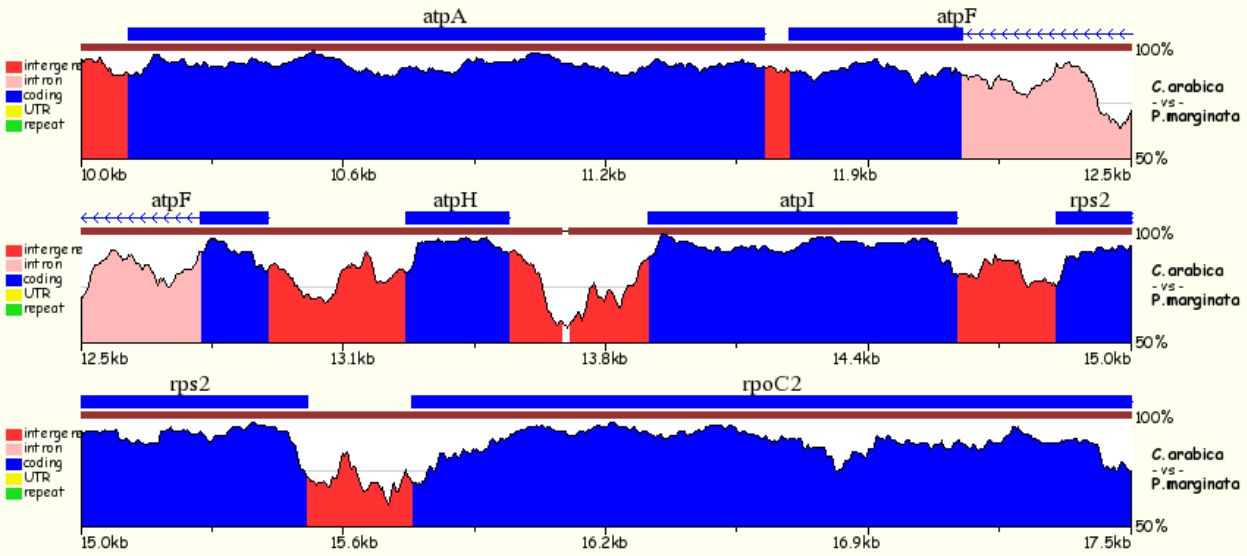


**Figure 2.2** Partial dynamic view of the whole plastome alignment of *P. marginata* and *C. arabica* (NC_008535.1). Coding regions are shown in blue with the gene names on top, and non-coding regions are shown in salmon (introns) or red (inter-genic spacers). Levels of sequence similarity (50% - 100%) are shown by height. The most variable non-coding regions were identified for marker development.

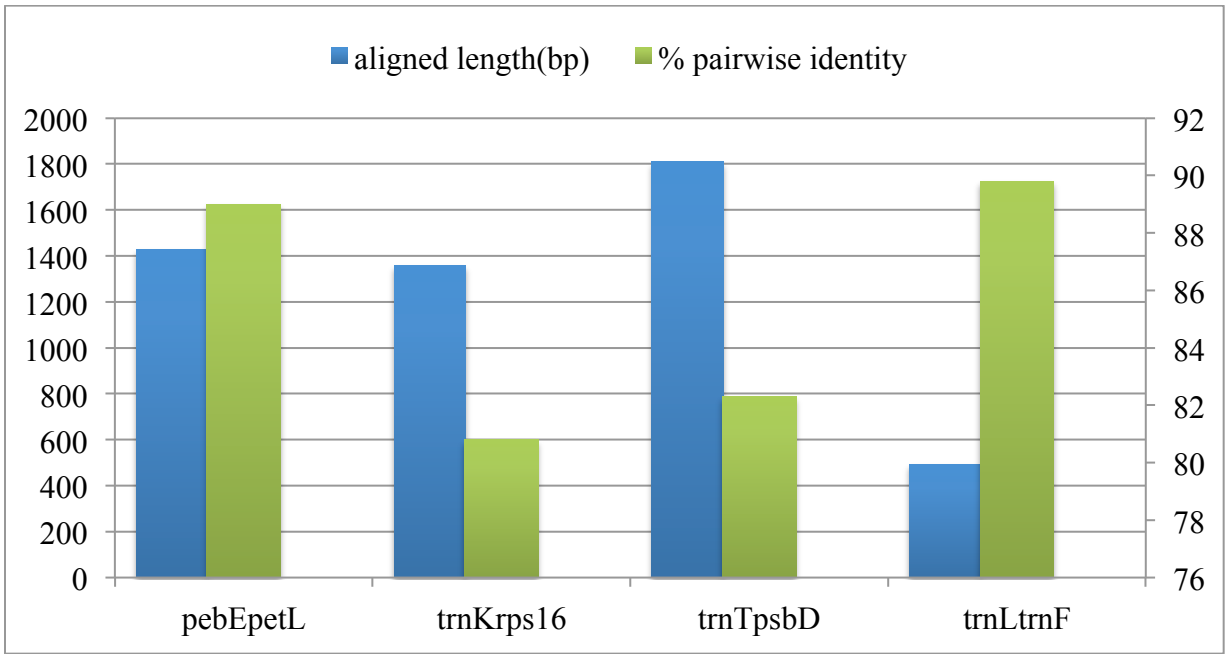


**Figure 2.3** Analysis of sequence variation across all 19 species for the three markers used in the phylogenetic analyses. The fourth locus is the trnL-trnF region as a comparison. Sequence lengths (in alignment) and percentage of identity are shown on the right and left Y axes, respectively.

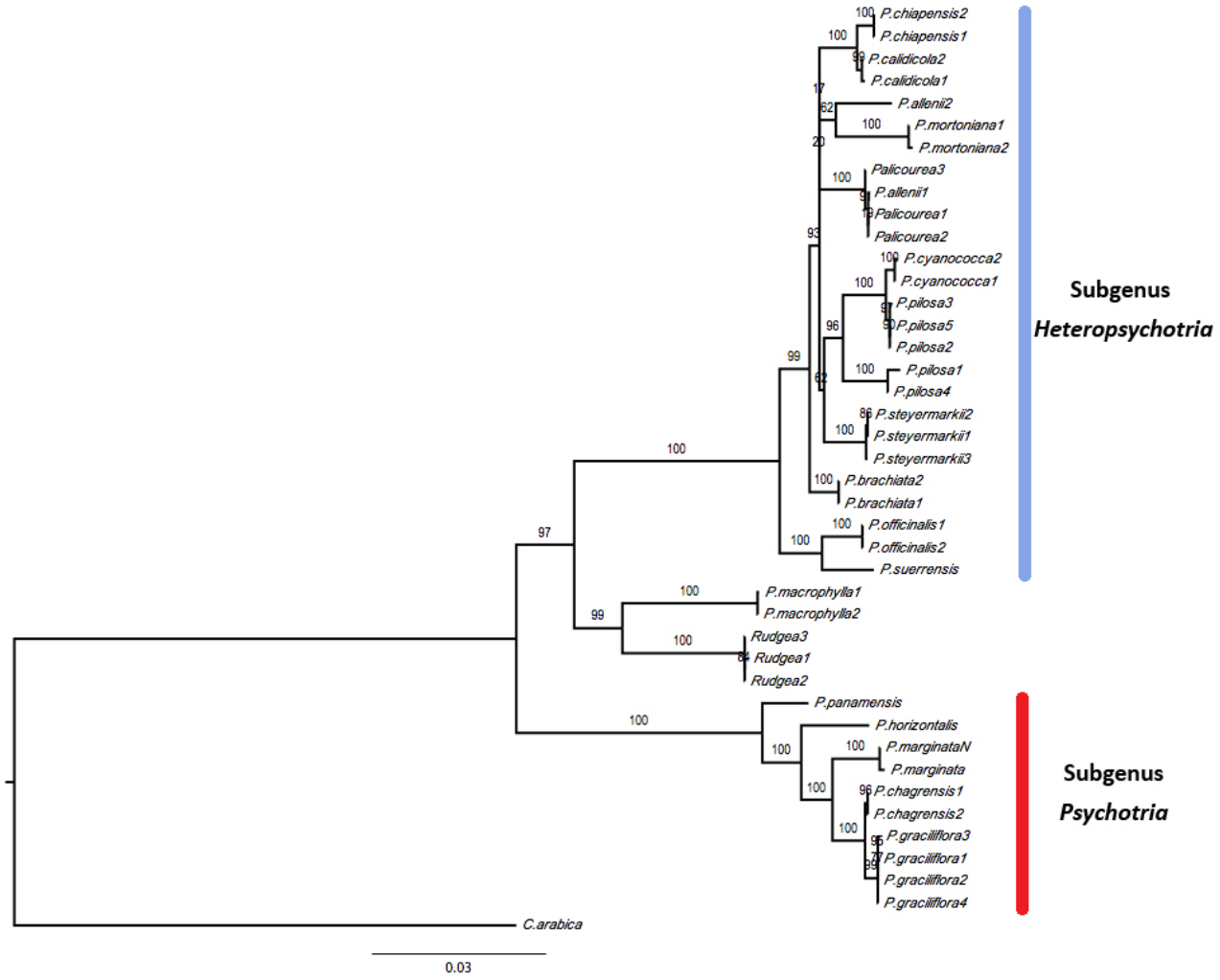


**Figure 2.4** Maximum-likelihood tree based on four concatenated chloroplast loci. The numbers after the names of a taxon indicate multiple accessions. Numbers above the internal branches are bootstrap support values based on 500 bootstrap replicates.

# CHAPTER 3

# MULTI-LOCUS PHYLOGENETIC ANALYSES ON A GROUP OF NEW WORLD *PSYCHOTRIA* (RUBIACEAE)[2]

[2] DONG, L. AND J.H. LEEBENS-MACK. To be submitted to *Systematic Botany*.

## ABSTRACT

Multi-locus phylogenetic analyses were performed for 18 species of neo-tropical *Psychotria* and its sister taxa *Rudgea* and *Palicourea*, using 5 nuclear (ITS, NL-57800, NL-217, NL-103, NL-A04) and 4 plastid ((*psbE-petL*, *trnT-psbD*, and *trnK-rps16* and *trnL-trnF*) non-coding sequence markers. Maximum-likelihood gene tree estimates as well as coalescent species-tree analyses both showed significant improvement of bootstrap support for the species relationships within *Psychotria*, though the tip branches still had low bootstrap values, further supporting a rapid radiation of the neo-tropical *Psychotria* species. Paraphyletic pattern of *P. pilosa* haplotypes on some of the gene trees strongly indicates incomplete lineage sorting; while the polyphyletic pattern of *P. allenii* haplotypes across all gene trees suggests the possibility of cryptic species complex as a result of rapid evolution. The subgenera-level relationships as well as the paraphyly of subgenus *Heteropsychotria* were well supported in both the gene trees and the species tree, and were also consistent with previous studies.


## 3.1 INTRODUCTION

The evolutionary history of species diversification includes a dramatic increase of complexity within and among organisms over millions of years. Understanding current biodiversity and the processes that promote speciation is a major goal for ecologists and evolutionary biologists. Species-rich angiosperm families often provide good model taxa to study evolutionary diversification, as species-richness often depends on both the innate traits of the species and the diversity of environmental conditions. *Psychotria* (tribe Psychotrieae, subfamily Rubioideae) is the largest genus within the family Rubiaceae, the fourth largest angiosperm family in the world (Govaerts et al., 2006). The genus contains over 1800 accepted species to

date, about 750 of which in the Americas, 350 in Africa, and 750 in Asian-Pacific regions (WCSP 2013); and is also considered as the 3rd largest flowering plant genus in the world (Frodin 2004). *Psychotria* species are largely woody shrubs to small trees with small, insect-pollinated white flowers, and some species produce important alkaloids (E.g., emetine (methylcephaeline) and cephaeline from *P. ipecacuanha*). With its remarkable species diversity and richness, *Psychotria* often comprises a large proportion of the understory vegetation in the wet tropical low-land forests across the world (Sohmer 1988), and has been considered as a model system for the study of speciation and evolutionary ecology in the tropics (Hamilton 1989a), especially for understanding factors driving speciation including niche specialization (Nepokroeff et al. 1999, Valladares et al. 2000, Sakai and Wright 2008) and neutral evolutionary processes (Hubbell 2001, Sedio et al. 2012, Sedio et al. 2013, Sterck et al. 2013).

Understanding phylogenetic relationships can be a key step in inferring patterns of evolution, but the phylogeny within Rubiaceae has been quite complex and poorly resolved at a variety of taxonomic levels. Over 50 phylogenetic studies have been published for this family over the past two decades, but many relationships remain ambiguous (Davis et al. 2007, Razafimandimbison et al. 2008, Bremer 2009, Bremer and Eriksson 2009, Paul et al. 2009). This is mainly due to the poor understanding of the relationships within the Psychotrieae complex. Tribe Psychotrieae is a monophyletic group containing 12 genera including *Psychotria*, but the relationships within the tribe are still unclear. Robbrecht and Manen (2006) proposed to split Psychotrieae into two tribes, but results from Bremer and Eriksson (2009) did not support this division and they suggest the tribe to be maintained as one. Within Psychotrieae, the genus *Psychotria* has been shown to be highly paraphyletic (Nepokroeff et al. 1999, Bremer and Manen 2000, Andersson 2002) with respect of other genera of the tribe. Specifically, subgenus

*Heteropsychotria* is paraphyletic with some species more closely related to *Palicourea* species than other *Psychotria* species; and the clade *Psychotria sensu stricto* is paraphyletic as species formerly assigned to subtribe Hydnophytinae are nested within this clade. In addition, the resolution and support values for the internal branches, particularly among the New World species, are usually low or not provided (Nepokroeff et al. 1999, Andersson 2002, Paul et al. 2009).

A major challenge in resolving the genus-wide phylogeny is the paucity of diagnostic characters for many groups within the family (Nepokroeff et al. 1999). This might be due partly to the species-richness of this group, and probably also high intra-specific variation for many traits in terms of both morphology and DNA sequence polymorphism. A lot of morphological characters have been explored to compensate for the lack of variation in floral morphologies within the Psychotrieae complex, including the anatomy of their pollen grains (Johansson 1993), stipules(Andersson 2002), petioles (Martínez-Cabrera et al. 2009), as well as a number of leaf micro-morphological characters such as domatia, vascular tissue arrangement of the petiole, epidermal characters, and presence/absence of styloid crystals (Moraes et al. 2011). However, they were either shown to in part contradict floral characters, or only useful for delimitation of inter-generic relationships instead of differentiating between species. At the molecular level, presence/absence and/or types of alkaloids had showed some chemotaxonomic implications (Lopes et al. 2004, Tan et al. 2012), and karyotypes of members in *Psychotria* as well as Rubiaceae were also studied (Correa and Forni-Martins 2004, Correa et al. 2010, Kiehn 2010). However, those characters were also insufficient to discriminate species, though they may help differentiate subgroups within the tribe. Molecular phylogenetic studies have used over 15 loci (Robbrecht and Manen 2006, Davis et al. 2007, Bremer 2009, Paul et al. 2009, Borhidi 2011,

Barrabe et al. 2012, Sedio et al. 2012) to resolve relationships within family Rubiaceae, from plastid genome (atpB-rbcL, ndhF, matK, rbcL, rps16, trn(T)L-F, trnSG, accD-psaI psbA-trnH) and/or from nuclear genome (ETS, ITS, nontranscribed spacer [NTS], pep-C large, pep-V small, Tpi, *nepGS*, *RPB2*). However, most phylogenetic studies on the Psychotrieae complex included analyses of fewer than three loci, with ITS and one or two plastid loci being the most commonly used. On the other hand, the chloroplast genome evolves slowly relative to the plant nuclear genome (Wolfe et al. 1987). With low numbers of plastid markers (and even combining ITS), it may not provide enough phylogenetic signals for the resolution of relationships within a rapidly and recently diversifying group like *Psychotria.* Thus it comes to a point where it is necessary to explore more possibilities in the faster-evolving nuclear genome to resolve additional molecular markers to be used in *Psychotria*.

Furthermore, there has been a paradigm shift in the field of molecular systematics after the realization that species relationships cannot be resolved by single gene phylogenies (Pamilo and Nei 1988, Maddison 1997, Rosenberg and Tao 2008). The process of speciation includes restriction of gene flow between diverging populations and genetic drift/selection acting on ancestral variation within the diverging populations. Even in the absence of interspecific gene flow, single-locus gene histories (including those for non-recombining plastid genomes) may not coincide with an actual species tree. In plants, there are usually three major sources of gene tree discordance: deep coalescence due to incomplete lineage sorting; interspecific gene flow including hybridization, horizontal gene transfer etc.; and misspecification of orthology due to reciprocal loss of duplicated gene copies after duplication events (Maddison 1997, Degnan and Rosenberg 2009). In addition, gene tree incongruence tends to be high when selected taxa have a short divergence time and/or large ancestral populations, due to the increased chance of gene

lineages not completely sorted after speciation, and/or increased chance of horizontal genetic exchange. Therefore, a better way to accommodate these factors when estimating species-level relationships is to sample multiple loci across the genome, taking into account variations among gene histories of different loci, rather than having only one or two loci to represent species relationships. Coalescent theory, which was originally developed to model population genealogies, has been adapted for phylogenetic/phylogenomic investigations (Rannala and Yang 2003) and is becoming widely incorporated into studies of species-level relationships. These methods often use a likelihood or Bayesian framework to estimate the species tree based on the probability distributions of gene trees under the coalescent model, assuming constant effective population sizes ($N_e$), random mating, and absence of selection and inter-specific gene flow. Some methods take the full dataset with parameter-rich algorithms under a Bayesian framework. E.g. BEST (Liu 2008) implements a Bayesian hierarchical model to jointly estimate gene trees and the species tree from multi-locus sequence data; a similar program *BEAST (Heled and Drummond 2010) differs from BEST in that it co-estimates the species tree and all gene trees in one Bayesian MCMC analysis, utilizing information from all gene trees simultaneously. Other methods use summary statistics, including (but not limited to) GLASS, STEM, STAR/STAEC, and MP-EST etc. The Global LAteSt Split (GLASS) method (Mossel and Roch 2010) and STEM [Species Tree Estimation using Maximum Likelihood (Kubatko et al. 2009)] clusters species using minimum coalescence; while STAR (Species Tree estimation using Average Ranks of coalescences) and STEAC (Species Tree Estimation using Average Coalescence times) both take a distance method [i.e. Neighbor Joining (Saitou and Nei 1987)] in estimating a species tree, using a newly defined summary statistic named ‘average ranks of coalescence’ or average coalescence time to construct distance matrices (Liu et al. 2009b). MP-EST uses a pseudo-

likelihood function developed by Liu et al. (2010) to obtain maximum pseudo-likelihood estimates (MPE) of species trees as an alternative of ML, with branch lengths in coalescent units, as they claimed that the original likelihood function under multi-species coalescent model by Rannala and Yang (2003) does not have a maximum likelihood estimate (Liu et al. 2010).

In this study, a multi-locus phylogenetic analysis was conducted to try to resolve the species-level relationships among a group of New World *Psychotria* species. A number of non-coding sequence markers from both the nuclear and chloroplast genomes were developed and combined with sequences from the ITS and *trnL-trnF* regions. Single gene phylogenies, concatenated phylogeny for the chloroplast data, as well as coalescent analyses for all loci were performed to assess among-locus variation in estimated species relationships and to infer the evolutionary history of the sampled group of species. Our data suggested that that rapid and recent speciation events have resulted in incomplete lineage sorting among closely related New World *Psychotria* populations.

## 3.2 METHODS

*Taxa Sampling & DNA extraction*

A total of 19 species, including 16 *Psychotria* species and three related species within Rubiaceae (Table 2.1), were sampled from different sources. Leaf tissue of 14 *Psychotria* spp., *Palicourea padifolia*, and *Rudgea skutchii* were collected and preserved in silica gel from La Selva Biological Station and Las Cruces Biological Station in Costa Rica during a fieldtrip in summer 2010. Leaf tissue of two other *Psychotria* (*P. marginata and P. horizontalis*) and *Coffea arabica* was collected from UGA Botany Greenhouses. For the Costa Rican samples, I tried to obtain multiple individuals for each species for the phylogenetic analysis, but due to the

limitation of time and sites visited, the number of individuals per species ranged from one to five, with most of the species represented by only one or two individuals (Table 2.1). Genomic DNA was extracted from each individual using a QIAGEN DNA extraction kit or a CTAB protocol modified from Doyle (1987).

*Development of non-coding nuclear and plastid markers*

Sequence markers in both nuclear and chloroplast non-coding regions were developed because these regions are thought to evolve in a largely neutral fashion (little or no selective constrains) so substitutions and indels can accumulate more rapidly than in coding regions (Palmer 1986, Wolfe et al. 1987). Intron-spanning nuclear markers were developed based on available *Psychotria* and *Coffea arabica* EST sequences. *Psychotria marginata* ESTs were generated previously in our lab and *P. ipecacuanha* ESTs have been generated by Toni Kuchan (Danforth Center, St. Louis, MO) and through the oneKp plant transcriptome sequencing project (www.onekp.com). The *Coffea arabica* sequences were retrieved from NCBI. Unigenes from these sources were sorted into gene families circumscribed in the PlantTribes database (Wall et al. 2008), and low-copy genes were identified. Intron positions were predicted for low copy genes with putative orthologs sampled from both *Psychotria* and *Coffea* using the SGN Intron Finder (Mueller et al. 2005). Exon-based primers for intron-spanning loci were designed from alignments of *Psychotria* and *Coffea arabica* homologs. The Prima-clade Primer Visualization (Gadberry et al. 2005) and Primer3 (Rozen and Skaletsky 2000) were used to identify conserved primer sites with favorable binding characteristics (i.e. no self-priming, no hairpins, and intermediate GC content). In addition, COS primers designed for asterids (Chapman et al. 2007) were tested, and the commonly used ITS region were sequenced for all sampled species.

Chloroplast markers were developed through comparison between the *Psychotria marginata* and *C. arabica* plastid genomes using a strategy that is similar to the nuclear marker development described above. The *C. arabica* plastome sequenced by Samson et al. (2007) was used as a reference. The chloroplast genome of *P. marginata* was sequenced using next-generation, short-read (70 bp) sequencing technology (Illumina). The short reads were assembled using both reference-based [Yasra (Ratan 2009), using *C. arabica* as reference genome] and *de novo* [Velvet (Zerbino and Birney 2008)] algorisms and merged into SEQUENCHER v4.2 (GeneCodes, Ann Arbor, MI, USA). The assembled contigs for the *P. marginata* plastid genome were aligned to the *C. arabica* plastome using the alignment and visualization tool zPicture (Ovcharenko et al. 2004). The most variable non-coding regions were identified and primers were designed in flanking conserved regions.

*PCR Amplification and Marker Sequencing*

PCR was performed for each primer set in a 25ul reaction mixture with 2 μl of 10× PCR buffer, 25 mM $MgCl_2$, 30 - 100 ng genomic DNA, 10 μM of each primer, 10 mM dNTPs, and 1 U *Taq* DNA polymerase (lab-made). The PCR program consisted of an initial denaturing step at 94°C for 5 min, followed by 30 cycles of amplification (94°C for 1 min, 54 – 57.5°C for 30sec, 72°C for 30sec – 45 sec with different primer sets), and a final elongation step at 72°C for 5 min. Amplification products were resolved by electrophoresis in 1.5% agarose gels and visualized by EtBr staining or GelRed™ Nucleic Acid Gel Stain (Biotium).

The markers were tested on all the sample species, and primers-pairs that produced single-band amplicons were sequenced using capillary (Sanger) sequencing with ABI BigDye at the Georgia Genomics Facility at the University of Georgia. For nuclear markers that produced low quality sequences (e.g. noisy signals, multiple peaks at the same position, etc.), molecular

cloning was performed using Invitrogen 's TOPO TA Cloning Kit for Sequencing and the cloned products were re-sequenced. In all, five nuclear loci (NL-103, NL-217, NL-57800, NL-A04 and ITS) and four chloroplast loci (*psbE-petL*, *trnT-psbD*, and *trnK-rps16* and *trnL-trnF*) were selected for further analyses.

*Phylogenetic analyses*

Both nuclear and chloroplast markers were aligned using MUSCLE (Edgar 2004). Maximum Likelihood analyses were performed for each nuclear locus as well as a concatenated plastid 4-loci dataset (*psbE-petL*, *trnT-psbD*, *trnK-rps16* and *trnL-trnF*). Because most of the sequence markers did not amplify in *C. arabica,* it was removed from further analyses. All the gene tree reconstructions were conducted using RAxML (ver. 7.2.8) (Stamatakis 2006) with the general time reversible model (GTR) and a gamma distribution for rate heterogeneity for each dataset, and clade support values were assessed with bootstrap analyses of 500 replicates. *Psychotria macrophylla* was used as an outgroup to generate rooted gene trees according to its position in the Nepokroeff et al. (1999) phylogeny.

*Coalescence Analyses & Species Tree Estimation*

To evaluate variation among gene trees and obtain a robust estimate of a species tree, coalescence analyses were performed using gene trees as input in the program MP-EST (Liu et al. 2010). MP-EST obtains maximum pseudo-likelihood estimates (MPE) of a species tree from a set of gene trees. To run the program, a gene tree file and a control file were created. The gene tree file contained rooted ML best trees from previous RAxML output; the control file contained parameter settings for running MP-EST, indicating the numbers of gene trees (six) and species (19), and particularly a matrix of species – allele association. Like other coalescent-based approaches, MP-EST is run under the assumption that there are no duplications and no inter-

specific gene flow following speciation. Therefore, we manually removed the duplicated sequences within the alignments of NL-103, NL-217, NL-57800 and NL-A04, and re-estimated gene trees for these loci to be used in the coalescence analyses with MP-EST. Statistical support for the species tree topology was assessed using multi-species bootstrap. RAxML had generated 500 bootstrapped gene trees for each locus by resampling the sequence alignments. These bootstrapped trees were then used to construct 500 bootstrapped MP-EST trees by resampling one gene tree per locus per replicate, and a consensus tree was built from those bootstrapped MP-EST trees using Majority-Rule-extension (MRe) in CONSENSE from the PHYLIP package (Felsenstein 2005).

## 3.3 RESULTS

*Gene Trees*

Gene tree estimations from RAxML showed that the six gene trees were generally consistent in topology, but none of them provided full resolution of the inter-specific relationships among sampled *Psychotria* species (Figure 3.1). The overall topologies across all gene trees were also consistent with previous studies on the genus-wide phylogeny (Nepokroeff et al. 1999), with strong bootstrap support for the paraphyly of this genus, as well as two major clades corresponding to subgenus *Psychotria* and subgenus *Heteropsychotria*. However, the branch lengths towards the tips of the trees are generally very short and usually resolved as polytomies for both subgenera. Haplotypes from different individuals of the same species were found to be monophyletic except in two species: *P. pilosa* and *P. allenii*. Despite the poor resolution in the *Heteropsychotria* clade that resulted in slightly different topologies, *P. pilosa* haplotypes were paraphyletic with *P. cyanococca* nested within them in four gene trees (ITS,

chloroplast concatenated, NL-57800 and NL-217); and *P. allenii* alleles were always polyphyletic, with a subset grouped with *Palicourea* and the rest grouped with *P. mortoniana* (Figure 3.1). *Psychotria macrophylla* was placed out of either subgenera for all gene trees and used as the outgroup. All but one gene tree (NL-A04) placed this species as sister to *Rudgea*. This relationship is consistent with the phylogeny by Nepokroeff et al. (1999), where *P. macrophylla* occurred within the same clade as two *Rudgea* species, and sister to the *Heteropsychotria* clade. The alternative topology in gene tree NL-A04 placed *P. macrophylla* basal to subgenus *Heteropsychotria* (Figure 3.1), indicating a closer relationship of *P. macrophylla* to the *Heteropsychotria* group than to *Rudgea.* Both topologies had moderate to strong bootstrap support values.

Moreover, nuclear gene trees showed evidence of gene duplications. Haplotypes from *P. calidicola*, *P. chiapensis*, *P. mortoniana*, *P. allennii* and *Palicourea padifolia, P. marginata, P. graciliflora, Rudgea skutchii* were split into two orthologous groups in some of the nuclear gene trees, especially NL-A04 (Figure 3.2), where *P. calidicola*, *P. chiapensis*, *P. mortoniana*, *P. allennii* and *Palicourea* haplotypes were placed into two clades (shown in boxes) in subg. *Heteropsychotria;* and *P. marginata* and *P. graciliflora* haplotypes were also grouped into two clades in subg. *Psychotria*. Similar but not-so-extreme cases occurred in three other gene trees – NL-57800, NL-217 and NL-103 (data not shown), a clear pattern of gene duplication events.

*Coalescent Species Tree Estimation*

Due to the duplications observed in some of the gene trees, we manually removed paralogous copies of the duplicated sequences for each of the four nuclear loci and re-estimated gene trees in RAxML with the reduced alignments. The output bootstrap trees, together with those of ITS and concatenated chloroplast data, were used as input for a multi-species bootstrap

analysis in MP-EST. The resulted majority-rule consensus species tree (Figure 3.3) showed an overall topology that was consistent with all gene trees. Both the Subg. *Psychotria* clade and the Subg. *Heteropsychotria* clade had very strong bootstrap support for their monophyly (500/500 respectively). The subg. *Heteropsychotria* clade still had mainly very short branches towards the tips, and the bootstrap support values for the internal nodes were generally low, indicating rapid yet relatively recent species diversification. *Psychotria macrophylla* is more closely related to *Rudgea* on the species tree, and the two species were not grouped within either subgenus *Psychotria* or *Heteropsychotria* with about 70% bootstrap support（347/500). As to the non-monophyletic species, *P. allenii* was sister to *Palicourea*, and *P. pilosa* and *P. cyanococca* formed a monophyletic group. Neither of these groupings had strong bootstrap support (256/500 and 222/500 respectively), suggesting lack of phylogenetic signal or conflicting signal. Compared to the earlier phylogeny by Nepokroeff et al. (1999), the placement of *P. macrophylla* and *Rudgea* is consistent with what they reported, and the relationships within subgenus *Psychotria* were much more robust in the MP-EST analysis. *Psychotria graciliflora, P. chagrensis, P. marginata* and *P. horizontalis* were among members of a unsupported clade of Subg. *Psychotria,* with zero bootstrap values for their internal relationships in the tree by Nepokroeff et al. (1999). On the other hand, MP-EST analysis showed that *P. graciliflora* and *P. chagrensis* are more closely related, forming a clade that is sister to *P. marginata* and then to *P. horizontalis*. These groupings were supported by 75.4% (377/500), 64.4%(322/500) and 72% (360/500) bootstrap values respectively, a very significant improvement of statistical support for the internal relationships within the Subgenus *Psychotria* relative to previous studies.

## 3.4 Discussion

*Phylogenetic Relationships*

Both the nuclear gene trees and the concatenated chloroplast phylogeny supported the paraphyly of the *Psychotria* species sampled in this study. This is consistent with previous studies based on a more comprehensive sampling of the genus (Nepokroeff et al. 1999, Paul et al. 2009). Specifically, *Rudgea* and *Palicourea* are nested within *Psychotria*, and *Palicourea* is nested within the *Heteropsychotria* clade, showing that members of *Psychotria* in this clade are more closely related to *Palicourea* species. Both morphological (Taylor 1996) and molecular studies (Andersson and Rova 1999, Nepokroeff et al. 1999, Andersson 2002) have shown that *Palicourea* is polyphyletic and derived from within Subg. *Heteropsychotria* clade. Its close relationship with *Psychotria* as well as some other genera within the tribe makes the generic circumscription of these groups plastic, and until recently, members of these taxa are still being placed into different groups according to different molecular studies (Borhidi 2011, Barrabe et al. 2012). Much more work is still needed to provide better understanding of the morphological and molecular identities in order to have a more meaningful delimitation of these groups. Although the tips of gene trees were still not fully resolved in our analyses, our results showed improved bootstrap support for many internodes at the sub-generic level compared to the phylogeny by Nepokroeff et al. (1999). The Nepokroeff phylogeny was estimated using nuclear ITS and chloroplast *rbcL* gene. Relationships were not well resolved; especially among New World species within Subg. *Psychotria*, with zero bootstrap for the relationships among *P. marginata*, *P. horizontalis*, *P. graciliflora*, *P. micrantha*, *P. brasiliensis*, *P. borjensis*, and *P. mapourioides*. Our analyses resolved the subgenus *Psychotria* clade with internal nodes supported with bootstrap support between 40 and 100. Relationships among species at the tips of our trees,

however, are poorly supported. Our results suggest a higher-level polymorphism in the non-coding markers used in this study that can provide more phylogenetically informative characters. In addition, combining five nuclear loci and four chloroplast loci, the MP-EST coalescent analysis resulted in a reasonably supported phylogeny for the species sampled in this study, showing that coalescent analysis can be helpful to accommodate information from different loci and provide a better way of species-level phylogeny reconstructions within lower taxonomic levels. As it is becoming cheaper to perform deep sequencing of plant genomes, combining gene capture and next-gen sequencing technology will be a very efficient way to identify hundreds and even thousands of variable non-coding orthologous markers at the same time across the genome. Therefore multi-species coalescent analysis could be a promising approach to be adapted widely into studies of shallower phylogenies.

*Rapid speciation in Psychotria*

Results from the multi-locus phylogenetic analyses clearly suggested that the New World *Psychotria* species experienced a relatively recent and rapid radiation, as indicated by the very short branches in subgenus *Heteropsychotria* observed in both the gene trees and MP-EST tree. Around 750 species of *Psychotria* are estimated to occur in the Neotropics, and we were able to see patterns of rapid diversification of this group with a limited sample size. This in turn supported that the New World *Psychotria* species have diverged relatively recently and rapidly, with extant species closely related with each other. One of the significant patterns resulting from recent speciation in large populations is incomplete lineage sorting/deep coalescence, where individual gene lineages split earlier than the species split while ancestral polymorphisms were kept into the daughter lineages. We observed that ILS is relatively common among these closely related species of *Psychotria*, with *P. pilosa* haplotypes showing inconsistent grouping patterns

among different gene trees. Those patterns indicate that the ancestral sequence polymorphism may be still maintained in some parts of the genome, and again highlights the usefulness of coalescent-based analysis, which specifically handles potential gene tree discordance caused by ILS.

Polyphyletic haplotypes of *P. allenii* across all the gene trees may be indicative of cryptic species. As the two sympatric individuals had extremely similar morphology, we identified them as the same species. However, our studies showed two distinct origins of these two individuals, suggesting that they do not share a common ancestor at the species level. The tropics are the most species-rich habitats on earth (Willig et al. 2003), and the sympatric co-existence of many species allows inter-specific interactions as well as parallel evolution of similar traits. Therefore it is possible that the tropics are more likely to harbor cryptic species.

Patterns of gene duplications were also observed for some of the species among the nuclear loci. Haplotypes from *P. calidicola*, *P. chiapensis*, *P. mortoniana*, *P. allennii* and *Palicourea padifolia, P. marginata, P. graciliflora, Rudgea skutchii* were split into two orthologous groups at some of the nuclear gene trees, especially NL-A04. Particularly, *P. allenii*, *P. pilosa* as well as some of their sister groups in the *Heteropsychotria* clade are among those species with gene duplications. These patterns probably correlate with the varying ploidy levels across species of both Subgenera *Psychotria* and *Heteropsychotria* reported by previous studies. Correa et al. (2010), together with an earlier work by the same group (Correa and Forni-Martins 2004) showed that members of *Psychotria* have a diversity of chromosome numbers ranging from n = 11, 12, 17, 22 to 34; and that there lacks a consistent karyotype pattern in both subgenera of this genus. A more comprehensive summary of neotropical Rubiodeae by Kiehn (2010) further showed that the frequency of polyploidy seems to be higher in Subg.

*Heteropsychotria* than in Subg. *Psychotria,* with of half of *Heteropsychotria* taxa (31 total) are tetraploid or higher compared to only two out of 10 *Psychotria* taxa that are tetraploid. Although most species included in our study were not counted in their studies, *Palicourea padifolia,* the only species from genus *Palicourea* used in our study, was confirmed to be tetraploid by Kiehn (2010). It is probable that the other species showing duplicating patterns (*Psychotria calidicola*, *P. chiapensis*, *P. mortoniana*, *P. allennii* ) in the *Heteropsychotria* clade are also polyploid given the relatively frequent occurrence of polyploidy species in this group, though further studies are needed to confirm their karyotype and ploidy level. The observed variation in chromosome numbers within members of *Psychotria* may reflect independent gene/genome duplication and loss events occurred in this group, and it would be interesting to test the relative intensity of selection versus drift that have contributed to these duplication/loss events in future studies.

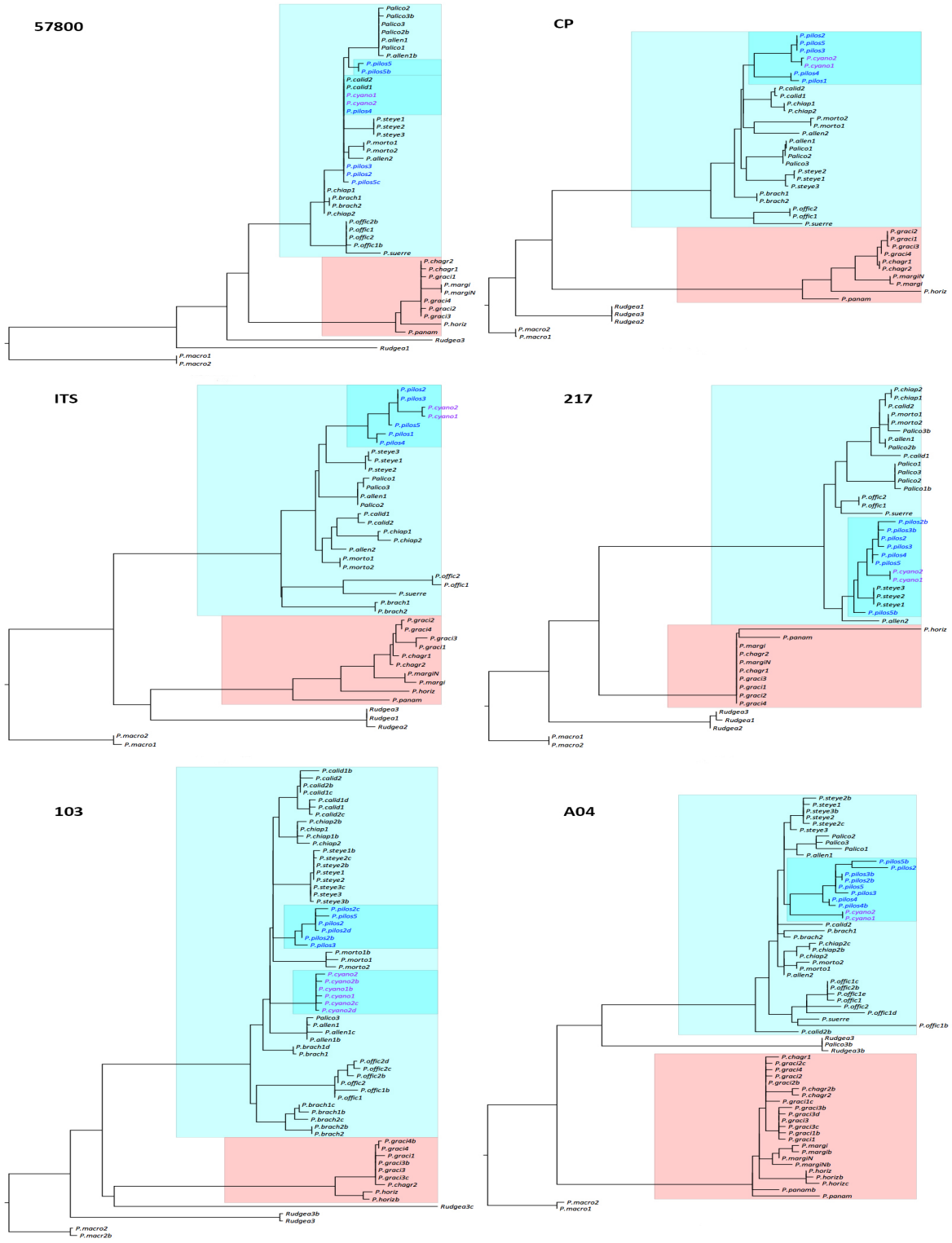


**Fig 3.1** ML gene trees from the modified dataset (after removal of duplicated copies of nuclear genes). Each locus is labeled at the top left corner of the tree. Blue and red highlights indicate the *Heteropsychotria* clade and *Psychotria* clade, respectively. *Psychotria pilosa* and *P. cyanococca* haplotypes are labeled blue and purple to show their relative positions on the tree.

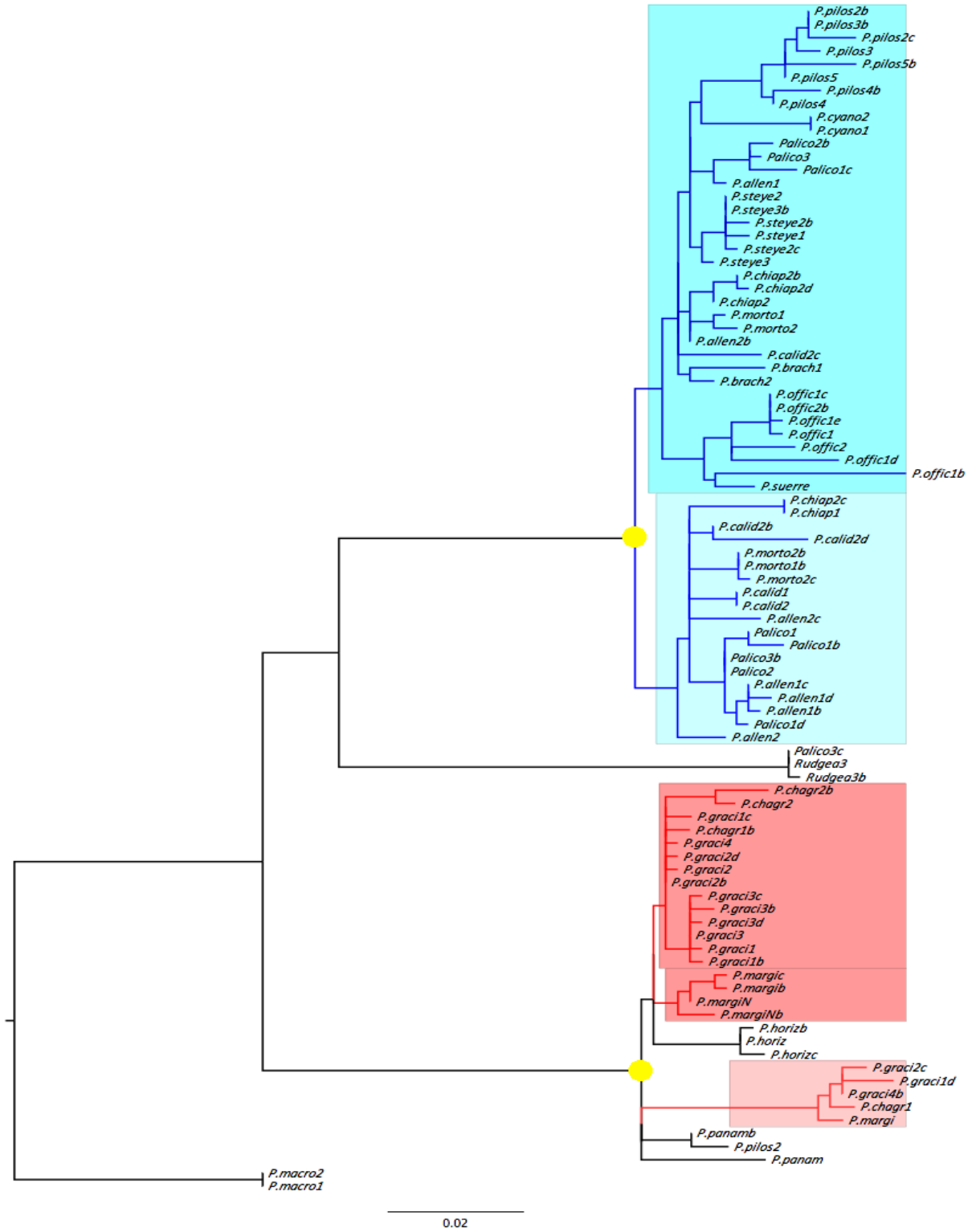


**Figure 3.2** An example of a nuclear gene tree (NL-A04, full data) showing gene duplications. Yellow dots denote the inferred node of gene duplication events. Blue and red highlights indicate the *Heteropsychotria* clade and *Psychotria* clade, with lighter highlighted regions indicating the duplicated clades within both subgenera.

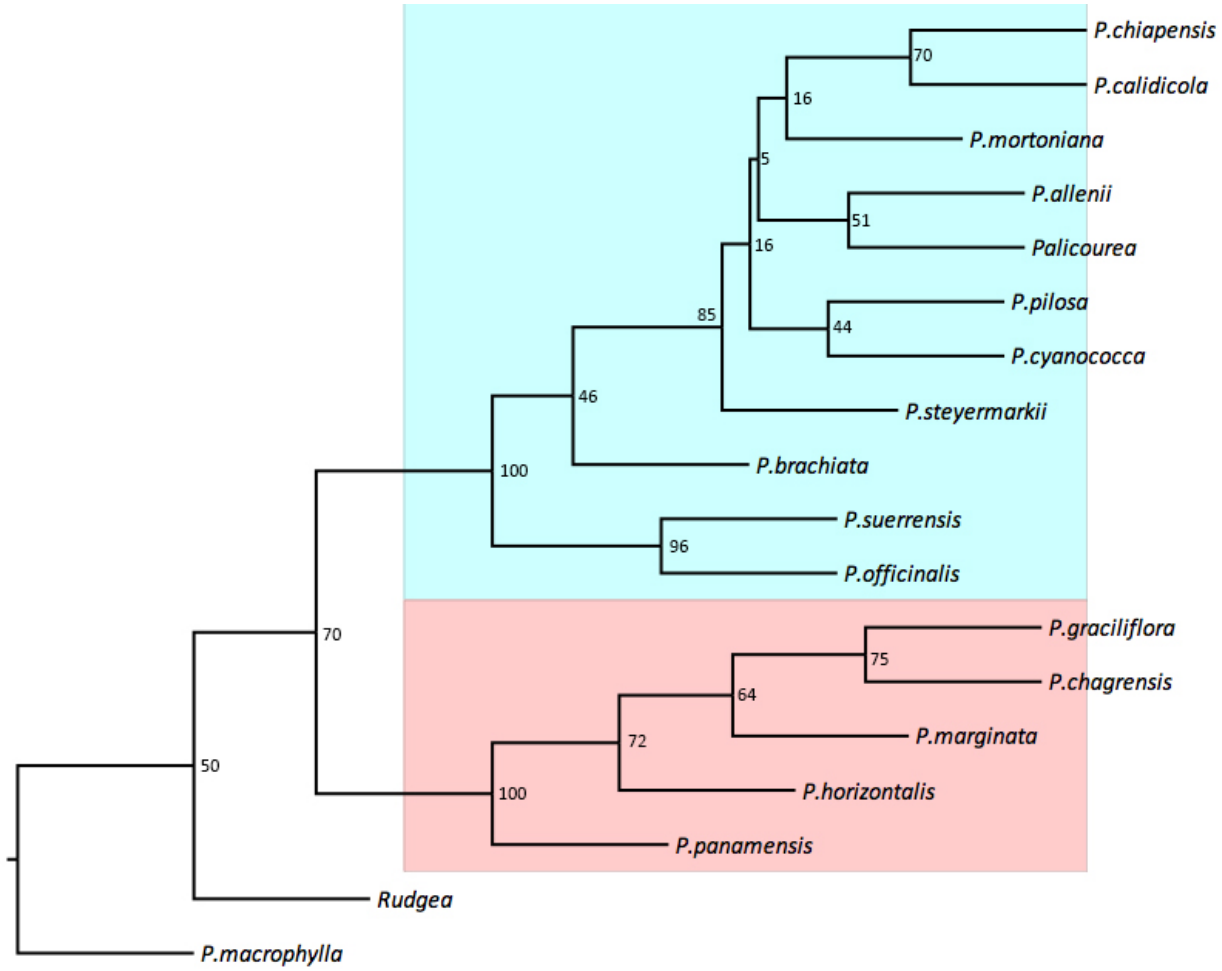


**Figure 3.3** The MP-EST consensus tree for 18 species. A multi-species bootstrap analysis by the MP-EST method was performed for 500 bootstrap datasets. The consensus tree was then constructed using Majority-Rule-extension (MRe) in the program CONSENSUS from the PHYLIP package. The numbers at each node are bootstrap values based on 500 replicates. Blue and red highlights show the *Heteropsychotria* clade and *Psychotria* clade, respectively.

# CHAPTER 4

# PARAPHYLETIC SPECIES DELIMITATION AND SPECIATION – A SIMULATION STUDY USING COALESCENT APPROACHES [3]

[3] DONG, L. AND J.H. LEEBENS-MACK. To be submitted to *Molecular Biology and Evolution.*

## ABSTRACT

Current coalescent models for species delimitation and species relationship inferences only deal with simple allopatric speciation events. However, sympatric speciation can be a much more common evolutionary process in the tropical forests, where species density is greatest. To examine the performance of coalescent-based approaches in the case of sympatric speciation versus allopatric speciation, coalescent simulations were performed in this study to assess whether incomplete lineage sorting by random drift alone can lead to paraphyletic species on gene trees under a regionally sympatric model with structured populations before speciation, as what we have observed in the neo-tropical species of *Psychotria* (Rubiaceae). In addition, we also examined the performance of species delimitation using the simulated data to see how the current coalescent methods handle sympatric versus allopatric speciation. The simulated data suggest that random sorting of alleles after speciation event alone could cause paraphyletic haplotypes on gene trees under the regionally sympatric model, and make species delimitation difficult, even under very small rate of gene flow between ancestral populations with large Ne. The incorrect species delimitation results for the reduced *Psychotria* dataset also suggest that incomplete lineage sorting after the initial divergence of *P. cyanococca* from *P. pilosa* alone can result in haplotypes from *P. pilosa* populations forming paraphyletic groups in some of the gene trees, although we cannot rule out the possibility of ongoing gene flow after speciation.


## 4.1 INTRODUCTION

Speciation is a fundamental source of biodiversity. In the field of biology, species are important for both taxonomists/systematists who study ‘characters and patterns’, and ecologists/population geneticists/evolutionary biologists who study ‘processes and

mechanisms'(Harrison 1998). The process of new species formation, i.e. speciation, is a significant outcome of evolution and the increase of biodiversity. Therefore, 'correctly' defining species circumscriptions is critical, both for the accurate characterization of biodiversity patterns and the study of evolutionary processes that generate biodiversity. However, researchers working in different areas tend to invoke different species concepts with species delimitation criteria that fit their own purposes, resulting in seemingly endless debates over how species should be defined (de Queiroz 2007, Hausdorf 2011). de Queiroz (1998) proposed a General Lineage Concept (GLC) that treats species as 'separately evolving meta-population lineages', emphasizing the distinction of the speciation process from secondary diagnostic biological attributes of organisms that can be used as criteria for quantitative empirical tests of hypothesized species delimitations (de Queiroz 2007). This conceptualization made an important clarification that species exist as evolutionary lineages regardless of any diagnostic attributes that might be used empirically delimiting them (Camargo and Sites 2013). However, people are still arguing about some issues with the GLC, including the notion that speciation can occur within structured populations in the face of ongoing gene flow; and challenges of species diagnosis, especially in the face of parallel or convergent evolution (Nosil 2008, Hausdorf 2011). In short, speciation and species delimitation are distinct issues. Species delimitation is essentially an empirical endeavor aiming to define diagnostic properties of distinct species, whereas speciation processes such as population divergence are typical continuous (Hey 2001). Despite the fact that species delimitation may be extremely challenging given a continuous (rather than abrupt) speciation process, species definitions are have important implications for systematics, ecology, evolutionary biology and conservation biology.

Species delimitation has become a very active area of research, particularly within the last five years with a sharp increase in publications (Camargo and Sites 2013). The increased interest in this topic is coincident with the incorporation of coalescent theory into phylogenetics and a paradigm shift in molecular systematics that has lead to greater appreciation of the distinction between gene trees and species trees (Edwards 2009). Similar to coalescent-based species tree reconstruction methods, coalescent-based species delimitation approaches do not require reciprocal monophyly of alleles within gene trees. Multi-locus data sets are often used to test alternative hypotheses of lineage divergence that allow for gene tree discordance as a result of random lineage sorting. The program BP&P 2.1 (Yang and Rannala 2010) estimates the posterior probability distribution of species delimitation models under a Bayesian reversible-jump MCMC framework, using sequence alignments and a guide tree as input. Another program SpedeSTEM (Ence and Carstens 2011) generates maximum-likelihood species trees of alternate species delimitation models and selects the best model using AIC. The program Brownie (O'Meara 2010) simultaneously estimates species delimitations and the optimum species tree by minimizing gene tree conflicts and intra-specific structures using heuristic searches. These model-based testing methods utilizing universal DNA sequence data have the advantage of good repeatability as well as easily avoiding investigator bias that may accompany identification of diagnostic morphological or ecological characters, though the actual performance of each approach varies under different conditions (Fujita et al. 2012).

Fujita et al. (2012) asserted that delimiting species among sympatric entities is generally non-controversial because reproductive isolation is often readily inferable based on morphological, behavioral, or ecological evidence, whereas delimiting allopatric species is more challenging since neutral divergence between allopatric populations may be gradual. Sympatric

speciation, however, may be more rapid and therefore more difficult to track by gene tree-based species delimitation approaches. In phylogenetics, allopatric speciation by subdivision is more likely to generate reciprocally monophyletic daughter species due to complete isolation, while locally occurring processes such as sympatric/parapatric speciation and founder effects are more likely to result in a paraphyletic progenitor and a monophyletic new species derived from it, as there is not an absolute barrier to inter-population gene flow at the time of species divergence. Reciprocal monophyly can require long periods of population divergence in the absence of gene flow. In plants, classical allopatric speciation may be less frequent than geographically local speciation events, as plant populations are generally continuous and widespread (Rieseberg and Brouillet 1994). In fact, studies have revealed a high incidence of paraphyletic species in angiosperm families such as Fabaceae, Aizoaceae, Asteraceae and Orchidaceae (Rieseberg et al. 1990, Freudenstein and Doyle 1994, Rieseberg and Brouillet 1994, Chandler and Crisp 1998, Thulin et al. 2012).

Our recent phylogenetic analyses on a group of Costa Rican *Psychotria* species (family Rubiaceae) also demonstrated a paraphyletic pattern of relationships with *P. cyanococca* haplotypes nested within a grade of *P. pilosa* haplotypes in some gene trees (Dong and Leebens-Mack, unpublished data). *Psychotria* is the largest genus in Rubiaceae and has a pan-tropical distribution. Members of this genus are mostly understory shrubs and are important components of tropical understory vegetation (Nepokroeff et al. 1999). *Psychotria pilosa* and *P. cyanococca* are both widely distributed in Central America, but the range of *P. pilosa* extends southward to Peru and Bolivia (Tropicos database, www.tropicos.org). Samples of *P. pilosa* came from two populations, one in the lowland tropical rainforest at La Selva Biological Station, and the other in the forest of about 1200m in elevation at Las Cruces Biological Station in Costa Rica. *P.*

*cyanococca* occurs sympatrically with *P. pilosa* in La Selva, but does not occur at Las Cruces. The two species are also morphologically distinct – *P. pilosa* is heavily pubescent with darker green, thick leaves; whereas *P. cyanococca* is glabrous with lighter green, thinner leaves. Results from our phylogenetic analyses showed that *P. pilosa* haplotypes were paraphyletic with *P. cyanococca* nested within them in gene trees based on ITS, two nuclear intron loci (NL-57800 and NL-217) as well as a concatenated 4-loci chloroplast sequence matrix, but were monophyletic as a sister group to *P. cyanococca* in two other nuclear gene trees.

The purpose of this study is to use coalescent simulations to model a simple speciation process similar to what we observed in the *Psychotria* data, with gene flow between the ancestral populations, and one population diverged from the others as a new species. Speciation was defined to occur when gene flow rate permanently dropped to zero. Our goal is to ascertain whether the coalescent processes alone, i.e. only neutral stochasticity without any selective forces, could affect species delimitation based solely on gene trees generated from the paraphyletic progenitor and the derivative species under a regionally sympatric model. If so, how would effective population size, migration rate and depth of split (divergence time) affect the probability of correctly recognizing the two species? Coalescent processes of a simple allopatric species split is also simulated as a comparison, and a reduced set of *Psychotria* empirical data was used to perform a species delimitation test based on the gene trees. Results from this study may provide insights into species delimitation using sequence-based gene trees，and an understanding of how coalescent-based species delimitation approaches implicitly model sympatric and allopatric speciation processes. This has implications for population/species sampling and interpretation of paraphyletic gene trees.

## 4.2 METHODS

*Simulation Models*

We modeled coalescent processes of a regionally sympatric speciation process (or 'geographically local speciation events' sensu lato (Rieseberg and Brouillet 1994) with gene flow between three ancestral populations of two species (Figure 1), and compared the efficacy of species delimitation given this process relative to an allopatric speciation process.

In the regional sympatric speciation model, the ancestral population A, B and C were of the same effective population size (Ne), and A & B were geographically closer than to C. The migration rates (*m*) were set to be symmetric between populations (i.e. $m_{ab}=m_{ba}$, $m_{ac}=m_{ca}$, $m_{bc}=m_{cb}$), with *m* between A and B twice as high as that between A and C as well as between B and C ($m_{ab}=m_{ba}=2m_{ac}=2m_{ca}=2m_{bc}=2m_{cb}$). At a specified time point ($t=t_0$), population B was set to form new species, so that gene flow ceased between B and the other two populations ($m_{ab}=m_{ba}= m_{bc}=m_{cb}=0$). The effective population size for each population was kept constant throughout the process. This model describes a sympatric speciation event occurring among three ancestral populations, resulting in a new species diverging from an existent ancestral species. The allopatric speciation model, simulated an ancestral population that was split into two daughter populations, A and B, by some type of geographic barrier. Population A and B have no gene flow since the time of split and have thus evolved independently into two new species.

*Coalescence Simulation*

In all simulations, five individuals were sampled from each population. For the regionally sympatric, 3-population model, parameters varied were the effective population size Ne, migration rate *m,* and depth of split (*t*, in generations). We also varied the number of randomly

sampled loci to test if more intensive sampling can help improve the result. We set three levels of Ne: 5000, 10000 and 100000. Under each Ne, all the combinations of four different *m*'s (1e-5, 1e-4, 0.001, 0.01) and three different *t*'s (1000, 5000 and 10000 generations) were simulated, except that for Ne=100000, m=0.01 were replaced by *m*=1e-6, and another *t*=50000 generation were added in the simulation. Ten replicates were simulated for each parameter combination, and gene trees were simulated with 5, 10, 50 and 100 loci, respectively. For the allopatric, 2-population model, all parameter settings were the same except that *m* is missing (*m*=0 for all conditions). All sets of gene trees were simulated using the program *msms* (Ewing and Hermisson 2010).

*Psychotria Data*

Sequences of *Psychotria pilosa* and *P. cyanococca* were extracted from the mega dataset together with those of *P. macrophylla* as an outgroup for all the loci (nuclear NL-103, NL-217, NL-57800, NL-A04, ITS as well as chloroplast *psbE-petL, trnK-rps16, trnT-psbD* and *trnL-trnF*). As Brownie requires each input gene tree with the same number of individuals for each species, we removed the redundant haplotypes for *P. cyanococca* and *P. pilosa* for some of the loci and kept the number of sequences for each species constant to reflect the original number of individuals collected: two for *P. cyanococca*, five for *P. pilosa* and two for *P. macrophylla*. Sequence for the six loci (the four chloroplast loci were concatenated into a single matrix) were aligned using MUSCLE (Edgar 2004), and the gene trees were reconstructed in RAxML (Stamatakis 2006).

*Species Delimitation*

The program Brownie (O'Meara et al. 2006, O'Meara 2010) was used to perform the species delimitation analyses on the simulated gene trees. As described in the introduction, this

program takes a heuristic search strategy to jointly perform species delimitation and species relationship inference. Although BP&P 2.1 showed reasonable performances alone or compared to other coalescent-based methods such as SpedeSTEM in a few empirical tests (Yang and Rannala 2010, Setiadi et al. 2011, Camargo et al. 2012, Niemiller et al. 2012), the guide tree seems to play an important role on affecting the accuracy of results, because mis-specification on the guide tree tend to result in overs-splitting the species (Setiadi et al. 2011). In a study of Southern Cavefish by Niemiller et al. (2012), many species groups resulted from Brownie were also supported by BP&P, though the number of loci and individuals sampled can influence the number of species delimited in Brownie. Therefore Brownie was chosen over BP&P 2.1 and SpedeSTEM, because it does not require any prior assignments of group membership or guide tree, which could hopefully reduce some bias introduced by those priors. For our datasets, the number of species inferred as well as the topologies of delimited species trees were recorded and compared for both the *Psychotria* data and simulated data. Particularly, ANOVA was performed in R (Version 2.15.1) in order to assess the effect of Ne, *m*, *t*, number of loci, and potential interactions among the four factors, on the accuracy of species delimitation results for the simulated dataset. According to the simulation model, the number of species expected should be two as a result of the speciation event for both the sympatric and allopatric models. Particularly, the output tree from Brownie is predicted to have a topology of (B, (A + C)) for the sympatric model, where all the individuals of B form a monophyletic clade while a mix of individuals of A and C form the other clade, supporting population B as a distinct species. Going backwards in time, the sampled lineages will take longer to coalesce, as Ne and *m* get larger. So we also expect that increasing *Ne*, *m* and decreasing *t* would make it more difficult to differentiating between species and to obtain the expected species delimitation from Brownie. On the other hand, under a

specific combination of Ne, *m* and *t*, increasing the number of loci sampled should increase the amount of informative signals contained in the gene trees, and thus help in getting the expected species delimitation.

## 4.3 RESULTS

*The Regionally Sympatric Model*

Figure 4.2 summarizes all types of delimited species trees from the simulated sympatric model, represented by trees of sampled individuals from each population. Each polytomy on the tree represents a species group recognized by Brownie. Under different scenarios of Ne, *m* and *t*, the number of species inferred varied between one (under-split) and three (over-split). In general, species delimitation under this 3-population model with gene flow between ancestral populations is problematic. In the majority of cases, Brownie was not able to recover the expected species delimitation, in terms of both the number of species recognized and the placement of different haplotypes into the right species group. Brownie did resolve the species number of two under some parameter combinations, but the internal relationships between individual haplotypes were usually chaotic. A few additional incorrect topologies, where the wrong type and/or number of individuals formed the new species clade, were quite often recovered. For the 10 replicates performed for each parameter combination, the proportion of correct species delimitation was usually 1/10 to 3/10, with a maximum of 5/10. In addition, the expected species delimitation only showed up in small populations (Ne = 5000). In the case of oversplit, the 3 populations were resolved as 3 different species, with varying relationships between haplotypes. Individuals from A, B and C were either correctly assigned to their own clade with different relationships [(B, (A, C)) or (C, (A, B))], or incorrectly put into clades of other populations, depending on the

parameter combinations (Figure 4.2). And finally, Brownie cannot differentiate between any populations in several instances and placed all individuals into a single species group.

The comparisons between different parameter combinations in terms of the average number of species inferred from 10 replicates for each parameter setting are shown in Figure 4.3. Generally as we expected, recovering just the correct species number became more difficult as the effective population size (Ne) and gene flow rate (*m*) increased and the depth of split (*t*) decreased. Particularly, Ne likely contributes more than the other two parameters to this pattern of change. When Ne was small, the correct number of species (N) could be easily recovered even when depth of split (*t*) was relatively small and migration rate *m* was relatively large; whereas when Ne was large, the correct species number was not obtained even under very small *m* and large *t*. For example, when Ne=5000, Brownie was able to return two species, occasionally with correct topology at *m* = 0.01 or *t* = 1000 gen; however when Ne = 100000, the correct species number was not returned until *m* = 1e-6 or *t* = 50000 gen. On the other hand, when the level of migration (*m*) is low, and/or depth of split (*t*) is large, oversplitting species tend to occur (Figure 4.3) but often with incorrect relationships. Interestingly, individuals from population B (the new species) were always correctly recovered as a clade in the oversplit cases, as long as *t* was long enough and the number of loci sampled was relatively large, indicating that Brownie was able to figure out that population B is the one that had changed.

The number of loci sampled also affects the inference of both species number and tree topology under the same parameter settings. The number of species returned by Brownie tends to increase under the same parameter settings as the number of loci sampled increases from five to 100. Specifically, in cases of undersplit with fewer loci, increasing the number of loci could increase the chance of obtaining the correct species number; whereas in cases of correct species

number recovery with fewer loci, increasing the sampling tends to result in oversplit. This trend is apparent with any combination of Ne, *m* and *t* (Figure 4.3), but seems to be masked to some extent by Ne. In smaller populations, this sampling effect was more obvious than in larger populations. For example, when Ne = 5000, the average N started to increase immediately from sampling five loci to 10, and continues for 50 and 100. However, when Ne = 100000, increasing the number of loci did not show an immediate effect from five to 10 loci, until the sampling increased from 50 loci to 100. These results suggest that increasing the number of loci does increase the amount of informative signals proportional to the effective population size, but also increases the proportion of complete sorted genes before the species diverge. On the other hand, increasing the sampling of loci also seems to obscure species boundaries and increase the chance of obtaining incorrect species groups with mixed individuals from different populations. This was only observed in the Ne = 5000 datasets, as there were hardly any correct grouping of individuals for every species in the Ne = 10000 or 100000 data given any number of sampled loci. Specifically, the correct species delimitation was recovered for six out of 10 total replicates for *m* = 0.001and *t* = 10000 gen with only five loci; seven out of 10 with 10 loci, but only one out of 10 with 50/100 loci. As expected, sympatric speciation results in greater challenges for coalescence-based species delimitation even with larger numbers of sampled loci.

*Allopatric Model*

Compared to the sympatric speciation model with structured ancestral populations, species delimitation for a single allopatric speciation event seems to be much easier. As with the regionally sympatric model, increasing Ne and/or decreasing *t* also decreases the chance of obtaining the correct species delimitation, and Ne still plays a major role affecting the accuracy of estimation. However, the allopathric species split can be recovered with large enough t or

number of loci even in large populations, including both the species number and the relationships between haplotypes. The average number of species recovered across 10 replicates for each parameter combinations are shown in Figure 4.4. In general, the accuracy of species delimitation is much higher for this 2-population model. Particularly, the frequency of obtaining the correct species relationships when N=2 increased significantly compared to the sympatric model, although there were still a few cases where incorrect relationships between individuals were resolved by Brownie (topologies not shown here). For Ne = 5000 and 10000, this only occurred in a number of replicates at $t$ = 1000 generations or $t$ = 5000 generations when the number of loci sampled was less than 50. For Ne = 100000, this was relatively more frequent across various parameter combinations, but no mis-identified haplotypes were found when t = 50000 and the number of loci >= 10. This suggests that under a simple allopatric speciation model without population structures, the correct species relationships are more readily identified. In addition, increasing the number of loci did not increase the chance of misplacement of haplotypes, but instead helped increase the chance of correct species delimitation by compensating for the inadequate signal caused by shallow divergence. For example, sampling 100 loci resulted in two out of 10 correct species delimitations when the depth of split was only 1000 generations for Ne = 10000 under this allopatric model, but the same parameter setting did not return any correct species delimitations under the sympatric model for any levels of gene flow. When Ne increased to 100000, sampling >=50 loci successfully recovered the species delimitation for all 10 replicates when t=50000 gen for the allopatric simulation, which is again a much better result compared to the sympatric model.

*ANOVA*

Single-factor effects on the average number of species recovered were plotted in Figure 4.5 using pooled data. In the sympatric model, migration rate (*m*) and number of loci seem to be two factors with the strongest effect on the accuracy of species delimitation (in terms of the average number of species recovered), followed by Ne and *t*. In the allopatric model, the number of loci does not have as strong effect as in the sympatric model, but shows a moderate effect similar to *t,* and followed by Ne. Note that the sample size for $m = 0.01$ as well as $t = 50000$ (generations) are much smaller because those values were applied only to part of the data (Ne=100000). So their corresponding values of average species number on the figure were not fully comparable with those under the same parameter of different values. However, there is a clear general trend of each parameter that is consistent with our predictions.

In order to statistically assess the effect of Ne, *m*, *t,* and number of loci, as well as possible interactions among the four factors, ANOVA was performed on both the sympatric and allopatric data, and the results were summarized in Table 4.1. For the 3-population sympatric speciation model, all four factors showed very significant effect ($P<0.001$) on the number of species recovered by Brownie. In addition, interactions between Ne and *t*, Ne and *m*, as well as *t* and number of loci were also very significant ($P < 0.01$). These results also make intuitive sense, because under the coalescent model, migration is proportional to Ne (4Ne*m*); while depth of split is in units of Ne (*t*/4Ne), and the program *msms* takes the transformed coalescent units (4Ne*m* and *t*/4Ne) as input to perform simulations. Therefore, the *t's* of the same absolute value (in generations) would become smaller in coalescent units for larger populations than smaller ones, resulting in shallower divergence. The *m*'s of the same absolute value would become larger in coalescent units for larger populations, producing more gene flow. Interactions between *t* and the

number of loci sampled probably indicate that informative signals from increased number of loci may compensate for lack of information in single genes due to short divergence times. For the 2-population allopatric model, Ne, t and number of loci all had a very significant effect on the number of species recovered ($P < 0.001$); Ne and *t* also showed very significant interactions ($P < 0.001$) as in the sympatric model. However, no significant interactions were detected between *t* and number of loci, while the joint interactions of Ne, t and number of loci was very significant ($P < 0.001$).

Psychotria *3-species Empirical Data*

Six gene trees were reconstructed from the reduced alignments of *P. cyanococca*, *P. pilosa* and *P. macrophylla* (Figure 4.6). *Psychotria macrophylla* was chosen as the outgroup because it is not in the clade of the two focal species, according to previous studies (Nepokroeff et al. 1999) as well as the results from our phylogenetic analyses on the full dataset. *Psychotria pilosa* individuals was paraphyletic in three out of the six gene trees and reciprocally monophyletic with *P. cyanococca* in the other three, with moderate to strong bootstrap support for each gene trees. *Psychotria macrophylla* always formed a monophyletic clade basal to the other two species across all six trees with 100% bootstrap support. However, species delimitation analysis did not result in reciprocally monophyletic relationships for the three species. The number of species recognized by Brownie was two. The two types of topologies returned as the best trees both had *P. macrophylla* grouped with *P. cyanococca*. One of them had both *P. macrophylla* individuals within the *P. cyanococca* clade, and the other had one individual within *P. cyanococca* clade and the second one within *P. pilosa* clade (Figure 4.7).

## 4.4 Discussion

Analyses on the simulation data suggest that species delimitation results varied under different combinations of Ne, *m* and *t*. As expected, large Ne and short divergence time under even low migration rate could make species delimitation difficult in the sympatric 3-population model. Under such parameter combinations, Brownie generated a single polytomy for all 15 individuals sampled, indicating that they came from a single species. This can be explained by the increased chance of incomplete lineage sorting plus more noise in the coalescent processes caused by gene flow between structured ancestral populations. Within the genome of a species, each locus has its own history of descendence. At the time of speciation, copies of different alleles from one locus will be randomly divided into the daughter species, and alleles that are more closely related might not be retained in a single daughter species. In turn, the daughter species might share some of the more closely related alleles between each other rather than within themselves. Therefore, a polyphyletic/paraphyletic pattern in the gene trees for the newly formed species would be expected immediately after speciation. Over time, each species evolves independently, and those ancient alleles may be lost or fixed by random drift or selection, while each species would produce their own new alleles by mutation. Eventually, each species would only retain certain types of ancestral alleles as well as private intraspecific variations, resulting in reciprocal monophyly for this locus. As the time required for achieving monophyly is proportional to Ne, it is more likely that the ancestral polymorphism will be retained at the time of speciation for many more loci and that the time since initial divergence to complete sorting of ancestral alleles will become longer with large Ne. Gene flow between ancestral populations in sympatry before the time of speciation would make the situation worse by blurring the inter-population structures. If the random sampling of alleles occurs a relatively short time after

speciation events, it is likely that individuals from different species will be taken as those from the same species based on sequence and gene tree data. On the other hand, smaller populations require much less time to complete lineage sorting, so recovering the correct species delimitations for smaller Ne is easier, even in the case of sympatric divergence of structured populations. However, since the effect of random drift tends to intensify in small populations, particularly when ancestral populations are highly structured, oversplitting species is likely to occur with small Ne/m and large t combinations. In sum, we observed both failure to differentiate between species (a single species) and oversplitting the populations into three species. The frequency of false negatives (undersplitting) and false positives (overspliting) in species delimitation varied with ratio of Ne, *m* and *t*.

Sampling also affected species delimitation. For all the Ne's, adding more loci increased the chance of obtaining the correct species delimitation, and the gradient pattern from under-splitting to over-splitting was more obviously shown in the 100-loci datasets comparing to the ones of 5-10 loci. Specifically, we see a more significant improvement of performance from sampling 10 loci to 50 loci in the pooled data plots for both the sympatric and allopatric models. This is not surprising because each locus represents an independent genealogy that is subject to many stochastic effects. Sampling only a few loci would provide limited informative signal from the stochastic noises, while increasing the number of loci would increase the amount of informative signal that reveal the true relationships, thus contributing to the correct species delimitation.

The magnitude of Ne did not show a strong effect on the number of species recovered when pooling the data from different *m, t* and number of loci, compared to other single-factor plots such as 'number of species' vs. '*m*' and 'number of species' vs. 'number of loci' (Figure

4.5). This is probably due to the significant interactions between Ne and *m*/*t* as supported by ANOVA. Specifically, increasing *m* generally decreased the number of species recovered, and increasing *t* generally increased the number of species recovered. Therefore when the data from all *m*'s and *t*'s under the same Ne are pooled, these antagonistic effects can offset each other.

The results from *Psychotria* dataset did not recover the correct species delimitation. Brownie returned only two delimited species groups with *P. macrophylla* individuals assigned to one or both of the other two species, though *P. cyanococca* and *P. pilosa* individuals were separated completely into two species groups in both types of trees. *Psychotria macrophylla* always formed its own clade in the gene trees, but Brownie did not recognize it as a separate species. This might be due to both the limited number of loci sampled and the varying branch lengths in each gene trees (data not shown). The branches shown in Figure 4.6 were not drawn to scale. The low bootstrap support for the internal relationships among *P. cyanococca* and *P. pilosa* individuals on some of the gene trees correspond with clades with very short tip branches that can be collapsed into polytomies. This topological uncertainty may negatively affect the performance of Brownie (O'Meara 2010), and thus result in unreliable species relationships in the output. The poorly resolved species delimitation of *Psychotria* data is consistent with the results for the simulated data of the regionally sympatric speciation model, suggesting that the paraphyletic placement of *P. pilosa* can be caused by incomplete lineage sorting alone after speciation. This does not mean that we can rule out the possibility of on-going gene flow AFTER speciation, which is another factor that can result in discordance among the gene trees. Although our model did not incorporate on-going gene flow between populations (as the gene flow rate was set to zero after the point of speciation), the simulated data clearly showed that paraphyletic / polyphyletic species are likely to occur in the gene trees following even regionally sympatric

speciation events in large populations. Therefore, incomplete sorting alone is sufficient to explain the paraphyletic pattern of *P. pilosa* seen in some of the gene trees in the *Psychotria* data, but further research is still needed to specifically examine the possible effect of on-going gene flow as well as selection.

The overall results also support the point that species delimitation as well as species relationship inferences under sympatric species divergence is difficult, and that current coalescent based approaches are not able to deal with such scenarios very well. Speciation is a continuous process, in which the newly formed species would acquire different properties (such as phenetic differences, reproductive isolation, ecological distinction, reciprocal monophyly, etc.) at different times along their history of divergence. The timing of each property being obtained depends on the interactions among random drift, natural selection and gene flow. These properties form an intermediate stage where species boundaries can be blurry before reciprocally monophyly is reached at the whole genome level for the new species. Therefore, depending on the time of sampling and the types of character examined, species delimitation and the reconstruction of species relationships can be difficult, just like what is observed in *Psychotria*, and maybe much more other species distributed in the tropics.

**Table 4.1** Summary of ANOVA results for the simulated data.

| **3-population model** | | | | **2-population model** | | | |
|---|---|---|---|---|---|---|---|
| **Term** | **Df** | **F value** | **P** | **Term** | **Df** | **F value** | **P** |
| Ne | 1 | 107.25 | <2e-16 *** | Ne | 1 | 81.042 | <2e-16 *** |
| *m* | 1 | 131.697 | <2e-16 *** | *t* | 1 | 77.767 | <2e-16 *** |
| *t* | 1 | 94.762 | <2e-16 *** | nloci | 1 | 101.365 | <2e-16 *** |
| nloci | 1 | 173.564 | <2e-16 *** | Ne:*t* | 1 | 51.902 | 3.01e-12 *** |
| Ne:*m* | 1 | 68.945 | <2e-16 *** | Ne:nloci | 1 | 0.451 | 0.50229 |
| Ne:*t* | 1 | 94.198 | <2e-16 *** | *t*:nloci | 1 | 0.362 | 0.54764 |
| m:*t* | 1 | 2.058 | 0.15164 | Ne:*t*:nloci | 1 | 11.37 | 0.00082 *** |
| Ne:nloci | 1 | 0.064 | 0.80023 | | | | |
| *m*:nloci | 1 | 0.639 | 0.42404 | | | | |
| *t*:nloci | 1 | 9.041 | 0.00268 ** | | | | |
| Ne:*m*:*t* | 1 | 0.749 | 0.38698 | | | | |
| Ne:*m*:nloci | 1 | 1.775 | 0.18293 | | | | |
| Ne:*t*:nloci | 1 | 0.521 | 0.47047 | | | | |
| *m*:*t*:nloci | 1 | 0.202 | 0.65348 | | | | |
| Ne:*m*:*t*:nloci | 1 | 0.04 | 0.84191 | | | | |

Significance Codes: P <=0.001 ***; P<=0.01 **; P<=0.05 *

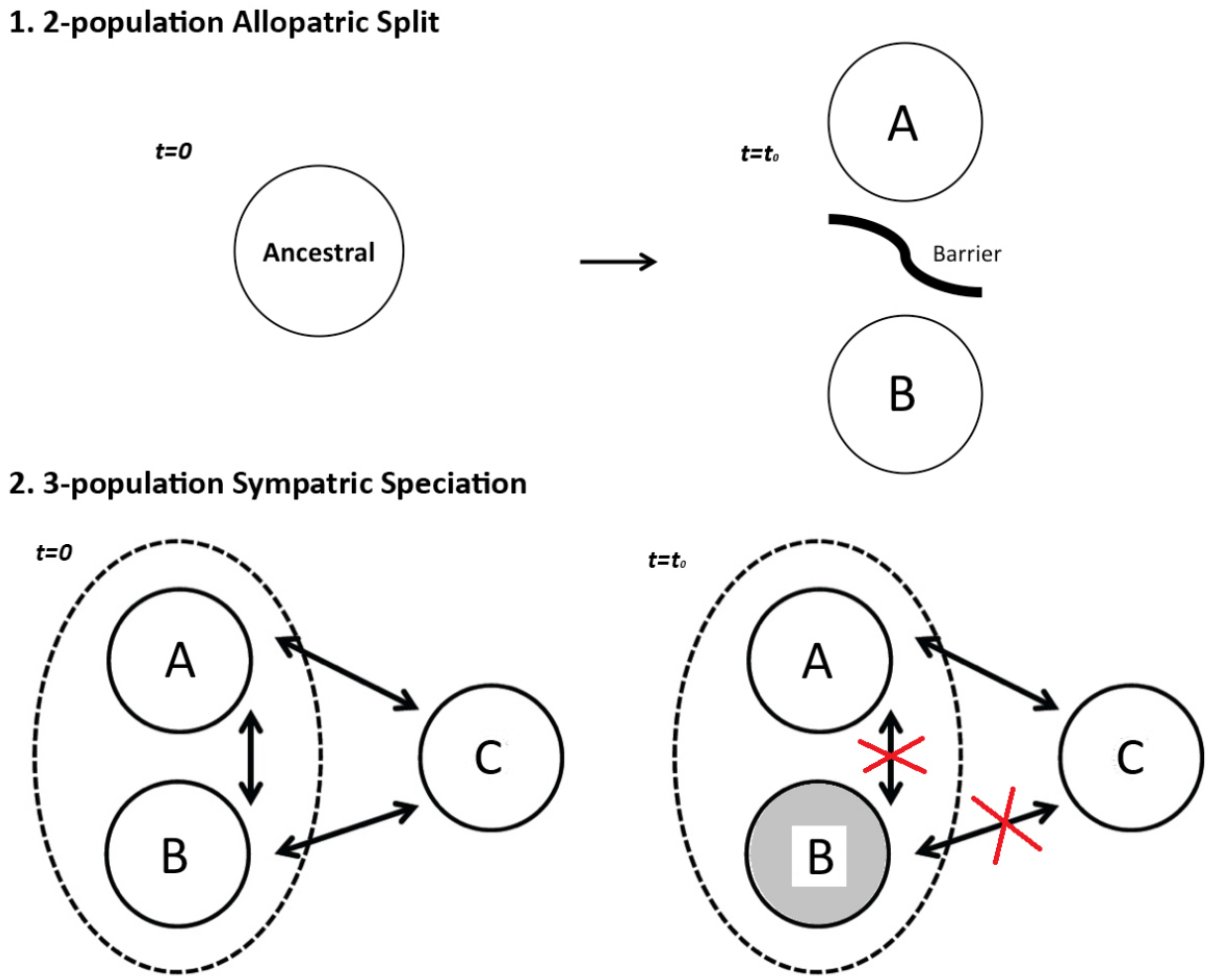


**Figure 4.1** Simulation models. 1. Two-population allopatric split (upper panel). At $t$=0 (left), the ancestral population were not separated. At $t$=$t_0$ (right), the population was split into daughter populations A and B permanently by a geographical barrier. A and B evolve independently into two species. 2. Three-population sympatric speciation (lower panel): At $t$=0 (left), populations A, B, C exist in a regionally sympatric area. A and B are sympatric as indicated by the dashed circle around them. Migration rate $m$ were set as $m_{ab}=m_{ba}=2m_{ac}=2m_{ca}=2m_{bc}=2m_{cb}$. At $t$=$t_0$ (right), population B became a new species (indicated by gray shade), and thus migration between B and the other two populations ceased. The time of the split ($t$), $m$ and Ne was varied across simulations. Gene trees for each combination of parameters were simulated for 5, 10, 50 and 100 loci respectively.

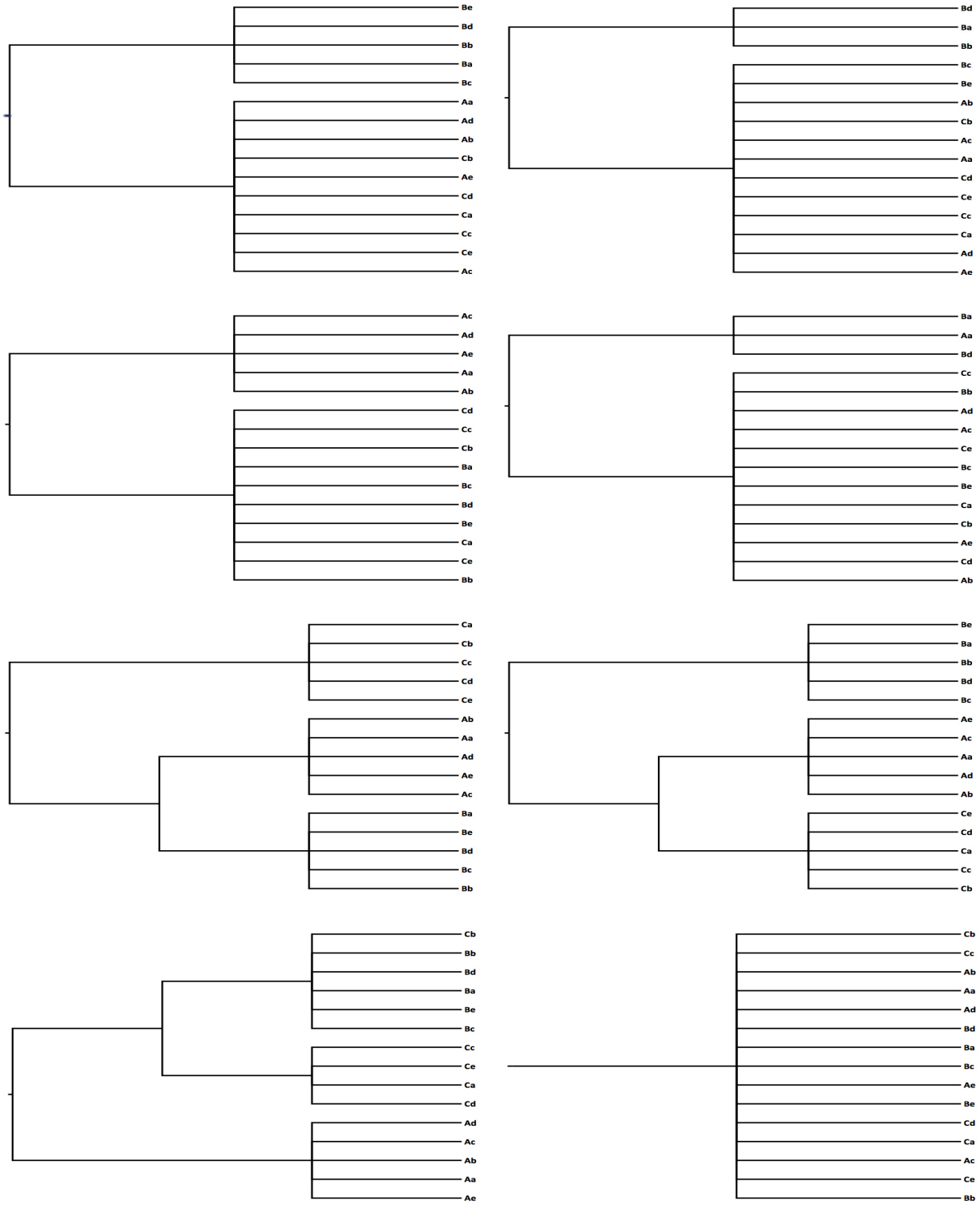


**Figure 4.2** Species delimitation results from the 3-population simulated data set. All types of output are shown represented by trees of samples. Upper case letters (A, B and C) indicate populations, and lower case letters (a – e) indicate different individuals. Each polytomy on the tree represents a single species. The "true" delimited species tree is shown on the top left.

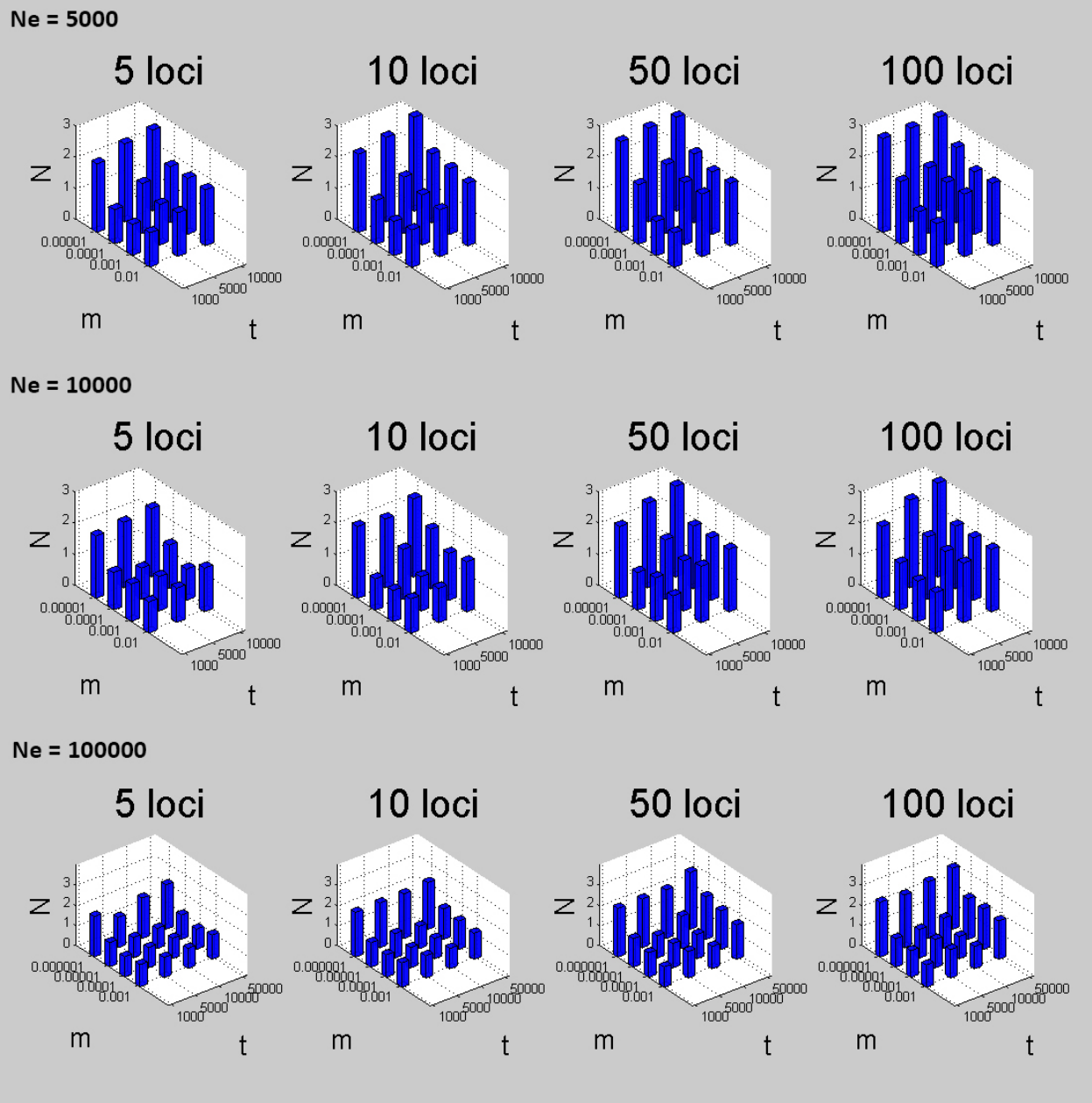


**Figure 4.3** Summary of the average number of species recovered for the 3-population simulated data across all parameter combinations. Each panel represents the results from one specific Ne (as indicated on top of the panel). Within each panel, each graph shows the results from a specific number of loci sampled (labeled on top of each graph). The horizontal axes represent the divergence time (*t*, in generations) and the level of gene flow (*m*); the vertical axis shows the average number of species recovered across 10 replicates for each parameter combination (N) from Brownie (note the different ranges of N, *m* and *t* on different graphs). Each column represents the result from a specific parameter combination.

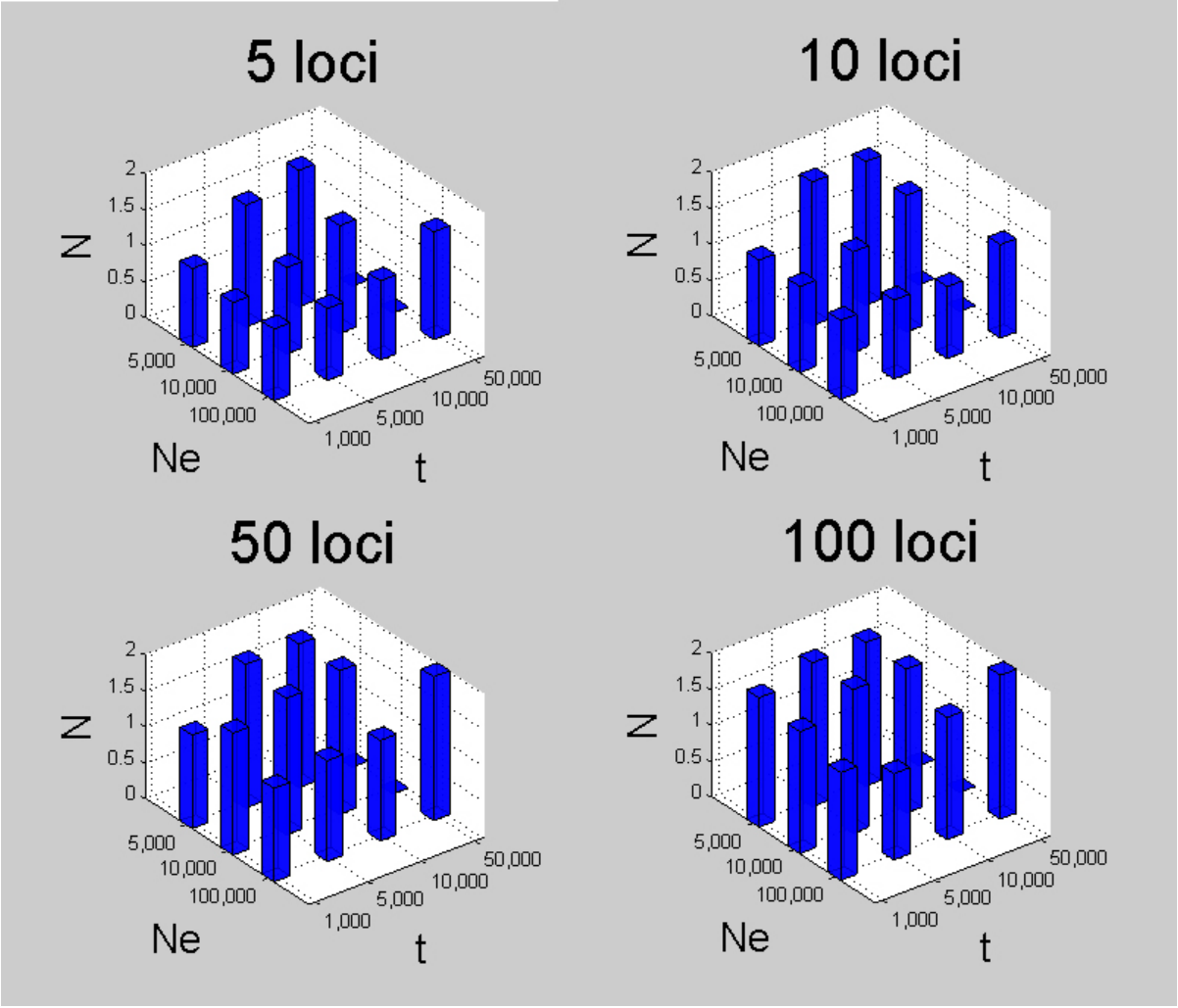


**Figure 4.4** Summary of the average number of species recovered for the 2-population simulated data across all parameter combinations. Each graph shows the results from a specific number of loci sampled (labeled on top of each graph). The horizontal axes represent the depth of the species split (t, in generations) and the effective population size (Ne); the vertical axis shows the number of species recovered (N) from Brownie. Each column represents the result from a specific parameter combination.

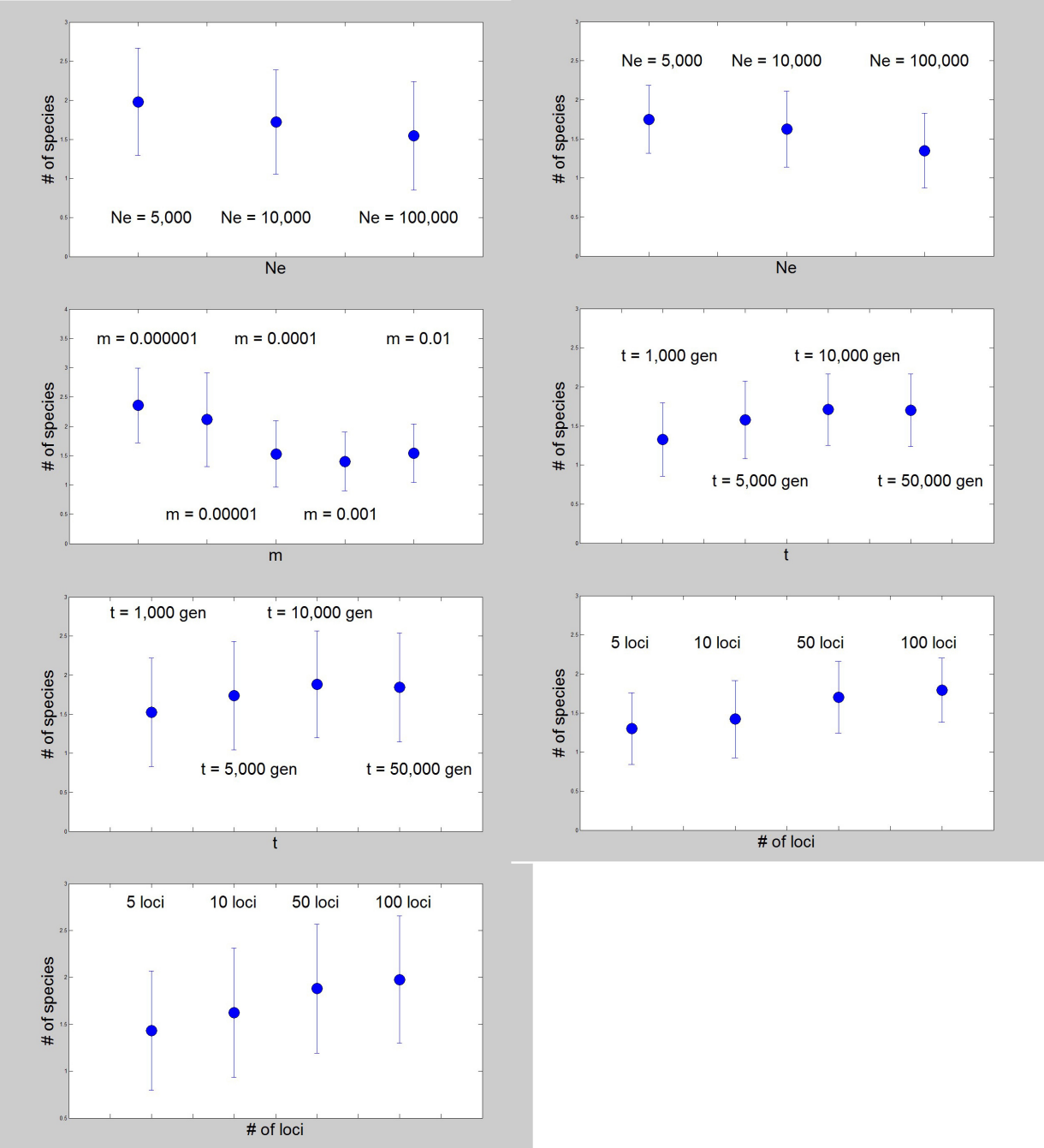


**Figure 4.5** Comparisons of single factor effects on the number of species retained. The left column shows the results from the 3-population simulation, and the right column shows the results from the 2-population simulation.

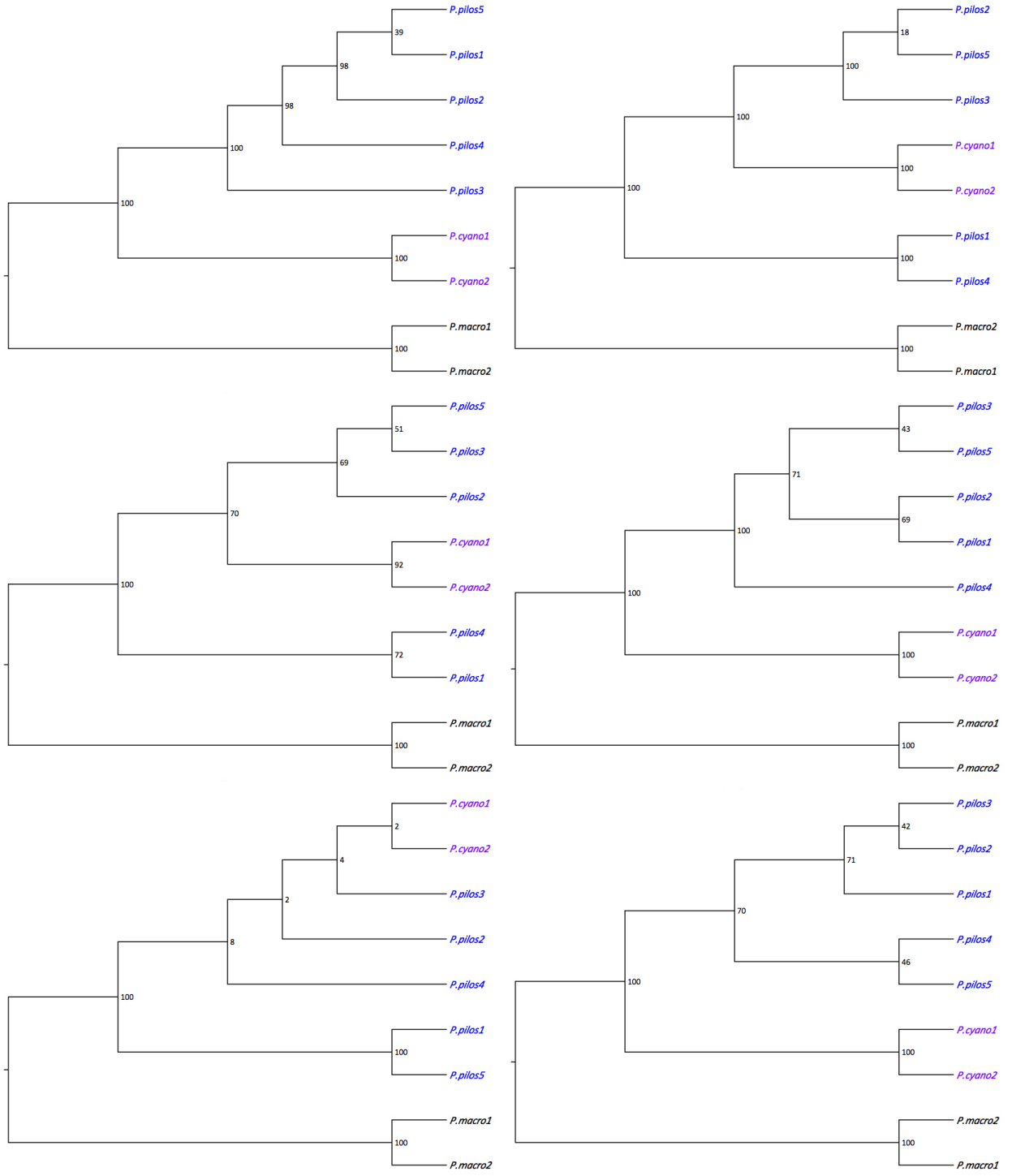


**Figure 4.6** Gene trees from the empirical *Psychotria* 3-species dataset. Tip labels: *P. cyano* – *P. cyanococca*; *P.pilos* – *P.pilosa*; *P.macro* – *P. macrophylla*. Numbers indicate different individuals from each species. *P. cyanococca* and *P. pilosa* are labeled in purple and blue, respectively. Note that branch lengths are not to scale.

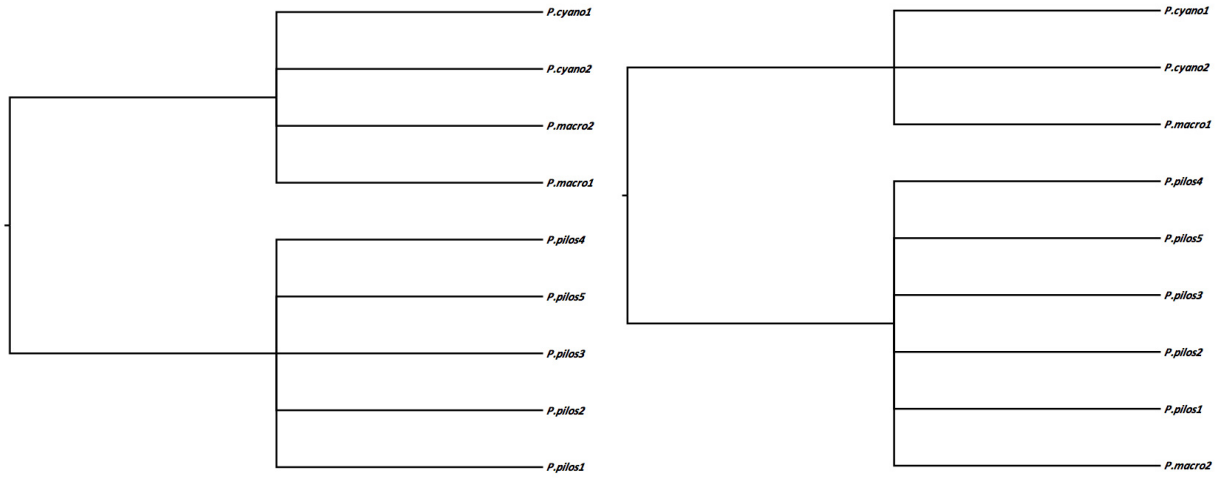


**Figure 4.7** Species delimitation results from the empirical *Psychotria* 3-species dataset. The two best trees resulted from Brownie are shown. Tip labels: *P. cyano – P. cyanococca*; *P. pilos – P. pilosa*; *P. macro – P. macrophylla*. Numbers indicate different individuals from each species.

# CHAPTER 5

# CONCLUSIONS

This dissertation contributes to the understanding of the evolutionary divergence in genus *Psychotria*. It is also the first phylogenetic study using coalescent approaches to resolve the species-level relationships among some New World species of *Psychotria*. In the first study, the complete chloroplast genome sequence of *P. marginata* was characterized, and universal markers flanking the non-coding regions of the genome were identified and developed for phylogenetic studies. The second study revealed the species-level relationships among 16 closely related New World *Psychotria* species. Paul et al. (2009) estimated that most lineages of Mesoamerican *Psychotria* species have diversified within the last 12 Ma. The recent and rapid radiation of this genus was also reflected in our gene trees and species trees by the ambiguous relationships towards the tips, and the polyphyletic patterns of haplotypes from *P. allenii* and *P. pilosa* shown in the gene trees provided evidence for cryptic species and incomplete sorting of lineages after the initial divergence of closely-related sister species. Coalescent simulations in the third study showed that paraphyletic species inferred in gene trees could be caused by random sorting process under (regionally) sympatric speciation model with structured ancestral populations (even with low rate of gene flow); and that species delimitation based on those gene tree data could be very difficult, particularly for species with large populations and short divergence times. Therefore, the paraphyletic pattern of *P. pilosa* and *P. cyanococca,* as

observed in our gene tree data, could be the result of a similar evolutionary process, although the possibility of on-going gene flow after speciation cannot be excluded and still needs to be tested.

The first study reported the whole chloroplast genome of *P. marginata*. It is the second species in family Rubiaceae to have the chloroplast genome sequenced. The circular chloroplast genome of *P. marginata* is 153430bp in total length, including a large (LSC, 84251bp) and small (SSC, 17607bp) single-copy regions separated by two inverted repeats (IRs, 25779bp each). Whole-genome sequence comparisons between *Psychotria* and *Coffea arabica* (Samson et al. 2007) showed that these two chloroplast genomes were totally co-linear in terms of gene content. *P. marginata* has part of *rps19* gene duplicated at the IRa–LSC boundary as a result of expansion of the IR, the same pattern as seen in *Coffea* (Samson et al. 2007) and also members of the related family Solanaceae (Chung et al. 2006). We observed in *Psychotria* that *infA* gene is a pseudogene, which is consistent with that in members of Solanaceae, although it was shown to be functional in *Coffea* (Samson et al. 2007). In addition, two more genes, *accD* and *rpl33* also contains stop codons in the middle of the ORFs so they might have become pseudogenes as well. Alignment of *P. marginata* and *C. arabica* chloroplast genome sequences yielded 10 primer sets designed flanking highly variable non-coding regions identified from the alignment, of which 9 of the primer pairs were successfully amplified in *Psychotria* and/or *Coffea.* Compared to the widely used *trnL-trnF* marker, all nine loci that had significantly longer sequences, and higher or at least similar level of sequence polymorphisms both within the genus and between genera. The Maximum-likelihood tree inferred from 3 of the loci, *psbE-petL*, *trnK-rps16*, *trnT-psbD* combined with *trnL-trnF* for 16 *Psychotria* and 3 other closely-related species in Rubiaceae showed significant improvement of bootstrap supports of the major clades than the ITS/*rbcL* tree estimated by Nepokroeff et al. (1999), suggesting that these sequence markers can be useful to

resolve inter-specific relationships within genus *Psychotria* and other closely related taxa.

In the second study, a multi-locus phylogenetic analysis was conducted to assess the pattern of diversification for 16 New World *Psychotria* species and 3 other species from closely related genera in Rubiaceae. Gene trees estimated from 5 nuclear (NL-103, NL-217, NL-57800, NL-A04 and ITS) and 4 chloroplast (*psbE-petL*, *trnT-psbD*, and *trnK-rps16* and *trnL-trnF,* concatenated) non-coding loci showed generally consistent topologies with previous studies on the genus-wide phylogeny (Nepokroeff et al. 1999). Strong bootstrap support was obtained for the two major clades corresponding to subgenus *Psychotria* and subgenus *Heteropsychotria,* but the tips of the tree were still relatively poorly resolved, indicating the rapid diversification of the genus. Moreover, some of the nuclear gene trees showed evidence of gene duplication (NL-A04, NL-57800, NL-217 and NL-103) that probably have occurred at the common ancestor of this genus with random loss in the descendant lineages, perhaps one of the driving forces of speciation. *P. pilosa* haplotypes were paraphyletic with *P. cyanococca* nested within them in 4 gene trees (ITS, chloroplast concatenated, NL-57800 and NL-217), a pattern of incomplete lineage sorting. *P. allenii* alleles are always polyphyletic, with a subset grouped with *Palicourea* and the rest grouped with *P. mortoniana,* suggesting that it is probably a cryptic species complex. Coalescent species tree analysis using ME-EST resulted in a consensus tree that is largely consistent with the gene trees. Short branches towards the tips with low bootstrap support further supported the rapid radiations of the New World *Psychotria* species, yet a significant improvement of statistical support for the clade of subgenus *Psychotria* was obtained compared to zero support for the relationships of the same taxa in Nepokroeff et al. (1999), with >60% bootstrap values for the internal relationships among *P. graciliflora, P. chagrensis, P. marginata* and *P. horizontalis* in the subgenus *Psychotria* clade.

Although the MP-EST species tree put *P. pilosa* and *P. cyanococca* as sister species, and grouped *P. allenii* with *Palicourea*, we saw fuzzy species boundaries in the gene trees for these taxa. As sympatric speciation probably occurs more frequently for species-rich groups in the tropics, we decided to generate a simple model of similar processes in coalescent simulations in the third study to compare with the patterns seen in *Psychotria* gene trees. With the reduced 3-species *Psychotria* empirical data (*P.pilosa*, *P.cyanococca* and *P.macrophylla*) from the six loci in the second study, we were not able to recover the correct species delimitation among those individuals. Simulated data revealed that for the speciation process with no absolute barrier to gene flow among sympatric populations, the combined effect of lineage sorting plus gene flow is likely to result in polyphyletic/paraphyletic species in the gene trees, and that species delimitation is quite difficult in terms of correctly assigning randomly sampled 'unknown' individuals from different populations into monophyletic species groups, particularly when the population size (Ne) is large and the divergence time (*t*) is short, even under very limited gene flow between ancestral populations. On the other hand, although Ne and *t* still had the same pattern of effect, allopatric speciation with no gene flow seems to be much easier to form clear species boundaries just by stochastic sorting, and random sampling of extant alleles is relatively less likely to result in incorrect species delimitation comparing to the sympatric scenario. In either case, increasing sample size from 5 loci to 100 loci is helpful for recovering the correct species delimitation, but the effect tends to be masked by large effective population sizes. These results support the idea that even under neutral evolutionary processes, the speciation event is just a starting point, and that newly formed sister species will require a time period for genetic drift to act upon before becoming reciprocally monophyletic. The length of this intermediate period will depend on the effective population size and the level of genetic structures among

ancestral populations, as well as possible on-going gene flow after the speciation event. If sampling occurs during this period, gene trees are likely to show paraphyletic patterns for one of the two species, at least for a portion of loci sampled across the genome, as seen in *P.pilosa* and *P. cyanococca*.

# REFERENCES

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Moest, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Vainola, J. B. W. Wolf, and D. Zinner. 2013. Hybridization and speciation. Journal of Evolutionary Biology **26**:229-246.

Alvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution **29**:417-434.

Andersson, L. 2002. Relationships and generic circumscriptions in the Psychotria complex (Rubiaceae, Psychotrieae). Systematics and Geography of Plants **72**:167-202.

Andersson, L., and J. H. E. Rova. 1999. The rps16 intron and the phylogeny of the Rubioideae (Rubiaceae). Plant Systematics and Evolution **214**:161-186.

Arnold, M. 2006. Evolution through genetic exchange. Oxford University Press.

Barniske, A. M., T. Borsch, K. Muller, M. Krug, A. Worberg, C. Neinhuis, and D. Quandt. 2012. Phylogenetics of early branching eudicots: Comparing phylogenetic signal across plastid introns, spacers, and genes. Journal of Systematics and Evolution **50**:85-108.

Barrabe, L., S. Buerki, A. Mouly, A. P. Davis, J. Munzinger, and L. Maggia. 2012. Delimitation of the genus Margaritopsis (Rubiaceae) in the Asian, Australasian and Pacific region, based on molecular phylogenetic inference and morphology. Taxon **61**:1251-1268.

Blaxter, M. L. 2004. The promise of a DNA taxonomy. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **359**:669-679.

Borhidi, A. 2011. Transfer of the Mexican Species of Psychotria Subg. Heteropsychotria to Palicourea Based on Morphological and Molecular Evidence. Acta Botanica Hungarica **53**:241-250.

Bremer, B. 2009. A Review of Molecular Phylogenetic Studies of Rubiaceae. Annals of the Missouri Botanical Garden **96**:4-26.

Bremer, B., and T. Eriksson. 2009. Time tree of Rubiaceae: phylogeny and dating the family, subfamilies, and tribes. International journal of plant sciences **170**:766-793.

Bremer, B., and J. F. Manen. 2000. Phylogeny and classification of the subfamily Rubioideae (Rubiaceae). Plant Systematics and Evolution **225**:43-72.

Brower, A. V. Z., R. DeSalle, and A. Vogler. 1996. Gene trees, species trees, and systematics: A cladistic perspective. Annual Review of Ecology and Systematics **27**:423-450.

Bryant, N., J. Lloyd, C. Sweeney, F. Myouga, and D. Meinke. 2011. Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in Arabidopsis. Plant Physiol **155**:1678-1689.

Cai, Z. Q., C. Penaflor, J. V. Kuehl, J. Leebens-Mack, J. E. Carlson, C. W. dePamphilis, J. L. Boore, and R. K. Jansen. 2006. Complete plastid genome sequences of Drimys, Liriodendron, and Piper: implications for the phylogenetic relationships of magnoliids. BMC Evolutionary Biology **6**.

Camargo, A., M. Morando, L. J. Avila, and J. W. Sites. 2012. Species Delimitation with ABC and Other Coalescent-Based Methods: A Test of Accuracy with Simulations and an Empirical Example with Lizards of the Liolaemus darwinii complex (Squamata: Liolaemidae). Evolution **66**:2834-2849.

Camargo, A., and J. J. Sites. 2013. Species Delimitation: A Decade After the Renaissance.

Carstens, B. C., and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from Melanoplus grasshoppers. Systematic Biology **56**:400-411.

Chandler, G. T., and M. D. Crisp. 1998. Morphometric and phylogenetic analysis of the Daviesia ulicifolia complex (Fabaceae, Mirbelieae). Plant Systematics and Evolution **209**:93-122.

Chapman, M., J. Chang, D. Weisman, R. Kesseli, and J. Burke. 2007. Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae (Compositae). Theoretical and Applied Genetics **115**:747-755.

Chase, M. W., R. S. Cowan, P. M. Hollingsworth, C. van den Berg, S. Madrinan, G. Petersen, O. Seberg, T. Jorgsensen, K. M. Cameron, M. Carine, N. Pedersen, T. A. J. Hedderson, F. Conrad, G. A. Salazar, J. E. Richardson, M. L. Hollingsworth, T. G. Barraclough, L. Kelly, and M. Wilkinson. 2007. A proposal for a standardised protocol to barcode all land plants. Taxon **56**:295-299.

Chase, M. W., N. Salamin, M. Wilkinson, J. M. Dunwell, R. P. Kesanakurthi, N. Haidar, and V. Savolainen. 2005. Land plants and DNA barcodes: short-term and long-term goals. Philosophical Transactions of the Royal Society B-Biological Sciences **360**:1889-1895.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, and Y.-L. Qiu. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Garden:528-580.

Chung, H. J., J. D. Jung, H. W. Park, J. H. Kim, H. W. Cha, S. R. Min, W. J. Jeong, and J. R. Liu. 2006. The complete chloroplast genome sequences of Solanum tuberosum and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. Plant Cell Rep **25**:1369-1379.

Chung, Y., and C. Ane. 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Systematic Biology **60**:261-275.

Comes, H. P., and R. J. Abbott. 2001. Molecular phylogeography, reticulation, and lineage sorting in Mediterranean Senecio sect. Senecio (Asteraceae). Evolution **55**:1943-1962.

Correa, A. M., and E. R. Forni-Martins. 2004. Chromosomal studies of species of Rubiaceae (A. L. de Jussieu) from the Brazilian cerrado. Caryologia **57**:250-258.

Correa, A. M., S. L. Jung-Mendacolli, and E. R. Forni-Martins. 2010. Karyotype characterisation of Brazilian species of the genus Psychotria L. - subfamily Rubioideae (Rubiaceae). Kew Bulletin **65**:45-52.

Cruzan, M. B. 1998. Genetic markers in plant evolutionary ecology. Ecology **79**:400-412.

Cullingham, C. I., P. M. A. James, J. E. K. Cooke, and D. W. Coltman. 2012. Characterizing the physical and genetic structure of the lodgepole pine jack pine hybrid zone: mosaic structure and differential introgression. Evolutionary Applications **5**:879-891.

Davis, A. P., M. Chester, O. Maurin, and M. F. Fay. 2007. Searching for the relatives of Coffea (Rubiaceae, Ixoroideae): The circumscription and phylogeny of coffeeae based on plastid sequence data and morphology. American Journal of Botany **94**:313-329.

Davis, A. P., R. Govaerts, D. M. Bridson, M. Ruhsam, J. Moat, and N. A. Brummitt. 2009. A Global Assessment of Distribution, Diversity, Endemism, and Taxonomic Effort in the Rubiaceae. Annals of the Missouri Botanical Garden **96**:68-78.

de Queiroz, A., and J. Gatesy. 2007. The supermatrix approach to systematics. Trends in Ecology & Evolution **22**:34-41.

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation. Pages 57-75 *in* D. J. H. S. H. Berlocher, editor. Endless Forms: Species and Speciation. Oxford University Press.

de Queiroz, K. 2007. Species Concepts and Species Delimitation. Systematic Biology **56**:879-886.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet **2**:e68.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution **24**:332-340.

Delannoy, E., S. Fujii, C. C. des Francs-Small, M. Brundrett, and I. Small. 2011. Rampant Gene Loss in the Underground Orchid Rhizanthella gardneri Highlights Evolutionary Constraints on Plastid Genomes. Molecular Biology and Evolution **28**:2077-2086.

Der Marderosian, A. H., K. Kensinger, F. Goldstein, and J. Chao. 1969. The use and hallucinatory principles of a psychoactive beverage of the Cashinahua tribe (Amazon basin). XI International Botanical Congress:45-45.

Doyle, J. J. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull **19**:11-15.

Doyle, J. J. 1992. Gene Trees and Species Trees - Molecular Systematics as One-Character Taxonomy. Systematic Botany **17**:144-163.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. Bmc Bioinformatics **5**:1-19.

Edwards, C. E., D. E. Soltis, and P. S. Soltis. 2008. Using patterns of genetic structure based on microsatellite loci to test hypotheses of current hybridization, ancient hybridization and incomplete lineage sorting in Conradina (Lamiaceae). Molecular Ecology **17**:5157-5174.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution **63**:1-19.

Ence, D. D., and B. C. Carstens. 2011. SpedeSTEM: a rapid and accurate method for species delimitation. Mol Ecol Resour **11**:473-480.

Ewing, G., and J. Hermisson. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**:2064-2065.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package). . Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fitch, W. M. 1995. Uses for evolutionary trees. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **349**:93-102.

Freudenstein, J. V., and J. J. Doyle. 1994. Plastid DNA, Morphological Variation, and the Phylogenetic Species Concept - the Corallorhiza-Maculata (Orchidaceae) Complex. Systematic Botany **19**:273-290.

Frodin, D. G. 2004. History and concepts of big plant genera. Taxon **53**:753-776.

Fujita, M. K., A. D. Leaché, F. T. Burbrink, J. A. McGuire, and C. Moritz. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology & Evolution **27**:480-488.

Gadberry, M. D., S. T. Malcomber, A. N. Doust, and E. A. Kellogg. 2005. Primaclade—a flexible tool to find conserved PCR primers across multiple species. Bioinformatics **21**:1263-1264.

Gielly, L., and P. Taberlet. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. Molecular Biology and Evolution **11**:769-777.

Golenberg, E. M., M. T. Clegg, M. L. Durbin, J. Doebley, and D. P. Ma. 1993. Evolution of a noncoding region of the chloroplast genome. Molecular Phylogenetics and Evolution **2**:52-64.

Govaerts, R. A., L; Robbrecht E et al. 2006. World Checklist of Rubiaceae.

Guisinger, M. M., T. W. Chumley, J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2010. Implications of the Plastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. Journal of Molecular Evolution **70**:149-166.

Guo, X., S. Castillo-Ramirez, V. Gonzalez, P. Bustos, J. L. Fernandez-Vazquez, R. I. Santamaria, J. Arellano, M. A. Cevallos, and G. Davila. 2007. Rapid evolutionary change of common bean (Phaseolus vulgaris L) plastome, and the genomic diversification of legume chloroplasts. BMC Genomics **8**:228.

Hamilton, C. W. 1989a. A Revision of Mesoamerican Psychotria Subgenus Psychotria (Rubiaceae) .1. Introduction and Species 1-16. Annals of the Missouri Botanical Garden **76**:67-111.

Hamilton, C. W. 1989b. A Revision of Mesoamerican Psychotria Subgenus Psychotria (Rubiaceae), Part II: Species 17-47. Annals of the Missouri Botanical Garden **76**:386-429.

Hamilton, C. W. 1989c. A Revision of Mesoamerican Psychotria Subgenus Psychotria (Rubiaceae), Part III: Species 48-61 and Appendices. Annals of the Missouri Botanical Garden **76**:886-916.

Hand, M. L., G. C. Spangenberg, J. W. Forster, and N. O. I. Cogan. 2013. Plastome Sequence Determination and Comparative Analysis for Members of the Lolium-Festuca Grass Species Complex. G3-Genes Genomes Genetics **3**:607-616.

Harrison, R. G. 1998. Linking evolutionary pattern and process. Pages 19-31 *in* D. J. H. S. H. Berlocher, editor. Endless Forms: Species and Speciation. Oxford University Press.

Hausdorf, B. 2011. Progress toward a general species concept Evolution **65**:923-931.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences **270**:313-321.

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution **27**:570-580.

Hey, J. 2001. The mind of the species problem. Trends in Ecology & Evolution **16**:326-329.

Hiratsuka, J., H. Shimada, R. Whittier, T. Ishibashi, M. Sakamoto, M. Mori, C. Kondo, Y. Honji, C.-R. Sun, and B.-Y. Meng. 1989. The complete sequence of the rice (Oryza sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Molecular and General Genetics MGG **217**:185-194.

Hollingsworth, P. M., L. L. Forrest, J. L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. W. Chase, R. S. Cowan, D. L. Erickson, A. J. Fazekas, S. W. Graham, K. E. James, K. J. Kim, W. J. Kress, H. Schneider, J. van AlphenStahl, S. C. H. Barrett, C. van den Berg, D. Bogarin, K. S. Burgess, K. M. Cameron, M. Carine, J. Chacon, A. Clark, J. J. Clarkson, F. Conrad, D. S. Devey, C. S. Ford, T. A. J. Hedderson, M. L. Hollingsworth, B. C. Husband, L. J. Kelly, P. R. Kesanakurti, J. S. Kim, Y. D. Kim, R. Lahaye, H. L. Lee, D. G. Long, S. Madrinan, O. Maurin, I. Meusnier, S. G. Newmaster, C. W. Park, D. M. Percy, G. Petersen, J. E. Richardson, G. A. Salazar, V. Savolainen, O. Seberg, M. J. Wilkinson, D. K. Yi, D. P. Little, and C. P. W. Grp. 2009. A DNA barcode for land plants. Proc Natl Acad Sci U S A **106**:12794-12797.

Huang, H., Q. He, L. S. Kubatko, and L. L. Knowles. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Systematic Biology **59**:573-583.

Huang, H., and L. L. Knowles. 2009. What Is the Danger of the Anomaly Zone for Empirical Phylogenetics? Systematic Biology **58**:527-536.

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pages 1-44 Futuyma, D. And J. Antonovics.

Jakob, S. S., and F. R. Blattner. 2006. A chloroplast genealogy of Hordeum (Poaceae): Long-term persisting haplotypes, incomplete lineage sorting, regional extinction, and the consequences for phylogenetic inference. Molecular Biology and Evolution **23**:1602-1612.

Jansen, R. K., Z. Cai, L. A. Raubeson, H. Daniell, C. W. dePamphilis, J. Leebens-Mack, K. F. Müller, M. Guisinger-Bellian, R. C. Haberle, A. K. Hansen, T. W. Chumley, S.-B. Lee,

R. Peery, J. R. McNeal, J. V. Kuehl, and J. L. Boore. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences **104**:19369-19374.

Jo, Y. D., J. Park, J. Kim, W. Song, C. G. Hur, Y. H. Lee, and B. C. Kang. 2011. Complete sequencing and comparative analyses of the pepper (Capsicum annuum L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome. Plant Cell Rep **30**:217-229.

Johansson, J. T. 1993. Pollen morphology in Psychotria (Rubiaceae, Rubioideae, Psychotrieae) and its taxonomic significance. A preliminary survey. Nordic Journal of Botany **13**:32-32.

Joly, S., and A. Bruneau. 2009. Measuring Branch Support in Species Trees Obtained by Gene Tree Parsimony. Systematic Biology **58**:100-113.

Kahlau, S., S. Aspinall, J. Gray, and R. Bock. 2006. Sequence of the Tomato Chloroplast DNA and Evolutionary Comparison of Solanaceous Plastid Genomes. Journal of Molecular Evolution **63**:194-207.

Kane, N. C., J. M. Burke, L. Marek, G. Seiler, F. Vear, G. Baute, S. J. Knapp, P. Vincourt, and L. H. Rieseberg. 2013. Sunflower genetic, genomic and ecological resources. Molecular Ecology Resources **13**:10-20.

Kiehn, M. 2010. Chromosomes of Neotropical Rubiaceae. I: Rubioideae. Annals of the Missouri Botanical Garden **97**:91-105.

Kingman, J. F. C. 2000. Origins of the Coalescent: 1974-1982. Genetics **156**:1461-1463.

Kode, V., E. A. Mudd, S. Iamtham, and A. Day. 2005. The tobacco plastid accD gene is essential and is required for leaf development. Plant J **44**:237-244.

Kress, W. J., K. J. Wurdack, E. A. Zimmer, L. A. Weigt, and D. H. Janzen. 2005. Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci U S A **102**:8369-8374.

Ku, C., W. C. Chung, L. L. Chen, and C. H. Kuo. 2013. The Complete Plastid Genome Sequence of Madagascar Periwinkle Catharanthus roseus (L.) G. Don: Plastid Genome Evolution, Molecular Marker Identification, and Phylogenetic Implications in Asterids. PLoS One **8**.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics **25**:971-973.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology **56**:17-24.

Lanave, C., G. Preparata, C. Sacone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution **20**:86-93.

Leache, A. D., and B. Rannala. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology **60**:126-137.

Liu, L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics **24**:2542-2543.

Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating Species Trees Using Multiple-Allele DNA Sequence Data. Evolution **62**:2080-2091.

Liu, L., L. Yu, and S. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology **10**:302.

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009a. Coalescent methods for estimating phylogenetic trees. Molecular Phylogenetics and Evolution **53**:320-328.

Liu, L., L. L. Yu, D. K. Pearl, and S. V. Edwards. 2009b. Estimating Species Phylogenies Using Coalescence Times among Sequences. Systematic Biology **58**:468-477.

Lopes, S., G. L. von Poser, V. A. Kerber, F. M. Farias, E. L. Konrath, P. Moreno, M. E. Sobral, J. A. S. Zuanazzi, and A. T. Henriques. 2004. Taxonomic significance of alkaloids and iridoid glucosides in the tribe Psychotrieae (Rubiaceae). Biochemical Systematics and Ecology **32**:1187-1195.

Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology **46**:523-536.

Maddison, W. P., and L. L. Knowles. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. Systematic Biology **55**:21-30.

Magee, A. M., S. Aspinall, D. W. Rice, B. P. Cusack, M. Semon, A. S. Perry, S. Stefanovic, D. Milbourne, S. Barth, J. D. Palmer, J. C. Gray, T. A. Kavanagh, and K. H. Wolfe. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res **20**:1700-1710.

Maier, R. M., K. Neckermann, G. L. Igloi, and H. Kössel. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. Journal of molecular biology **251**:614-628.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends in Ecology & Evolution **20**:229-237.

Mao, X., G. He, P. Hua, G. Jones, S. Zhang, and S. J. Rossiter. 2013. Historical introgression and the persistence of ghost alleles in the intermediate horseshoe bat (Rhinolophus affinis). Molecular Ecology **22**:1035-1050.

Martínez-Cabrera, D., T. Terrazas, and H. Ochoterena. 2009. Foliar and petiole anatomy of tribe hamelieae and other rubiaceae. Annals of the Missouri Botanical Garden **96**:133-145.

McCormack, J. E., H. Huang, and L. L. Knowles. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Systematic Biology **58**:501-508.

Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Gray, C. W. Morden, P. J. Calie, L. S. Jermiin, and K. H. Wolfe. 2001. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell **13**:645-658.

Moore, M. J., P. S. Soltis, C. D. Bell, J. G. Burleigh, and D. E. Soltis. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci U S A **107**:4623-4628.

Moraes, T. M. D., G. R. Rabelo, C. R. Alexandrino, S. J. D. Neto, and M. Da Cunha. 2011. Comparative leaf anatomy and micromorphology of Psychotria species (Rubiaceae) from the Atlantic Rainforest. Acta Botanica Brasilica **25**:178-190.

Mossel, E., and S. Roch. 2010. Incomplete Lineage Sorting: Consistent Phylogeny Estimation from Multiple Loci. Ieee-Acm Transactions on Computational Biology and Bioinformatics **7**:166-171.

Mueller, L. A., T. H. Solow, N. Taylor, B. Skwarecki, R. Buels, J. Binns, C. Lin, M. H. Wright, R. Ahrens, Y. Wang, E. V. Herbst, E. R. Keyder, N. Menda, D. Zamir, and S. D. Tanksley. 2005. The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond. Plant Physiology **138**:1310-1317.

Nepokroeff, M., B. Bremer, and K. J. Sytsma. 1999. Reorganization of the genus Psychotria and tribe Psychotrieae (Rubiaceae) inferred from ITS and rbcL sequence data. Systematic Botany **24**:5-27.

Nepokroeff, M., K. J. Sytsma, W. L. Wagner, and E. A. Zimmer. 2003. Reconstructing ancestral patterns of colonization and dispersal in the Hawaiian understory tree genus Psychotria (Rubiaceae): A comparison of parsimony and likelihood approaches. Systematic Biology **52**:820-838.

Newton, A. C., T. R. Allnutt, A. C. M. Gillies, A. J. Lowe, and R. A. Ennos. 1999. Molecular phylogeography, intraspecific variation and the conservation of tree species. Trends in Ecology & Evolution **14**:140-145.

Nie, X. J., S. Z. Lv, Y. X. Zhang, X. H. Du, L. Wang, S. S. Biradar, X. F. Tan, F. H. Wan, and W. N. Song. 2012. Complete Chloroplast Genome Sequence of a Major Invasive Species, Crofton Weed (Ageratina adenophora). PLoS One **7**.

Niemiller, M. L., T. J. Near, and B. M. Fitzpatrick. 2012. Delimiting species using multilocus data: diagnosing cryptic diversity in the southern cavefish, Typhlichthys subterraneus (Teleostei: Amblyopsidae). Evolution **66**:846-866.

Nosil, P. 2008. Speciation with gene flow could be common. Molecular Ecology **17**:2103-2106.

O'Meara, B. C. 2010. New Heuristic Methods for Joint Species Delimitation and Species Tree Inference. Systematic Biology **59**:59-73.

O'Meara, B. C., C. Ane, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution **60**:922-933.

Olmstead, R. G., and J. D. Palmer. 1994. Chloroplast DNA Systematics: A Review of Methods and Data Analysis. American Journal of Botany **81**:1205-1224.

Openshaw, H. 1970. The Ipecacuanha. D Alkaloids. Pages 85-115 *in* S. W. Pelletier, editor. Chemistry of the Alkaloids. Illus. Van Nostrand Reinhold Co., Div. of Litton Educational Publishing, Inc., New York, N.Y., U.S.A.

Ovcharenko, I., G. G. Loots, B. M. Giardine, M. M. Hou, J. Ma, R. C. Hardison, L. Stubbs, and W. Miller. 2005. Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. Genome Research **15**:184-194.

Ovcharenko, I., G. G. Loots, R. C. Hardison, W. Miller, and L. Stubbs. 2004. zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. Genome Research **14**:472-477.

Page, R. D. M., and M. A. Charleston. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem. Molecular Phylogenetics and Evolution **7**:231-240.

Palmer, J. D. 1985. Comparative organization of chloroplast genomes. Annu Rev Genet **19**:325-354.

Palmer, J. D. 1986. Isolation and structural analysis of chloroplast DNA. Pages 167-186 *in* H. W. Arthur Weissbach, editor. Methods in Enzymology. Academic Press.

Pamilo, P., and M. Nei. 1988. Relationships between Gene Trees and Species Trees. Molecular Biology and Evolution **5**:568-583.

Parks, M., R. Cronn, and A. Liston. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. Bmc Biology **7**.

Paul, J. R., C. Morton, C. M. Taylor, and S. J. Tonsor. 2009. Evolutionary Time for Dispersal Limits the Extent but Not the Occupancy of Species' Potential Ranges in the Tropical Plant Genus Psychotria (Rubiaceae). American Naturalist **173**:188-199.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread Discordance of Gene Trees with Species Tree in <italic>Drosophila:</italic> Evidence for Incomplete Lineage Sorting. PLoS Genet **2**:e173.

Rannala, B., and Z. H. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics **164**:1645-1656.

Ratan, A. 2009. Assembly algorithms for next-generation sequence data. PhD Thesis. The Pennsylvania State University.

Raubeson, L. J., RK. 2005. Plant diversity and evolution : genotypic and phenotypic variation in higher plants. CABI Pub., Wallingford, Oxfordshire, UK ; Cambridge, MA.

Razafimandimbison, S. G., C. Rydin, and B. Bremer. 2008. Evolution and trends in the Psychotrieae alliance (Rubiaceae) - A rarely reported evolutionary change of many-seeded carpels from one-seeded carpels. Molecular Phylogenetics and Evolution **48**:207-223.

Rieseberg, L. H., and L. Brouillet. 1994. Are Many Plant-Species Paraphyletic. Taxon **43**:21-32.

Rieseberg, L. H., R. Carter, and S. Zona. 1990. Molecular Tests of the Hypothesized Hybrid Origin of 2 Diploid Helianthus Species (Asteraceae). Evolution **44**:1498-1511.

Robbrecht, E., and J.-F. Manen. 2006. The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of Coptosapelta and Luculia, and supertree construction based on rbcL, rps16, trnL-trnF and atpB-rbcL data. A new classification in two subfamilies, Cinchonoideae and Rubioideae. Systematics and Geography of Plants:85-145.

Rogalski, M., M. A. Schottler, W. Thiele, W. X. Schulze, and R. Bock. 2008. Rpl33, a nonessential plastid-encoded ribosomal protein in tobacco, is required under cold stress conditions. Plant Cell **20**:2221-2237.

Rosenberg, N. A. 2013. Discordance of Species Trees with Their Most Likely Gene Trees: A Unifying Principle. Molecular Biology and Evolution **30**:2709-2713.

Rosenberg, N. A., and R. Tao. 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. Systematic Biology **57**:131-140.

Rousseau-Gueutin, M., X. Huang, E. Higginson, M. Ayliffe, A. Day, and J. N. Timmis. 2013. Potential Functional Replacement of the Plastidic Acetyl-CoA Carboxylase Subunit (accD) Gene by Recent Transfers to the Nucleus in Some Angiosperm Lineages. Plant Physiol **161**:1918-1929.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132**:365-386.

Rubinoff, D., S. Cameron, and K. Will. 2006. Are plant DNA barcodes a search for the Holy Grail? Trends in Ecology & Evolution **21**:1-2.

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

Sakai, S., and S. J. Wright. 2008. Reproductive ecology of 21 coexisting Psychotria species (Rubiaceae): when is heterostyly lost? Biological Journal of the Linnean Society **93**:125-134.

Samson, N., M. G. Bausher, S. B. Lee, R. K. Jansen, and H. Daniell. 2007. The complete nucleotide sequence of the coffee (Coffea arabica L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. Plant Biotechnology Journal **5**:339-353.

Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends in Ecology & Evolution **13**:105-109.

Sedio, B. E., J. R. Paul, C. M. Taylor, and C. W. Dick. 2013. Fine-scale niche structure of Neotropical forests reflects a legacy of the Great American Biotic Interchange. Nat Commun **4**.

Sedio, B. E., S. J. Wright, and C. W. Dick. 2012. Trait evolution and the coexistence of a species swarm in the tropical forest understorey. Journal of Ecology **100**:1183-1193.

Setiadi, M. I., J. A. McGuire, R. M. Brown, M. Zubairi, D. T. Iskandar, N. Andayani, J. Supriatna, and B. J. Evans. 2011. Adaptive radiation and ecological opportunity in Sulawesi and Philippine fanged frog (Limnonectes) communities. American Naturalist **178**:221-240.

Shaw, J., E. B. Lickey, J. T. Beck, S. B. Farmer, W. Liu, J. Miller, K. C. Siripun, C. T. Winder, E. E. Schilling, and R. L. Small. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany **92**:142-166.

Shaw, J., E. B. Lickey, E. E. Schilling, and R. L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. American Journal of Botany **94**:275-288.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayshida, T. Matsubayasha, N. Zaita, J. Chunwongse, J. Obokata, K. Yamaguchi-Shinozaki, C. Ohto, K. Torazawa, B. Y. Meng, M. Sugita, H. Deno, T. Kamogashira, K. Yamada, J. Kusuda, F. Takaiwa, A. Kata, N. Tohdoh, H. Shimada, and M. Sugiura. 1986. The complete nucleotide sequence of the tobacco chloroplast genome. Plant Molecular Biology Reporter **4**:111-148.

Slowinski, J. B., A. Knight, and A. P. Rooney. 1997. Inferring Species Trees from Gene Trees: A Phylogenetic Analysis of the Elapidae (Serpentes) Based on the Amino Acid Sequences of Venom Proteins. Molecular Phylogenetics and Evolution **8**:349-362.

Sohmer, S. 1988. The non-climbing species of the genus Psychotria (Rubiaceae) in New Guinea and the Bismarck Archipelago. Bishop Museum Bulletins in Botany. Bishop Museum Press, Honolulu, HI.

Sohmer, S. H. 1978. Morphological Variation and Its Taxonomic and Evolutionary Significance in the Hawaiian Psychotria (Rubiaceae). Brittonia **30**:256-264.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.

Sterck, F. J., R. A. Duursma, R. W. Pearcy, F. Valladares, M. Cieslak, and M. Weemstra. 2013. Plasticity influencing the light compensation point offsets the specialization for light niches across shrub species in a tropical forest understorey. Journal of Ecology **101**:971-980.

Sugiura, M. 1992. The chloroplast genome. Pages 149-168 *in* R. Schilperoort and L. Dure, editors. 10 Years Plant Molecular Biology. Springer Netherlands.

Taberlet, P., L. Gielly, G. Pautou, and J. Bouvet. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol Biol **17**:1105-1109.

Tan, M. A., J. A. Eusebio, and G. J. D. Alejandro. 2012. Chemotaxonomic implications of the absence of alkaloids in Psychotria gitingensis. Biochemical Systematics and Ecology **45**:20-22.

Tang, J., H. a. Xia, M. Cao, X. Zhang, W. Zeng, S. Hu, W. Tong, J. Wang, J. Wang, J. Yu, H. Yang, and L. Zhu. 2004. A Comparison of Rice Chloroplast Genomes. Plant Physiology **135**:412-420.

Taylor, C. M. 1996. Overview of the Psychotrieae (Rubiaceae) in the neotropics. Pages 261-270 *in* E. Robbrecht, C. Puff, and E. Smets, editors. Opera Botanica Belgica; Second International Rubiaceae Conference: Proceedings.

Taylor, H. R., and W. E. Harris. 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources **12**:377-388.

Taylor, J. S., and J. Raes. 2004. DUPLICATION AND DIVERGENCE: The Evolution of New Genes and Old Ideas. Annu Rev Genet **38**:615-643.

Thulin, M., J. Thiede, and S. Liede-Schumann. 2012. Phylogeny and taxonomy of Tribulocarpus (Aizoaceae): A paraphyletic species and an adaptive shift from zoochorous trample burrs to anemochorous nuts. Taxon **61**:55-66.

Toews, D. P. L., and A. Brelsford. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology **21**:3907-3930.

Valladares, F., S. J. Wright, E. Lasso, K. Kitajima, and R. W. Pearcy. 2000. Plastic phenotypic response to light of 16 congeneric shrubs from a Panamanian rainforest. Ecology **81**:1925-1936.

Van de Peer, Y., J. A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele. 2009. The flowering world: a tale of duplications. Trends in Plant Science **14**:680-688.

Vijayan, K., and C. H. Tsou. 2010. DNA barcoding in plants: taxonomy in a new perspective. Current Science **99**:1530-1541.

Wall, P. K., J. Leebens-Mack, K. F. Muller, D. Field, N. S. Altman, and C. W. dePamphilis. 2008. PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Res **36**:D970-976.

Willig, M. R., D. M. Kaufman, and R. D. Stevens. 2003. LATITUDINAL GRADIENTS OF BIODIVERSITY: Pattern, Process, Scale, and Synthesis. Annual Review of Ecology, Evolution, and Systematics **34**:273-309.

Wolfe, K. H., W. H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences **84**:9054-9058.

Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20**:3252-3255.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution **39**:306-314.

Yang, Z., and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences.

Zerbino, D. R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research **18**:821-829.