A PHYLOGENETIC ANALYSIS OF THE *Andira* CLADE BASED ON CHLOROPLAST AND ITS SEQUENCES

VICTORIA CABRERA

2015

Thesis submitted in partial fulfilment for the

MSc in the Biodiversity and Taxonomy of Plants

ABSTRACT

The *Andira* Clade is nested within the 50-kb inversion clade of papilionoid legumes. The lineage includes *Hymenolobium* Benth. and *Andira* Lam. com. cons., two predominantly Neotropical genera associated with tropical rain forests. In this study I generated phylogenies for the two genera based on nuclear ribosomal internal transcribed spacer sequences (ITS), and chloroplast sequences *mat*K and *trn*L. The chloroplast based phylogenies were unresolved for both genera but the ITS tree was more resolved and strongly supported. Phylogenetic signals from the chloroplast genes trees and the nuclear gene tree were not conflicting so the datasets were combined in a total evidence analysis. There is an ecological and geographical pattern shared in the phylogenies in *Hymenolobium* and *Andira*. Both genera have Amazonian lineages nested in a basally divergent position, they have apically nested species from the Mata Atlantica with low sequence divergence suggestive of recent radiation, and they have Central American species apically nested in the tree, suggesting three migrations from South to Central America.

TABLE OF CONTENT

List of Tables

List of Figures

1. INTRODUCTION

There is an evident misbalance in the distribution of species around the world. Some regions, like those located in the tropics, are remarkably more species rich than others. Even in tropical regions the biodiversity is heterogeneously distributed: the Neotropics alone contains around 90,000-110,00 species of seed plants, a greater number than in the tropical regions of Africa, Asia and Oceania combined (Antonelli and Sanmartín, 2011). The disparity in geographical patterns of biodiversity is more outstanding in some lineages more than others. For example, Bardon et al. (2012) reported that the Neotropics possesses around 80% of the estimated species of the family Chrysobalanaceae in. Which are the factors that explain the overwhelming biodiversity in this region? This is a question that naturalists and biogeographers have asked themselves for a long time without a clear answer (Hughes et al 2013).

Historically the discussion about the reasons of high diversity in the Neotropics had been focused on two main ideas: the museum hypothesis, which suggests a stable accumulation of diversity in an environment that had been exposed to relatively few perturbations, allowing the amassing of species because of low extinction (Stebbins 1974); and the cradle model in which the idea of a more dynamic environment leading to faster generation of species; the refuge hypothesis suggested by Haffer (1969) is an example of this approach. Currently the dichotomy that biodiversity in the Neotropics has originated by either recent or ancient diversification has been abandoned: the new insights in this matter propose that the actual biodiversity patterns are consequent to a series of scenarios with a combination of old and recent events as result of a mixture of rapid and slow diversification rates (Rull 2008, Antonelli and Sanmartín, 2011, Särkinen et al 2012). In addition to this, Hughes and colleagues (2013) suggest that a more biome-centered approach that investigates the patterns of diversification separately in different ecological settings may be more effective in the attempt to explain diversification in tropical South America.

When it comes to plant diversity, the Legume family immediately stands out; it is the third most species rich family within the angiosperms, with an estimate of 19,500 species; only surpassed by Orchidaceae and Asteraceae (Lewis *et al*., 2005; LPWG, 2013). The variety in species number and their outstanding geographical and ecological span (Lewis et al., 2005) make this family an excellent group to study evolution and diversification,

particularly in those lineages distributed in an area as evolutionary complex as the Neotropics.

In the traditional taxonomy of the Legume family there are three subfamilies: Caesalpinoideae, Mimosoideae and Papilionoideae. Papilionoid legumes are the largest and most widespread of the three, represented by an estimated 478 genera and 13,800 species (Lewis *et al.*, 2005). This diverse group has great importance given its value as a source of food (e.g. soy beans, chickpeas, domestic beans), timber, dyes and fuel, as well as having an essential role in medicinal and chemical industry. Moreover, from an ecological point of view the group also constitutes a central element in the floristic composition of several biomes, in addition of being environmentally relevant in biochemical processes at global scale involved in nitrogen cycling (Sprent, 2001). Papilionoideae has proven to be a monophyletic group (Doyle 1995; Kajita *et al.* 2001; Wojciechowski *et al.* 2004, Lavin *et al* 2005; Cardoso *et al*. 2012; LPGW, 2013); their grouping is supported by a set of features related to wood anatomy, initiation of flower structures and seed anatomy that are shared by all the species (Wojciechowski et al 2004). However the lineage is widely recognised by a more evident morphological character, which is the structure of the flower, described its own qualitative term of "papilionate flower". Characters of flower symmetry were used in the past to separate the group into "advanced" and "primitive" groups (Polhill, 1981), where a strong bilateral symmetry and fusion of the structures was considered a derived character in contrast with the usually undifferentiated and radially flowers of the basal group.

The classification of Papilionoideae *sensu* Polhill (1981) has been modified since based on phylogenetic studies using molecular evidence (Lavin *et al*., 2001; Pennington *et al.*, 2001; Wojciechowski *et al*., 2004). The phylogeny of the subfamily generated by Cardoso and colleagues (2012) using the chloroplast genes *mat*K and *trn*L, resulted in a monophyletic clade where most of the species are included in a group defined by a structural arrangement within the chloroplast genome, called the 50-kb inversion clade (Doyle *et al.* 1996; Cardoso *et al*., 2012). Nine groups are nested in the unresolved 50-kb inversion lineage; the *Andira* clade, a group consisting in two genera (*Andira* and *Hymenolobium*), is one of them.

Both genera used to be included in Dalbergieae (Bentham 1860), a tribe comprising tropical trees with strong fusion of the floral structures and indehiscent pods. Based upon molecular phylogenetics, the monophyly of the *Andira-Hymenolobium* clade, and its

separation from Dalbergieae has been strongly supported in several works (Lavin *et al.*, 2001; Penninton *et al.*, 2001; Cardoso *et al.* 2012). The results obtained by Lavin and colleagues (2001) implied that *Hymenolobium, Andira, Vatairea and Vataireopsis* did not belong to the dalbergioid clade within which most of Dalbergieae reside; this separation of the four genera from the others included in the tribe is supported by disparities in the characteristics of the wood, since they lack the storied structure and uniseriate rays that the rest of the dalbergioids have (Lavin *et al.* 2001).

The *Andira* clade consist of predominantly Neotropical species (only *A. inermis* reaches Africa), generally associated with tropical rain forests. The close relationship of *Andira* and *Hymenolobium* has been suggested from phylogenetic analysis of cpDNA restriction sites and nrDNA ITS sequences (Pennington 1995, 2003; Simon *et al.* 2009), and is reinforced by 1) a distinguishing nodule structure that is uncommon in other papilionoid legumes, 2) their shared inflorescence features of densely flowered terminal panicles, 3) the presence of the distinctive papilionate flower with the petals differentiated as keel petals, wing petals and standard petal, and 4) the clustered disposition of the leaves (Pennington, 2003). The two genera differ greatly in fruit type and mode of germination: *Hymenolobium* fruits are samaras and its germination is phanerohypogeal, whereas the fruits in *Andira* are drupes, and the germination is cryptohypogeal (Pennington, 1995, 1996).

1.1. The genus *Hymenolobium* Benth.

Hymenolobium is a genus originally described by Bentham (1876), which is comprises 17 species (Mattos, 1979; Lima, 1982), all restricted to the tropical regions in South America with the exception of one species, *H. mesoamericanum* that is distributed in gallery forests in Central America (Lima, 1988). The genus is characterized as an Amazonian taxon since most of the species are big emergent trees found in the rain forests of the Amazon, particularly from Brazil (Mattos, 1979). The dimensions of *Hymenolobium* trees and their aesthetic beauty during flowering periods have motivated their use as ornamentals (Ducke, 1936). In addition, many species of *Hymenolobium,* or "angelim", as is commonly known in Brazil, possess a hard wood that has a high commercial value, used industrially for naval and civil construction (Ducke, 1936; Mattos, 1979). Other usages reported for the species of the genus include the consumption of the boiled seeds of *H. heterocarpum* as food (Ducke, 1936) and the use of *H. excelsum* for medicinal purposes (Silva *et al.*, 1977).

The species from the genus are medium to big trees, with compound leaves and oblong, obovate-oblong or lanceolate leaflets. The calyx is campanulate with truncate apex, 5-lobed. Flowers are purple to pink. The corolla has an orbicular standard petal that recurves above the calyx when the flower opens; wing petals are obliquely oblong and straight; keel petals are free. The 10 stamens are united by their filament forming one group. Ovary is shortly stalked, linear to linear-lanceolate. Fruits are samaras, oblong to oblong-lanceolate, compressed, membranaceous, reticulately veined with one or two prominent nerves (Mattos, 1979; Lima 1988).

Several authors (Ducke, 1936; Mattos, 1979; de Lima, 1982) have described the systematic study of *Hymenolobium* as a challenging work for different reasons. Firstly, there is the problem related the collection of specimens. Because many species are from the Amazon, accessibility may be difficult and even when a specimen has been found, the great heights reached by *Hymenolobium* trees makes the collection of herbarium specimen a complex task. Secondly, the majority of the species of *Hymenolobium* lose their foliage with the advent of the flowers and the new leaves start to grow just after the fall of the ripen fruits (Ducke, 1936); as a consequence the possibilities of collecting an "ideal" specimen (i.e. including both vegetative and reproductive structures) are low, which is why most of the material of *Hymenolobium* available in herbaria is vegetative (Lima, 1982). Lastly, the fact that morphological characters of the leaves (that are the most frequent features used for the determination) may vary depending on the age of the specimen and the environmental stress to which it has been exposed represents an additional difficulty in the study of the group.

The study I present here includes sampling of accessions of virtually all species of *Hymenolobium* for the chloroplast regions *mat*K and *trn*L and the nuclear ribosomal internal transcribed spacer (ITS). This is the first phylogeny to sample comprehensively across the genus, greatly increasing taxon sampling compared to previous studies (Cardoso *et al.* 2012, 2013), which were also based only on chloroplast sequence data.

1.2. The genus *Andira* Lam. *nom. cons.*

In a monograph for the genus, Pennington (2003) explained the problems faced tracing the first valid publication of the name *Andira*, resulting finally in the proposal and acceptance of Lamark as an author of the name (Pennington, 2002). The genus comprises 29

species distributed in tropical America, with one of them, *A. inermis* also occurring in Africa. As for *Hymenolobium*, *Andira* species (also vernacularly named "angelim") have been commercialized because of the quality of the wood (Ferreira *et al.* 2004). Pennington (2003) mentioned that other species of *Andira* have been used for medicinal purposes related to the anthelmintic properties of the seeds (*A. fraxinifolia*) and bark (*A. inermis* and *A. surinamensis*).

Species of *Andira* are generally medium to large trees (rarely geoxylic suffrutices), with spirally disposed, compound, imparipinnate, up to 9-jugate leaves. The calyx is shallowly to deeply 5-lobed, with the upper lobes differentiated from the lower ones. The flowers are purple, pink or white; with wing petals sculptured or not, keel petals are overlapping but not fused. There are 10 stamens 10, with the vexillary stamen free. The ovary is distinctly stipitate. And the fruits are a globose to elongated drupe (Pennington, 2003).

Previous systematic works of the genus include studies focused at different taxonomic levels that contrast phylogenetic relationships of *Andira* based on morphological and molecular data (e.g. Pennington 1995, 1996, 2003; Pennington *et al.* 2001; Lavin *et al.*, 2001; Simon *et al.*, 2009; Cardoso *et al.* 2012, 2013). A dated phylogeny of the genus based on ITS sequences and the ecological factors related to its diversification in the savannas of central Brazil is also reported in the study of Simon and colleagues (2009). The problem with this study is that it is based upon a single nuclear locus. Therefore a critical goal of my study involves estimation of *Andira* phylogeny using an independent estimate from cpDNA (matK and trnL), and exploration of the possibility of a phylogeny using total evidence from a combined nuclear and chloroplast dataset.

1.3. Using nuclear and chloroplast regions to infer phylogenetic relationships

1.3.1. Internal Transcribed Spacer (ITS) regions

The Internal Transcribed Spacer (ITS) regions sequences are among the most frequent data used to infer evolutionary relationships in plants at generic and infrageneric level (Baldwin et al. 1995; Soltis and Soltis, 1998; Hughes at al. 2006). The two spacers of this region (ITS-1 and ITS-2) are part of the nuclear ribosomal DNA transcript, but are not incorporated into ribosomes, and seem to be involved in the maturation of nuclear rRNA. The popularity of this region to reconstruct phylogenies in angiosperm relates to the fact that 1) it

is relatively small (approximately 600 bp) and is flanked by conserved sequences of the 18S-5.8S-26S cistron, which facilitates its amplification using universal primers, 2) in a majority of cases variation in ITS sequences is attributed to point mutations, so the alignment of its sequences is usually unambiguous, and 3) the rapid nucleotide substitution rates of the region makes it a valuable to resolve relationships at lower taxonomic levels (Baldwin et al., 1995; Alvarez and Wendel, 2003). However, the nrDNA where the ITS region is located has a complex organization in multiple arrays (with hundreds to thousands of transcription units present within an array), that allow the existence of polymorphism, pseudogenes, and paralog sequences; which complicates the understanding of the evolution of the region, and may lead incorrect phylogenetic estimates (Bailey *et al* 2003; Feliner and Rosello, 2007)

1.3.2. The chloroplast regions *mat*K and *trn*L

Chloroplast DNA sequences have been extensively used to reconstruct phylogenies in the past (Alvarez and Wendel, 2003; Shaw *et al.* 2005; Hughes *et* al., 2006). Within the chloroplast genome, two regions have that are used in this thesis have been commonly employed in molecular systematic studies: *mat*K and *trn*L. The protein-coding region *matK* consists of approximately 1500 bp, is positioned within an intron of 2600bp flanked by the 5' and 3' exons of the transfer RNA gene for lysine *trn*K, and encodes *mat*K, a maturase involved in splicing of type II introns (Wolfe *et al.* 1992). The *mat*K region is among the coding regions with the fastest nucleotide substitution rate within the chloroplast (Wolfe 1991), which is evidenced by the high resolution provided in the reconstruction of phylogenies obtained based on this region (Johnson and Soltis, 1995; Soltis and Soltis 1998). In contrast to *mat*K, the *trn*L regions are part of the non coding sequences of the chloroplast genome; they consist of the *trn*L (UAA) intron and the intergenic spacer between the *trn*L (UAA) 3' exon and the *trn*F (GAA) gene, also known as the *trn*L-*trn*F spacer (Taberlet *et al.* 1991, Gielly and Taberlet, 1996). These relatively small regions (approximately 350-600 bp the intron and 120-350bp the spacer) are easy to amplify and sequence using universal primers (Taberlet *et al* 1991). Shaw and colleagues (2005) give a detailed review of systematic studies in angiosperms at different taxonomic levels including analysis for the *trn*L regions. Both *mat*K and *trn*L regions have worked well resolving phylogenetic relationships in legumes at different taxonomical levels (e.g. Hu *et al.*, 2000; Lavin et al.,

2001, 2005; Pennington *et al.* 2001, Steele and Wojciechoski, 2003; Wojciechowski *et al.* 2004; Cardoso *et al.* 2012; LPWG, 2013).

There are different factors that can potentially contribute to phylogenetic inaccuracy in analyses based on cpDNA; introgression as consequence of hybridization is particularly relevant in studies at infrageneric level. In other words, the transfer of chloroplast genes from one species to another as a result of hybridization and subsequent introgression leads to a potentially erroneous estimation of phylogenetic relationships (Rieseberg and Brunsfeld 1992, Soltis and Kuzoff, 1995). Such problems can be highlighted by sampling numerous accessions per species and contrasting a chloroplast-based phylogenetic hypothesis with one derived from a nuclear locus, both of which strategies are employed in this thesis.

1.4. Potential conflict between nuclear and chloroplast phylogenetic inferences

One fundamental issue in the reconstruction of phylogenies based on molecular data is the potential disparity between the gene tree and the species tree (Pamilo and Nei 1988; Doyle 1992; Wendel and Doyle 1998). That is, each phylogenetic reconstruction based on a particular gene (gene tree) represents a hypothesis of the real evolutionary relationships of the group (species tree); species trees are inferred form gene trees but that does not imply that they are synonymous (Doyle 1992). Gene trees do not always tell the same evolutionary history; some of the factors responsible for phylogenetic incongruence include: 1) hybridisation and introgression, (Wendel et al 1991; Hardig et al 2000), 2) lineage sorting (Pamilo and Nei, 1988; Doyle, 1992), 3) gene duplication (Bailey et al 2003) 3) disparate diversification rates and 4) taxon sampling (Wendel and Doyle, 1998). All of these factors suggest that the study of single gene phylogenies could be misleading, which is why it is imperative to estimate phylogenetic relationships combining the information provided by different genes from independent genomes, i.e. contrasting nuclear and chloroplast regions (Pamilo and Nei, 1988).

1.5. Aims of the study

The purpose of this thesis is to generate an entirely new phylogeny for *Hymenolobium* based upon nuclear (ITS) and chloroplast (trnL, matK) markers. In addition, the research presented here provides the first chloroplast DNA-sequence based phylogeny for *Andira*.

Specific ojectives are:

(i) For both *Andira* and *Hymenolobium*, to assess congruence between ITS and chloroplast phylogenies. In the case of Andira, this allow assessment of the accuracy of the published phylogeny (Simon et al., 2009), which is based only upon ITS.

(ii) If the chloroplast and ITS phylogenies are congruent, to generate a phylogeny for both genera using total evidence.

(iii) To use the phylogenies to study character evolution and biogeography in *Andira* and *Hymenolobium*.

2. MATERIALS AND METHODS

2.1. Taxon Sampling

The phylogenies were based on the analysis of three regions: the internal transcribed spacer (ITS) regions of 18S to 26S nuclear ribosomal DNA, the plastid gene *matK*, and the *trn*L choroplast intron and spacer. The samples assembled for ITS included 17 accessions representing 10 species of *Hymenolobium* and 48 accessions represented by 25 species of *Andira*; for *matK* 25 accessions were gathered from 13 species of *Hymenolobium* and 29 accessions from 21 species of *Andira;* whereas the matrix for trnL compiled 21 sequences from 12 species of *Hymenolobium* and 21 species of *Andira*. Seven outgroup accessions accounting for four species (one of *Sweetia* and three of *Vatairea*) were used for the analysis. The outgroup selection was drawn from the unresolved 50-kb inversion clade in the phylogeny of papilionoid legumes of Cardoso *et al*. (2012). In total, 18 ITS, 23 *mat*K and 23 *trn*L sequences were generated during de preparation of this thesis.

Differences in taxon sampling amongst the regions were due to sample accessibility and laboratory work issues related to the quality of sequences. Regarding the access to material, *H. elatum* and *H. velutinum* were not available to process in the laboratory since

there were no samples (neither from herbarium nor previous DNA extractions) deposited in RBGE; *H. velutinum* was a particular case since the only material known for the taxon is the species holotype (R. T. Pennington pers. comm.). No DNA material was accessible for *H. excelsum* either; the only data for this species was a *matK* sequence that had already been generated by Alexandra Clark. Just one accession of *H. modestum* was processed in the laboratory but poor quality sequences, for both ITS and *matK* were obtained*;* however, *matK* data generated previously was used in the analyses. *H. petraeum* is also missing from the ITS data due to a failed attempt at sequencing DNA for the region. The samples used for *Andira* were chosen from a previous phylogenetic and biogeographic study of the genus based on ITS sequences (Skema, 2003; Simon et al 2009) to allow, ideally, the comparison of the ITS tree topology with the phylogeny obtained from the chloroplast loci *matK* and *trnL*, which gives an additional estimate of relationships. In the case of *A. praecox, A. tervequinata, A. trifoliolata* and *A. unifoliolata* ITS sequences from Skema's (2003) study were the only data at hand, so these species are not represented in the phylogenies based on chloroplast DNA.

2.2. Molecular data

The ITS region has been widely used to build phylogenies of species since it is easy to amplify in the laboratory using universal primers, and it often provides sufficient variation at lower taxonomic levels to resolve species-level relationships (Baldwin et al., 1995; Soltis and Soltis, 1998; Hughes et al., 2006). In this study I generated *Hymenolobium* ITS sequences that complement those already available for *Andira*, resulting in a broader molecular data set for the *Andira-Hymenolobium* clade within papilionoids. The idea of combining a nuclear ribosomal locus with chloroplast DNA regions was to determine if it gave a better resolved species-tree (Nixon and Carpenter, 1996). The plastid genes *matK* and *trnL* were chosen in this study since they have proven to be sufficiently variable in phylogenetic studies at intergeneric and interspecific level (Soltis and Soltis, 1998). Studies such as those of Whitten et al. (2000), Pennington et al. (2001), Chen et al. (2005), and Torke and Schaal (2008) highlight the utility of *trnL* at different taxonomic levels within angiosperms, while the chloroplast locus *matK*, in particular, has been useful for resolving lineages within legumes (Wojciechowski et al., 2004; Lavin et al 2005, Cardoso et al 2012, Cardoso et al. 2013). Also, *matK* has now been so widely sequenced in legumes (LPWG, 2013), that sequences generated in this study can contribute to wider datasets being developed for the entire family.

Genomic DNA was isolated from silica-dried leaves and extracted using DNeasy Plant Minikit (Qiagen, Santa Clara, California, USA) for some *Hymenolobium* specimens. The rest of the *Hymenolobium* material was available as total DNA that was previously extracted from silica-dried and herbarium samples following the CTAB method (Doyle and Doyle, 1987). The existing DNA aliquots of *Andira* were extracted between 1991, 1995 (Pennington, 1995), and 2003 (Simon *et al*. 2009), from fresh material, herbarium specimens, and leaves dried in silica, in anhydrous calcium sulphate or over a drying frame (Pennington, 1995). The extraction method followed for those samples was a modification of the CTAB procedure (Doyle and Doyle, 1990), after which the aliquots were stored at -80C in the interim.

The polymerase chain reactions (PCR) involved Bioline Taq and reagents (Bioline, London NW2, UK). The primers used for the amplifications and sequencing of the ITS region were forward 5P and reverse 8P as described by Moller and Cronk (1997). The *mat*K gene was amplified using primers trnK685F and trnK2R*, then it was sequenced using forward primers trnK685F and matK4La, and the reverse primers matK1932R and trnK2R* according to Wojciechowski and colleagues (2004). Primers "c" and "f" of Taberlet et al. (1991) were used to amplify the *trn*L-F intron and spacer. Primer sequences of the regions amplified are shown in detail in Table 1.; details of composition of the PCR reaction are shown in Table 2.

Table 1. Names and sequences of primers used for amplification and sequencing of ITS, *mat*K and *trn*L.

Region	Primer name	Sequence
ITS	5 P	GGAAGGAGAAGTCGTAACAAG
ITS	8 P	CACGCTTCTCCAGACTACA
matK	trnK685F	GTATCGCACTATGTATCATTTGA
matK	trnK2R*	ACACGGCTTTCCCTATGTCTAC
matK	4La	CCTTCGATACTGGGTGAAAGAT
matK	1932R	CCAGACCGGCTTACTAATGGG
trnL	c	CGAAATCGGTAGACGCTACG
trnL	f	ATTTGAACTGGTGACACGAG

The quality of the PCR amplification product of all the samples was checked running electrophoresis on a 1% agarose gel, and visualized using GeneSys Image Acquisition Software. Amplifications were repeated for those samples that did not have signal for the band of the double stranded DNA. Although most of the extracted DNA used in this study had been stored frozen over a decade, degradation of the material seems to have been minor since there majority of the accessions were processed successfully.

All amplifications were performed in one reaction. The ITS PCR reaction was that followed by Skema (2003) to sequence ITS region of *Andira*, consisting of a three minutes denaturation step at 94 °C; followed by 30 cycles of 1 minute at 94 °C denaturation, annealing at 55 °C for 1 minute and extension at 72 °C for 1.5 minutes; finished by 5 minutes at 72 °C extension. Cycle sequencing conditions for *mat*K were 3 minutes at 95 °C preceding 35 cycles of 30 seconds at 95 °C, 45 seconds at 50 °C and 1 minute at 72 °C, with a final extension of 72 °C for 7 minutes. The PCR reaction for *trn*L was performed for 3 minutes at 94 °C, followed by 30 cycles of 94 °C for 45 seconds, 55 °C for 45 seconds and 72 °C for 2 minutes, with a final extension of 7 minutes at 72 °C. PCR product of all regions was purified using the enzymatic PCR clean-up solution ExoSAP-IT (Exonuclease I and Shrimp Alkaline Phosphatase) according to the manufacturer's protocol.

Table 2. Components and proportions of reagents used in Polymerase chain reactions mix to amplify ITS and *matK* regions.

Reagents	ul added per sample
Deoxynucleotide solution mix (dNTPs) 2mM	2
10x Buffer	2
Magnesium chloride ($MgCl_2$) 2mM	1
Primer A	0.65
Primer B	0.65
Taq Polymerase	0.1
CES buffer	4
Template DNA	0.8
Distilled water	8.8

2.3. Sequence editing and alignment

Strands of forward and reverse sequences were grouped in contigs and edited using Sequencher 5.1 software (Gene Codes, Ann Arbor, Michigan, USA). Sequences of the loci were aligned by eye in Mesquite 2.75 and assembled in separate matrices that included outgroup sequences. The latter consisted of sequences taken from GenBank from the study of Cardoso and colleagues (2013); other sequences that had been previously generated and were incorporated in the analysis were those of ITS region for *Andira* (Simon et al., 2009) and some *matK* sequences generated by Alexandra Clark. Please refer to Appendix I for further information related to the accessions used in this study.

2.4. Phylogenetic analysis

Parsimony analysis of individual and combined datasets were carried out in PAUP* version 4.0a145 (Swofford, 2002). Heuristic searches for maximum parsimony analysis were performed in two stages: the first one with default settings modifying the random stepwise addition to 10,000 replicates, no swapping and steepest descent on; the second search was run using all trees in memory obtained from the first search, selecting tree bisection-reconnection as branch swapping algorithm (TBR), saving multiple trees, keeping the option of swapping on all optimal starting trees (steepest descent) off, and saving no more than 100 trees of score greater than or equal to 1 in each replicate to minimise computation time. All parsimonious searches were set to collapse branches, which had a minimum length of zero. The support for the phylogenies was estimated by calculating bootstrap values from 10,000 replicates, obtaining the starting tree via stepwise addition, with the steepest descent option not in effect, and saving only one tree per replicate. Bootstrap support is defined as: 50-70% weak, 71-85% moderate and 86-100% strong. For all analyses gaps were treated as missing, all characters were considered unordered and were equally weighted.

Chloroplast genes presented one region of ambiguous alignment of 19 and 16 base pairs for *mat*K and *trn*L, respectively. These fragments of sequences consisted in a single nucleotide repetition of variable length among the taxa sampled; allowing different alignment combinations, therefore the assessment of homologies in this region was questionable. Exploratory parsimony analyses were performed in two sets of data: one including and other excluding the regions of uncertainty in the alignment, in order to evaluate the potential effect of these characters in the phylogenetic relationships of the species before carrying on with further analyses.

Data from different sources (e.g. dried material, DNA extractions, GenBank sequences) was compiled in this study to provide a wide sampling of material available for the three loci selected to compare between *Hymenolobium* and *Andira*. When studying the data, it was evident that there were cases of accessions with too many missing data, poor quality sequences, or uncertainties in the determinations of the samples. Preliminary analyses considering the inclusion or exclusion of those accessions were performed in an attempt to keep a conservative approach using the most reliable data in the analyses without compromising too much the number of characters or taxa.

Analyses were done individually for each of the loci as well as combining the data in three sets, which contained sequences from: 1) only the chloroplast loci (*mat*K and *trn*L), 2) the three genes including all accessions compiled (ITS, *mat*K and *trn*L), and 3) the three genes including only those accessions that were sequenced for all the regions. The incongruence Length Difference (ILD) test (Mickevich and Farris, 1981), or homogeneity partition test, was run for all combined data sets to assess the potential incongruence of the data, and whether or not they should be combined to carry on with further analyses.

Bayesian analyses of individual and combined data sets were completed on MrBayes v 3.2.5. Four Markov chains started from a random tree and ran for up to 10 million generations, with a sampling frequency of 1000; discarding the initial 10% of generations as burn-in samples. The ITS region was divided in two partitions, spacers and 5.8S, following the region boundaries suggested by Yokota and colleagues (1989) for *Vicia faba*. The *mat*K intron sequence boundaries were identified as those suggested for *Andira galeottiana* (GenBank AF142681.1) by Hu et al. (2000), while *trn*L was partitioned into intron and spacer, setting the boundaries according to the *trn*L Genbank sequence for Tobacco Z00044 from Shinozaki et al. (1986). The partition of the data into separate regions (e.g. spacers, intron) allowed estimating the model of nucleotide substitution that best fitted the data more accurately. For further information about the models that best fitted each of the partitions of the regions please refer to table X3. The Bayesian majority-rule consensus was visualised and edited using Figtree v.. Posterior probability support is defined as: 50-80% weak, 81-95% moderate and 96-100% strong.

3. RESULTS

3.1. Phylogeny based on ITS

The ITS region did not present any fragment of sequence that would be considered ambiguous to align; the only characters excluded for the analyses were those belonging to the 18S and the 26S genes because of differences in the length of the sequences. *Andira galeottiana* LR *sn.* was removed from the original matrix because it lacked data for the other loci. The strict consensus trees including and removing this accession had identical topologies. Additionally, the overall bootstrap values for the supported nodes of *Andira* increased 1-3%. In the case of *Hymenolobium,* the only change produced from the extraction of the sample was the loss of bootstrap support of *H. heterocarpum* as sister taxon of the clade consisting of *H. sp, H. flavum, H. petraeum, H. alagoanum, H. grazielanum, H. heringerianum* and *H. janeirense.* However, the previous support for the node before removing *Andira galeottiana* was only 50%.

The phylogeny based on ITS was the most resolved of all the trees obtained in this study. A summary of the characteristics of the ITS matrix and the statistics obtained from the parsimony analysis are summarized in Table 3. The monophyly of the genera has a moderate bootstrap support (*Hymenolibum:* BS 78%; *Andira*: 81%) (Figure1). Two lineages could be identified within *Hymenolobium*: a well-supported clade A, which occupies a basally divergent position in the phylogeny of the genus, including *H. pulcherrimum, H. mesoamericanum, H. nitidum, H. sericeum* and *H. sp* (DMN 1965); and clade B, weakly supported by a bootstrap value of 55%, containing *H. heterocarpum* and the unidentified *H. sp* (TDP 16995) which are sister to the rest of the *Hymenolobium* species. The latter are grouped in three lineages forming a polytomy of a highly supported clade of *H. flavum* and *H. petraeum,* a clade composed of all the *H. alagoanum* accessions, and an unresolved group clustering *H. grazielanum, H. heringerianum* and *H. janeirense*.

Relationships in the clade of *Andira* remain without resolution at the base of the tree, where *A. unifoliolata, A. parviflora* and *A. micrantha* are positioned, together with a 100% supported clade of *A. cordata* and *A. cujabensis*. Accessions of *A. macrothyrsa* are all clustered together, as well as *A. grandistipula;* both situated in the unresolved base of the *Andira* clade. Two major lineages are evident within the genus: clade C, containing a subclade of 1) a highly supported polytomy with *A. praecox, A. taurotesticulata, A. tervequinata* and *A. trifoliolata,* and 2) a 93% supported group with the paraphyletic *A.*

inermis together with *A. multistipula* and *A. jaliscensis*; while clade D is formed by *A. marauensis* diverging from the rest of the *Andira* species. These taxa are organised in 1) a more closely related group of *A. galeottiana, A. humilis, A. macrothyrsa* (TDP 13550), *A. surinamensis* and *A. vermifuga*, and 2) an unresolved clade of *A. anthelmia, A. fraxinifolia, A. humilis, A. legalis, A. ormosioides* and *A. nitida*. One accession of *A. nitida* (AMC 3309) is situated outside the former clade, together with *A. carvalhoi*.

Table 3. Statistics from the overall parsimony analysis for each of the regions and the combined data sets: cpDNA (*mat*K + *trn*L), all accessions combined (ITS + *mat*K + *trn*L), and shared accessions (only those accessions that were sequenced for the three regions).

	ITS	*mat*K	*trn*L	cpDNA	All accessions	Shared accessions
Number of taxa	71	55	47	55	82	32
Number of characters	806	2812	1065	3877	4683	4683
Eliminated characters	144	20	27	47	191	191
Variable characters	201	274	56	329	530	429
Parsimony informative characters	186	210	49	258	444	364
Number of most parsimonious trees	49	13	22	4	6	138
Length of most parsimonious tree	413	287	56	344	778	602
Consistency Index (CI)	0.6247	0.9791	1	0.9767	0.7635	0.7973
Retention Index (RI)	0.9045	0.9935	1	0.9927	0.9324	0.9246
P value for partition-homogeneity test				0.1700	0.3300	0.9100

Bayesian analysis resulted in a phylogenetic hypothesis almost identical to that of the parsimony analysis (Figure 2). The outcome of *H. mesoamericanum* as sister species of *H. sericeum* is one of the differences between the Bayesian majority-rule consensus tree and the strict consensus tree from the parsimony analysis. Regarding *Andira*, Bayesian inference strongly supports the position of *A. micrantha* as sister species of the *A. cordata* and *A. cujabensis* clade; even when this relationship between the three species is not evident in the strict consensus tree obtained from parsimony, the bootstrap consensus tree support this clade as well, although poorly. Finally, a more resolved clade D1 is present in the phylogeny generated by Bayesian methods, where *A. surinamensis* is placed in a basal position with *A. macrothyrsa* (TDP 13550) diverging next, followed by an unresolved relationship of *A. galeottiana* and *A. vermifuga*, with the apically placed accessions of *A. humilis.*

3.2. Phylogenetic reconstruction based on *mat*K

The inclusion or exclusion of the ambiguous alignment regions resulted in trees of uniform topologies and similar bootstrap support. The major difference that resulted from the exclusion of the ambiguous region was a low supported clade of *H. excelsum, H. flavum* and *H. petraeum* (in contrast with the broad *Hymenolobium* polytomy obtained when the region was included for analysis), and the lack of support of *A. taurotesticulata* as sister species of the group containing *A. inermis, A. jaliscensis* and *A. multistipula*. The strict consensus tree for the data set that excluded the ambiguous region turned out to be more resolved for both genera. Considering both data sets had clades with similar support and that there was a stronger phylogenetic signal in the strict consensus tree, the data set excluding the ambiguous alignment was selected for further analysis.

Although the strict consensus tree excluding those accessions with too many missing data or uncertain identifications resulted in the collapse of many branches for both genera (Figure 3), bootstrap support values were higher for all nodes, except for the one supporting clade C (*A. inermis, A. jaliscensis, A. multistipula* and *A. taurotesticulata*). However, even with a lower bootstrap value, the C clade was still highly supported (86%) when the conflicting taxa were removed. The dataset chosen to perform the rest of the analyses was the one excluding those accessions.

The phylogeny obtained with *mat*K also supports with a value of 100% (BS and pp) the monophyly of the *Hymenolobium* - *Andira* clade, as well as the monophyly of each genus separately. In both genera the trees obtained for the *mat*K gene by parsimony analysis were less resolved than those inferred based on ITS. In the case of *Hymenolobium* most branches collapsed; Clade A (*H. pulcherrimum, H. nitidum, H. sp* (DMN 1965), *H. sericeum* and *H. mesoamericanum*) is no longer present, and the only lineages that remained supported are the group of *H. excelsum, H. flavum* and *H. petraeum* (with a support of 51%), and the group of *H. alagoanum* (RTP 224) – *H. sp.* (TDP 16995). The genus *Andira* also comes out as a big polytomy with only two lineages retained: an unresolved clade C with 86% of support, and a less supported group with *A. cujabensis* (DMN 1889), *A. galeottiana, A. ormosioides* and *A. vermifuga.*

The Bayesian majority-rule tree (Figure 4) was more resolved than the strict consensus tree from the parsimony analysis. The major difference between the topologies was the support of the group *H. alagoanum* and *H. janeirense* var. *janeirense*, and the

resolution of two clades within *Andira*: the first one comprising the "basal diverging species" (comparing with the ITS results) of *A. grandistipula, A. macrothyrsa, A. micrantha, A. cordata, A. cujabensis* (RTP 503) and *A. parviflora*; and the second one including some of the species from Clade D (*A. marauensis, A. cujabensis* DMN 1889, *A. galeottiana, A. ormosioides* and *A. vermifuga*). Both lineages were weakly supported by low posterior probabilities values in the Bayesian analysis.

3.3. Phylogenetic inferences based on *trn*L

The monophyly of the *Hymenolobium – Andira* clade is supported by 100% in the *trn*L phylogeny; nevertheless, there is no support for the monophyly of *Andira*. The strict consensus tree including and excluding the region with uncertain alignment was identical. Bootstrap consensus trees removing this region was better supported and resolved one lineage of *Andira* with some species of Clade D (*A. galeottiana*, *A. ormosioides, A. vermifuga*). Just as for the *mat*K locus, the data set selected to continue with the rest of analyses was the one excluding the region of ambiguity due to the more resolved and better supported relationships of the taxa.

The parsimony analysis removing the "conflictive" accessions resulted in strict consensus tree with the same topology as the tree obtained when those taxa were included. Excluding those accessions did not have a major impact on the bootstrap support values for the nodes. In order to be as consistent as possible regarding the accessions included for each region, the dataset chosen to continue the rest of the analyses was the one without the "conflictive" taxa.

The phylogenetic hypothesis obtained based on *trn*L sequences was the least resolved compared to that of the other two regions considered in this study. In the parsimony analysis *Hymenolobium* is weakly supported as monophyletic (BS of 63%); within the genus the relationships of the species are unresolved, and there is only one clade, weakly supported, grouping the accessions of *H. alagoanum* and *H. janeirense* var. *stipulatum* (Figure 5). The relationships between the species of *Andira*, and between this genus and *Hymenolobium* are completely unresolved. The only lineage that is supported clusters *A. galeottiana, A. ormosioides* and *A. vermifuga* with a bootstrap support of 61%.

The topologies of the Bayesian majority-rule consensus tree and the strict consensus tree from the parsimony analysis differed in that the Bayesian inference resolved one more

clade for *Hymenolobium*, and three groups in the case of *Andira* (Figure 6). In the case of *Hymenolobium*, the group of *H. alagoanum* and *H. janeirense* var. *stipulatum* was supported by a posterior probability of 1; the clades supported for *Andira* consisted of the strongly supported group of *A. galeottiana, A. ormosioides* and *A. vermifuga,* and the two less supported lineages of 1) *A. cordata* and *A. fraxinifolia* with a posterior probability of 0.57, and 2) *A. inermis MC 3579, A. jaliscensis, A. multistipula* and *A. taurotesticulata,* supported by a probability of 0.65.

3.4. Combined datasets

All of the combined matrices had insignificant p-values for the Incongruence Length Difference (ILD) homogeneity test: 0.17 for the cpDNA dataset, 0.33 for the combined matrix of all the available sequences, and 0.91 for the "shared" dataset that included only those accessions that were sequenced for the three loci. If the ILD test is non significant ($p>0.05$) it means there is no substantial incongruence between the partitions and the data could be combined without introducing conflict into the analysis. The monophyly of *Andira* and *Hymenolobium* was supported in all the analyses performed with combined data sets.

The results of the ILD test are supported by visual inspection of the phylogenies, which reveal no differences in topology that are strongly supported by parimony bootstrap of Bayesian posterior probabilities. Differences in topology largely reflect lack of resolution in the trnL and matK trees.

3.4.1 Plastid genes: *mat*K and *trn*L

The parsimony analysis of the combined chloroplast loci did not have a major impact in the resolution of the phylogenetic relationships within each genus. The only difference in the topology, compared to the more resolved obtained by *mat*K, was one extra clade grouping *A. grandistipula* and *A. macrothyrsa*. The analysis of combined cpDNA did not contributed to the resolution of the *Andira* clade, thus the trees were excluded from the text.

3.4.2. Accessions of all regions combined

The strict consensus tree from the parsimony analysis resembled, in resolution level, the one obtained based on the ITS region. In the case of *Hymenolobium* there is a basally divergent group "A" (including the clade of *H. nitidum, H. sp.* DMN 1965, *H. mesoamericanum* and *H. sericeum*) next to *H. heterocarpum* LPQ 13905, *H. pulcherrimum* WR 11181 and *H. sp.* WR 11183 forming a polytomy. Clade "B", the second major lineage, includes *H. heterocarpum* (LPQ 13899) and *H. sp.* 16995 separated from a cluster that compiles the subgroups: 1) *H. excelsum, H. flavum,* and *H. petraeum, 2)* H. *heringerianum HCL 7460 - H. modestum* and *H. alagoanum – H. janeirense* var. *stipulatum*, and 3) *H. heterocarpum* LPQ 13899, *H. sp.* TDP 16995, *H. grazielanum*, *H. heringerianum* (HCL 3266, AEHS 4265) and *H. janeirense* var. *janeirense.*

The strict consensus tree from the parsimony analysis, with a basally divergent clade of *A. cordata, A. cujabensis* RTP 503, *A. micrantha* and *A. parviflora* KGD 6920 is resolved, though with weak support (Figure 7). The clade including *A. grandistipula, A. unifoliolata* and *A. macrothyrsa* has weak support and is sister to Clade "C". The remaining major group within *Andira*, Clade "D", has *A. marauensis* as sister to the rest of the taxa conforming the group, then *A. nitida* and *A. carvalhoi* are placed in a polytomy with the group "D2" (*A. anthelmia, A. fraxinifolia, A. humilis, A. legalis, A. ormosioides* and a paraphyletic *A. nitida*). The remaining clade, "D1", is composed of *A. surinamensis* in a basally divergent position, then *A. legalis* HCL 7460, *A. macrothyrsa* TDP 13550, all accessions of *A. humilis*, and the cluster of *A. cujabensis* DMN 1889, *A. galeottiana* and *A. vermifuga.*

The phylogenetic hypothesis derived from Bayesian analysis differed from the one obtained by parsimony in resolving the relationship between the basally divergent group of *A. micrantha, A. cordata* and *A. cujabensis* (Figure 8). In the Bayesian tree there is also a shift of *A. micrantha* and *A. parviflora* from the basal position of the *Andira* clade to a more derived position nested within the "D" group, supported with a weak value of 50%. *Andira legalis* HCL 7480 is also moved from the "D1" subgroup to a basal position in relation to the "D2" clade.

3.4.3 Combined data of only accessions sequenced for the three regions

This dataset was the most conservative of all, in the sense that it was restricted only to 32 taxa; each of the accessions were sequenced for the three regions, so there is more confidence in the signal obtained from the combination of the sequences from all the loci.

The parsimony and Bayesian analyses presented the two major clades ("A" and "B") of *Hymenolobium* and the relationships within each of them are consistent with those suggested by the more resolved ITS phylogeny. The lineages follow the same pattern as in the other phylogenies, even when these analyses used considerably fewer taxa than the other analyses done for the rest of the datasets. The trees generated from the analyses of this dataset are not included since they did not provide new information in matters of resolution or support compared to the results of the other analyses.

Figure 1. One of the 49 most parsimonious trees resulting from the analysis of the ITS region. Values show bootstrap support >50%; * indicates nodes collapsed in the strict consensus tree. In blue Clade A, in green Clade B, in orange Basically Divergent Taxa in *Andira*, in red Clade C, in purple Clade D. CI= 0.62, RI=0.90.

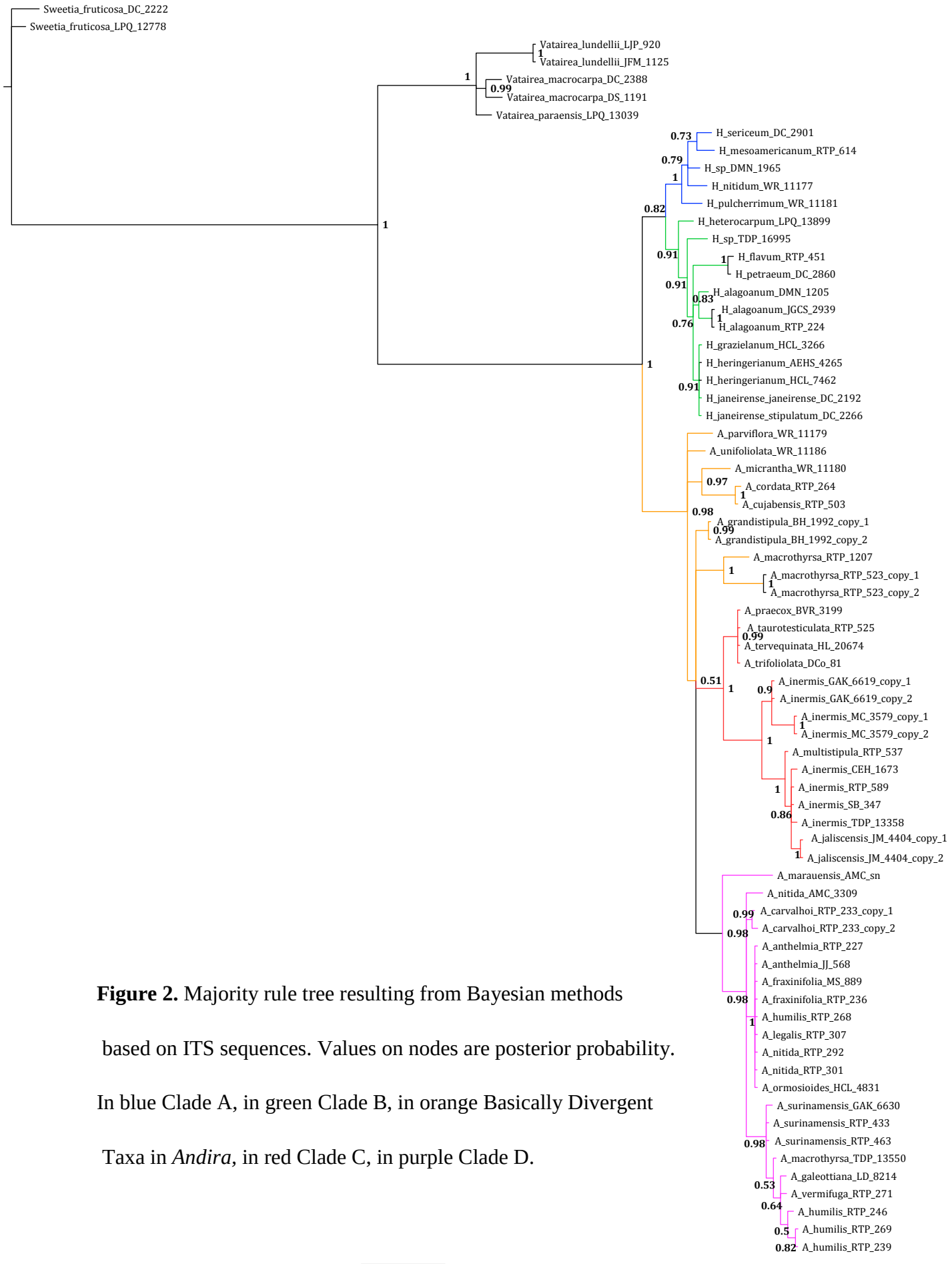

Figure 2. Majority rule tree resulting from Bayesian methods based on ITS sequences. Values on nodes are posterior probability. In blue Clade A, in green Clade B, in orange Basically Divergent Taxa in *Andira*, in red Clade C, in purple Clade D.

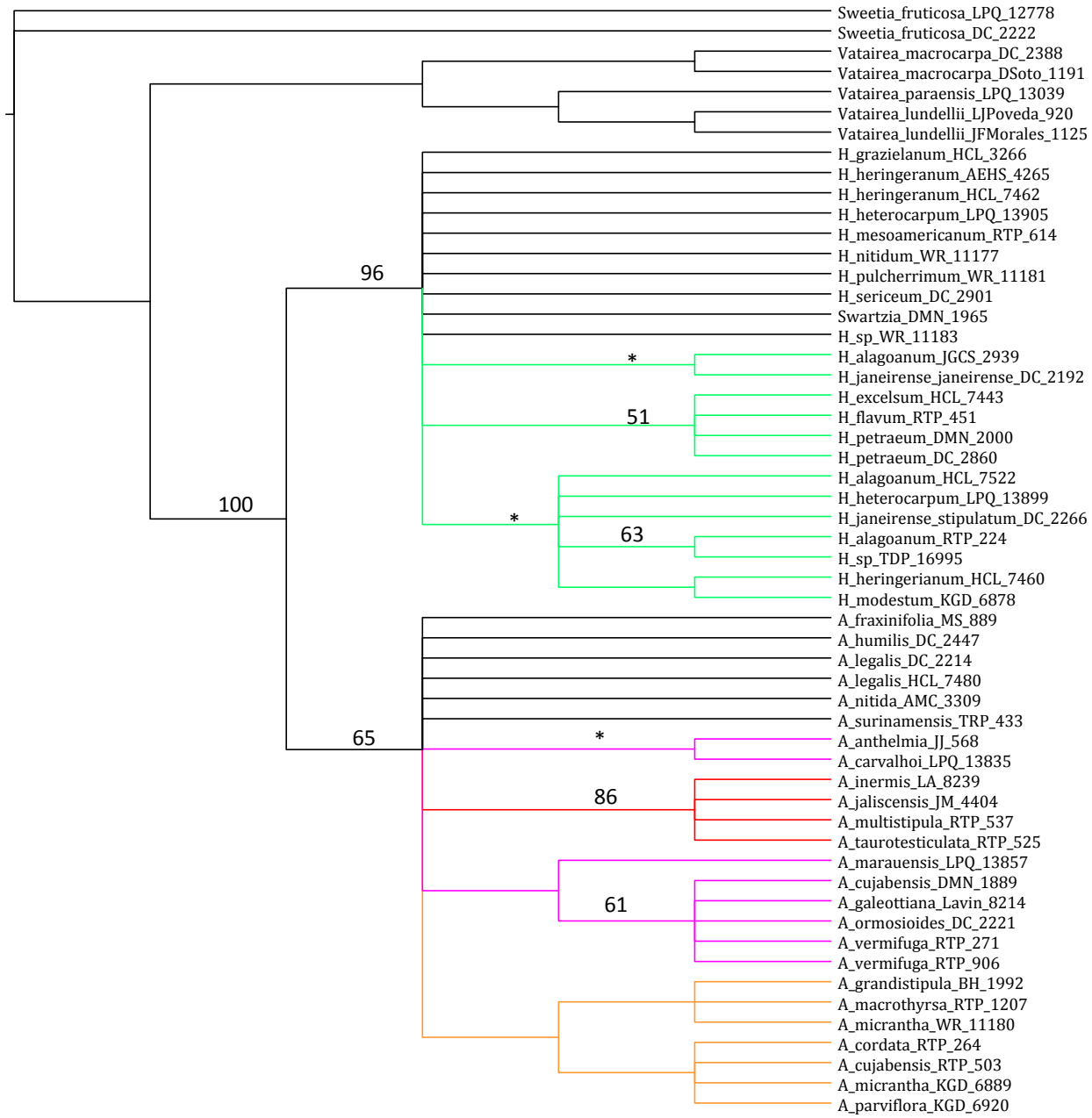

Figure 3. One of the 13 most parsimonious trees resulting from the analysis of the *mat*K region. Values show bootstrap support >50%; * indicates nodes collapsed in the strict consensus tree. In colours: clades also present in Clade B (green), Basically Divergent Taxa in *Andira* (orange), Clade C (red), Clade D (purple). CI= 0.98, RI=0.99.

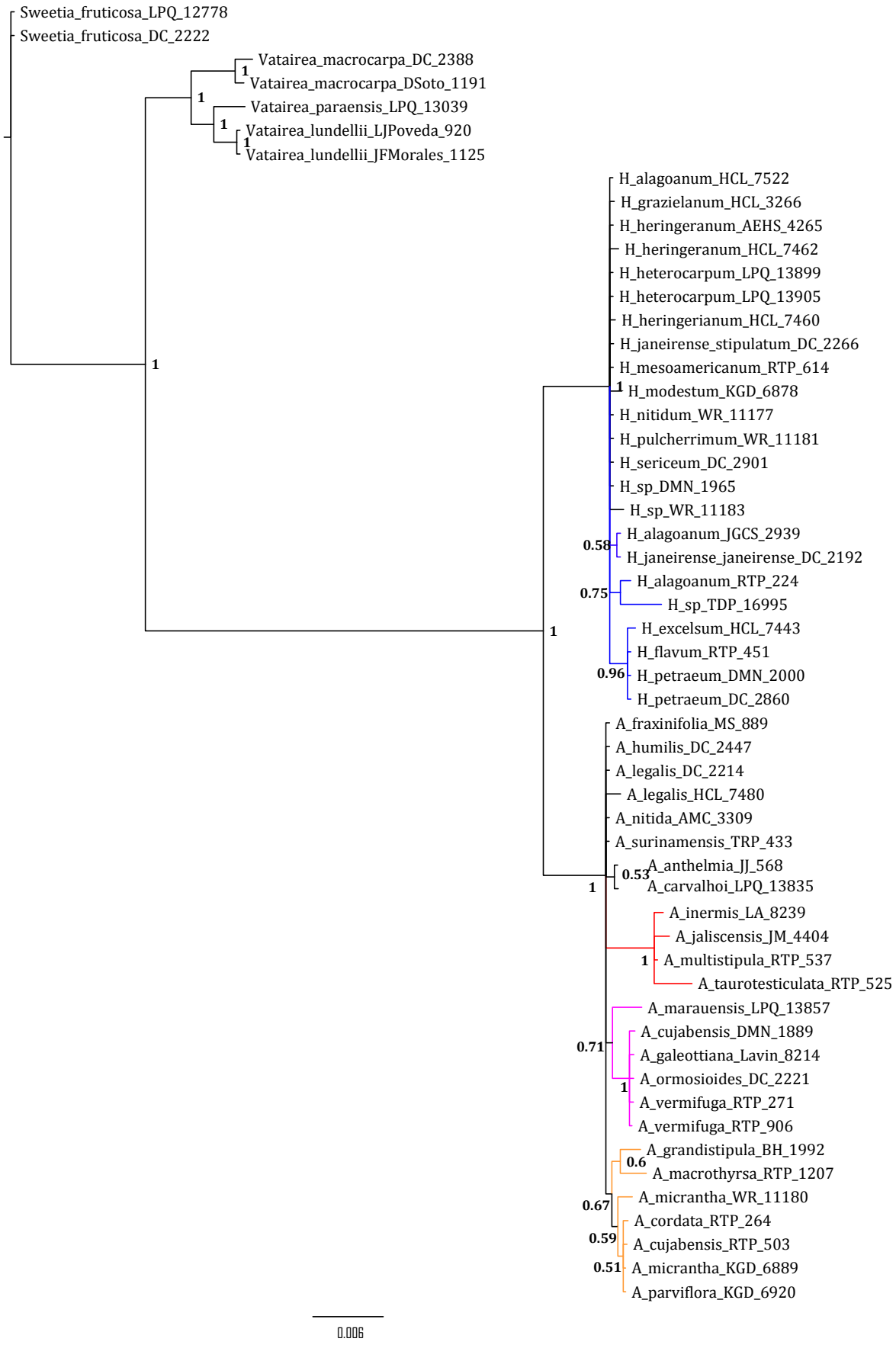

Figure 4. Majority rule tree resulting from Bayesian analysis based on *mat*K sequences. Values on nodes are posterior probability. In colours: clades also present in Clade B (green), Basically Divergent Taxa in *Andira* (orange), Clade C (red), Clade D (purple).

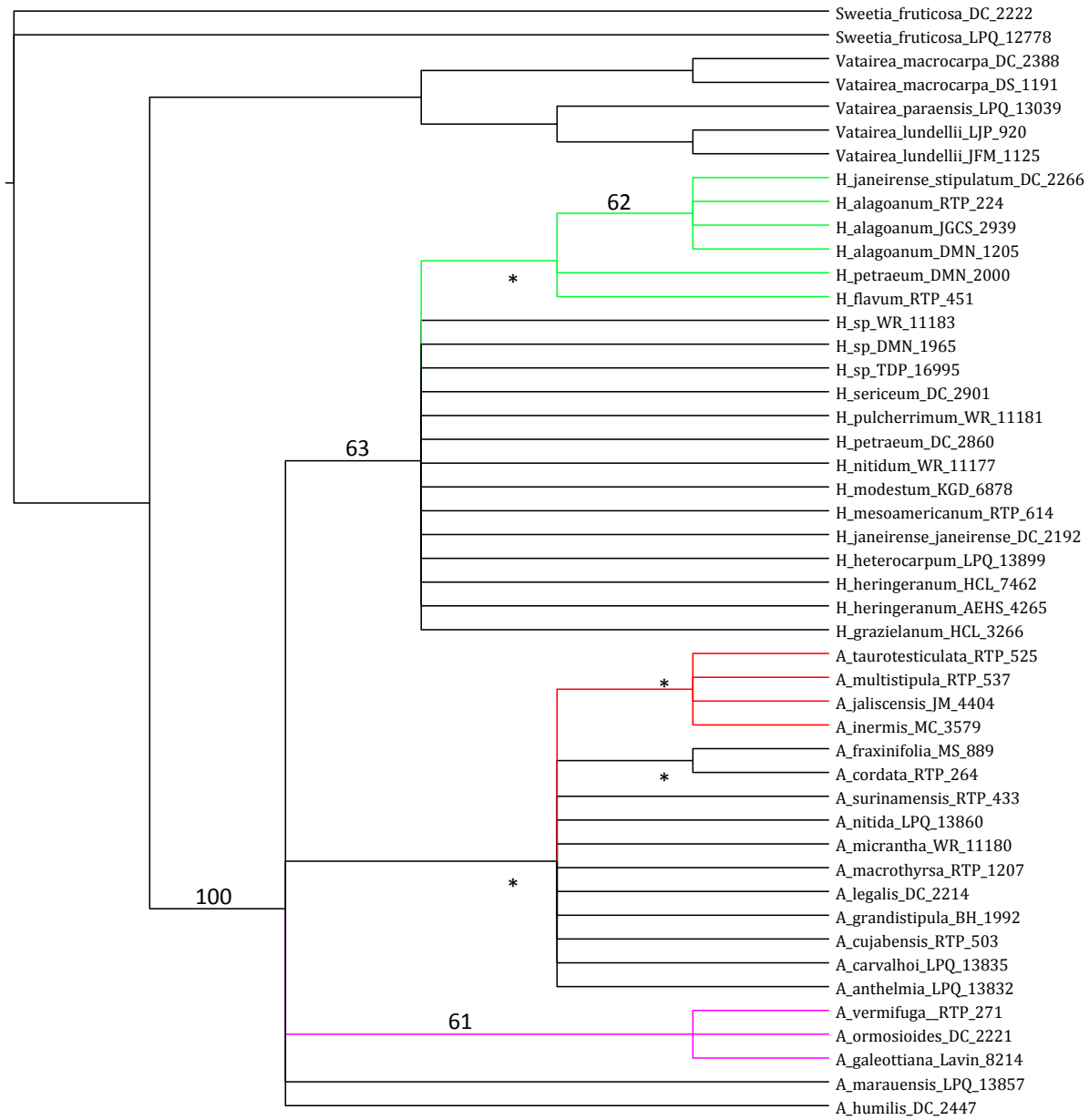

Figure 5. One of the 22 most parsimonious trees resulting from the analysis of the *trn*L region. Values show bootstrap support >50%; * indicates nodes collapsed in the strict consensus tree. In colours: clades also present in Clade B (green), Basically Divergent Taxa in *Andira* (orange), Clade C (red), Clade D (purple). CI= 1, RI=1.

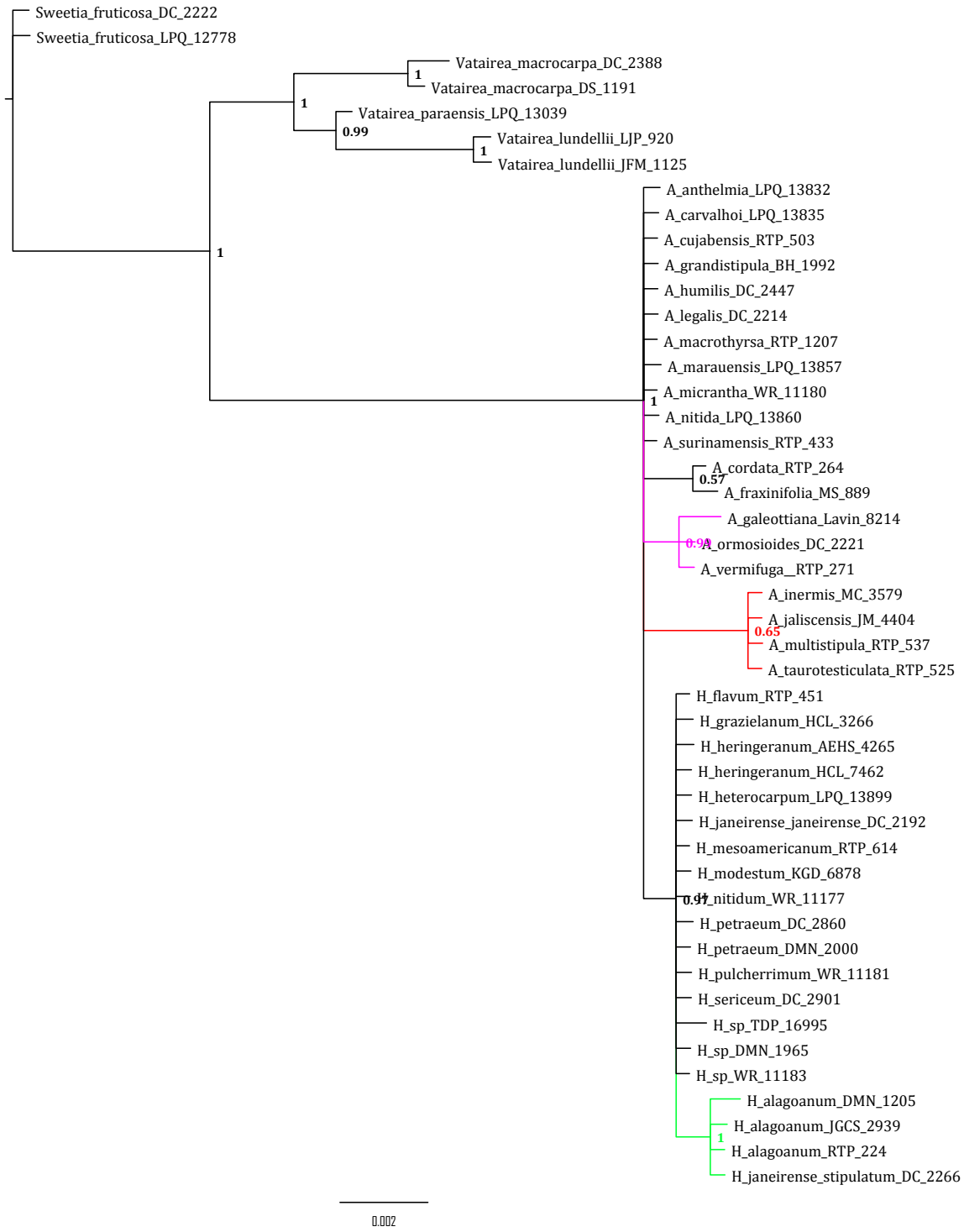

Figure 6. Majority rule tree resulting from Bayesian methods based on *trn*L sequences. Values on nodes are posterior probability. In colours: clades also present in Clade B (green), Clade C (red), Clade D (purple).

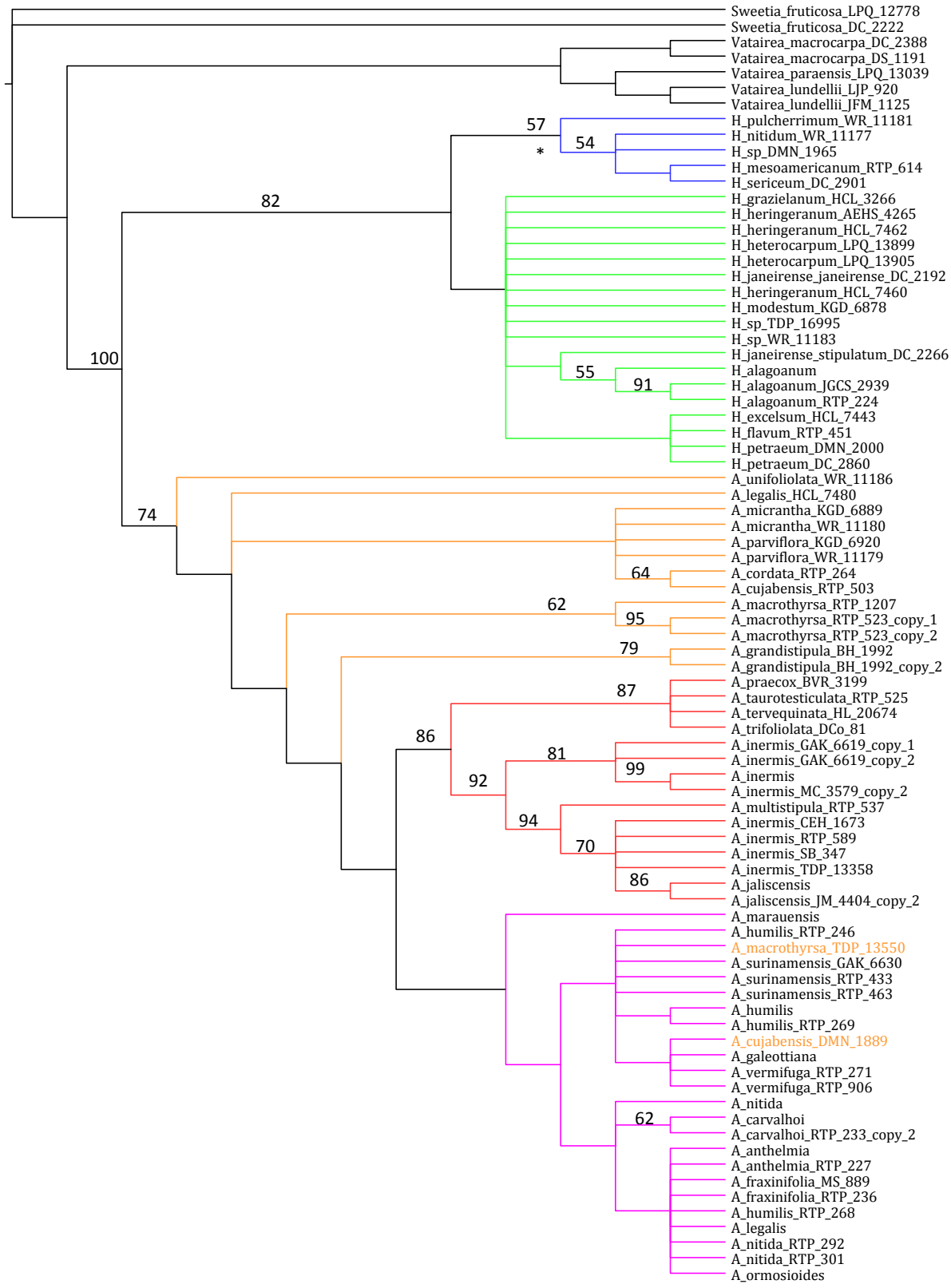

Figure 7. One of the 6 most parsimonious trees resulting from the combined analysis of the ITS and chloroplast regions. Values show bootstrap support >50%; * indicates nodes collapsed in the strict consensus tree. In blue Clade A, in green Clade B, in orange Basically Divergent Taxa in *Andira*, in red Clade C, in purple Clade D. CI= 0.76, RI=0.93.

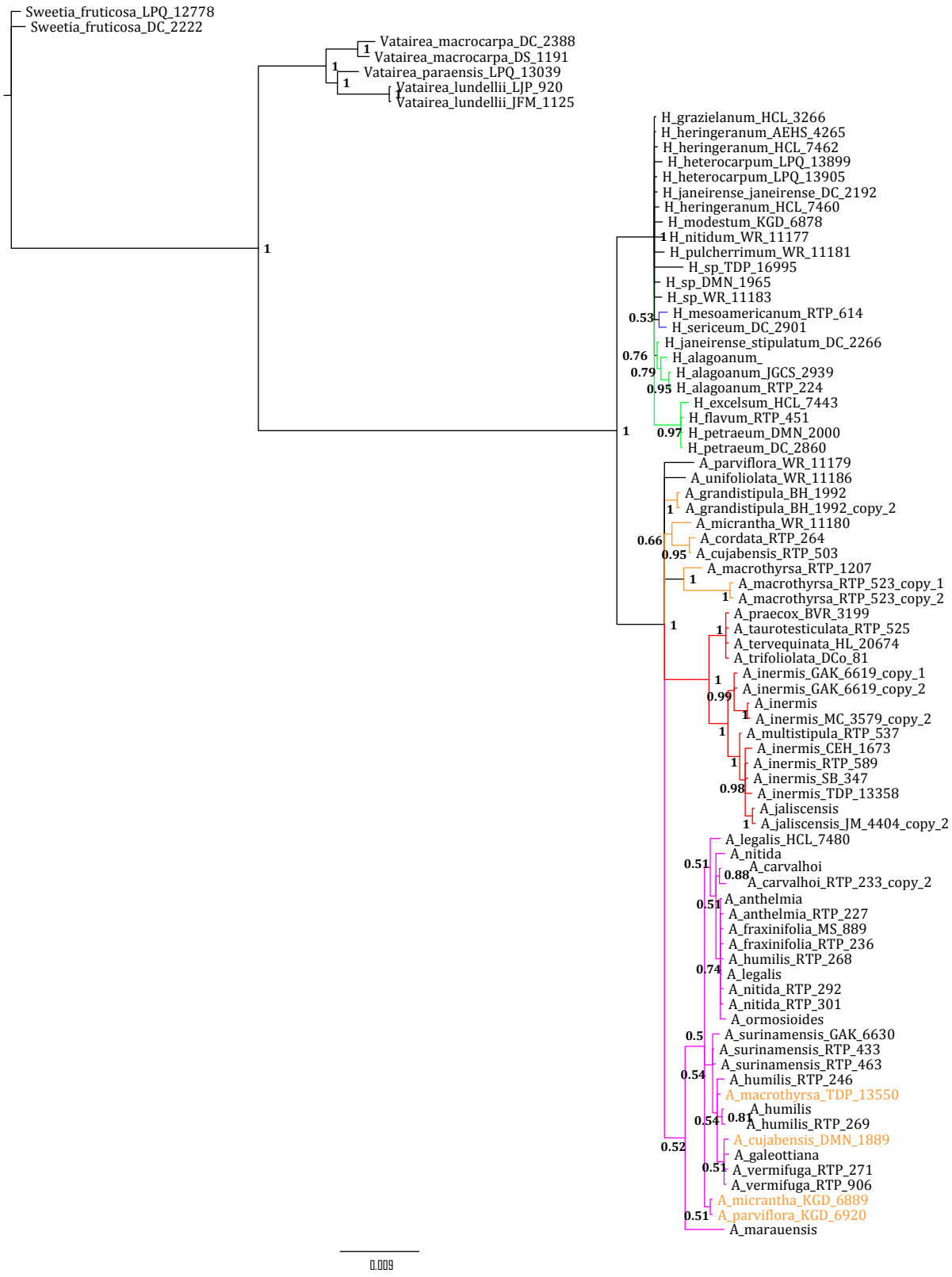

Figure 8. Majority rule tree resulting from Bayesian methods based on the combined analysis of ITS, *mat*K and *trn*L sequences. Values on nodes are posterior probability. In colour clades also present in: blue Clade A, in green Clade B, in orange Basically Divergent Taxa in *Andira*, in red Clade C, in purple Clade D.

4. DISCUSSION

4.1. Selection of data for the analyses

The exploratory analyses of the initial matrices allowed examination of different configurations (i.e. including or excluding portions of sequences or accessions) of the same data set in order to select one that would maximise the support for clades and the resolution within them. I attempted to compile a homogenized data set in terms of taxa sampled for each of the loci, and length and quality of sequences, in order to decrease the potential sources of conflict in the analyses (Wendel and Doyle 1998); however some difficulties arose in the process. The major drawback was, a consequence of the effort to achieve broad taxon sampling, since it lead to the compilation of material from varied sources (e.g. dried leaves, extracted DNA, GenBank sequences) for different taxa for the three loci. This was considered inconvenient in the sense that it hindered the comparison of some results, thus the potential explanation of some of the patterns observed in the topologies of the trees. I will illustrate this with an example: *Hymenolobium excelsum* was a species that was sequenced only for *mat*K, and it was represented only by one accession. The problems in this particular case were that the phylogenetic relationships between this accession and other *Hymenolobium* sequences could not be contrasted with the generally more informative ITS regions (Johnson and Soltis 1995, Soltis and Soltis 1998). Additionally, the availability of only one accession for the species reduce the ability to assess the accuracy of the placement of the species within the tree by recognition of potential misidentifications or laboratory contaminations. Additionally, with a single accession for a species, we miss the possibility to uncover biological processes that may be influencing the pattern observed in the *mat*K topology such as hybridization and lineage sorting. These processes can only be uncovered by inclusion of at least two accessions of the same species in the tree.

4.2. Contrasting phylogenetic hypothesis based on different genes

4.2.1. Resolution obtained with nuclear genes versus chloroplast genes

Chloroplast regions were not successful clarifying the species relationships within each of the genera. One common issue for both chloroplast data sets compared to that of ITS was the reduced number of taxa included in the analyses due to availability of material, problems in the laboratory, and most particularly time limitations during this research project.

The taxon sampling disparity among regions is an important issue to keep in mind when comparing the resolution of the trees obtained for each locus.

In the tree based on *trn*L the whole *Andira* lineage is unresolved with the exception of the group *A. galeottiana*, *A. ormosioides* and *A. vermifuga*, and *Hymenolobium* remains almost entirely as a polytomy. Although this region has proven to work well between and within genera (e.g. Gielly and Taberlet 1996, Bellstedt et al. 2001), that was not the case in this study. It must be acknowledged that not all sequences of *trn*L available for the analyses included both the *trn*L intron as well as the *trn*L-*trn*F spacer. A considerable number of the samples were sequenced only for the intron, missing potentially informative characters present in the more variable spacer (Shaw et al. 2005), thus perhaps contributing to an unresolved *Andira* – *Hymenolobium* clade. When it comes to *mat*K, despite of being one of the most rapidly evolving protein-coding regions (Wolfe, 1991), it gave low phylogenetic signal for both genera compared to that resulting from the analysis of ITS. As in any protein-coding region it is expected that the nucleotide sequences could be divided into sets of three bases (codons) that would translate into amino acid sequences. It is possible to corroborate this quality of a protein-coding region by checking the reading frame; however when the matrix of *mat*K was translated into protein, it was evident that there were alignment errors, leading to an incorrect translation of the gene, and potentially affecting the results of the phylogenetic analysis (Ritland and Clegg, 1987). These errors could not be fixed before performing the analyses because of time limitations; most of them were related to one region comprised between the characters 1727-1758 of the matrix. Given that the inclusion and exclusion of alignment ambiguous regions in matK did not affect results, it seems highly likely that these alignment issues in such a short region of the alignment would also have no effect on the inferred phylogeny.

Regardless of the potential errors that could have been present in the analysis of the chloroplast data, the contrasting difference in the resolution of the trees may also be attributed to the fact that *Hymenolobium* and *Andira* have gone through recent diversification events (Skema, 2003). Therefore, for these genera, the ITS region, with its high nucleotide substitution rate possesses great information content that is more useful to infer phylogenetic relationships at lower taxonomic levels compared to the chloroplast genes (Baldwin et al. 1995; Johnson and Soltis 1995; Soltis and Soltis 1998).

4.2.2. Comparison with previous chloroplast phylogenies of *Andira* and *Hymenlobium*

Previous phylogenetic analyses for both genera have been done, yet this is the first study that attempts to combine a broad sampling of taxa to contrast evolutionary hypotheses of *Andira* and *Hymenolobium* based on nuclear as well as chloroplast genes. In a revision of the phylogeny of early branching papilionoids, Cardoso et al. (2012) generated a tree based on the combined data from *trn*L and *mat*K sequences. The study of Cardoso et al. (2012) was focused in resolving relationships at subfamily level and it is necessary to do a more exhaustive compilation of species per genus, as the one intended in my study, in order to obtain species-level phylogenetic hypotheses. Yet the result of including a broader number of taxa for each of the regions, particularly for *mat*K, did not clarify too much the relationship of species within *Hymenolobium.* The genus remained mostly unresolved with only one clade with high support (the one including *H. excelsum, H. flavum,* and *H. petraeum*). The case of *Andira* was different, the inclusion of 13 more species in my study in addition to those considered by Cardoso et al (2012), generated a more resolved *mat*K phylogeny, that implied the grouping of some taxa that can be contrasted with the results from the cladistics analysis of Pennington (2003) using chloroplast restriction site data (Appendix II).

The phylogenetic hypothesis obtained using *mat*K data in my work supported the group of *A. inermis, A. multistipula, A. jaliscensis* and *A. taurotesticulata* (Clade C), which was called Clade I by Pennington (2003). The Clade II in his phylogeny comprises *A. cordata, A. cujabensis, A. grandistipula, A. parviflora* and *A. unifoliolata*, which would be equivalent to my Basal grade, with the modification that *A. unifoliolata* is not present since it was not sequenced for *mat*K, and the inclusion of *A. macrothyrsa* with this group of species (which in Pennington's phylogeny was located in the Plastome Group I clade). My phylogeny from *mat*K did not resolve the group of species from Clade D as the monophyletic lineage obtained in the ITS analysis; instead it resulted in a polytomy of a combination of species from Clade D1 and D2, and a poorly supported group with *A. marauensis, A. cujabensis, A. galeottiana, A. ormosioides* and *A. vermifuga.* The placement of *A. cujabensis* in this lineage is discussed in more detail further in this chapter. The suggestion of the close relationship of *A. marauensis* with the rest of the species of the group has low support; still the clustering of *A. galeottiana, A. ormosioides* and *A. vermifuga* was reinforced with a posterior probability of 1 in the Bayesian analysis. In the chloroplast restriction data analysis of Pennington (1995, 2003) these species differ in their plastome type, resulting in their placement in separate clades: *A. ormosioides* was characterized with a plastome type of Clade III, while both *A.*

galeottiana and *A. vermifuga* happened to have restriction site mutations characteristic of the clade Plastome Group I.

4.2.3. *Hymenolobium* phylogeny based on ITS

Since the phylogeny of *Andira* based on the ITS region has previously been described in detail in the study of Skema (2003), in this section I will focus on the results obtained for *Hymenolobium*.

The clade of *Hymenolobium* resolves in two major groups, Clade A and Clade B, which seem to correspond with the geographical distribution of the species. In Clade A, with the exception of *H. mesoamericanum* which is restricted to gallery forests of Costa Rica, Nicaragua and Panama (Lima 1988), all other taxa are reported to be distributed in the Amazonian rain forest in Brazil (*H. nitidum, H. pulcherrimum* and *H. sericeum*) and Peru (*H. nitidum*). As for Clade B, the accessions located in a basally divergent in the clade (*H. heterocarpum* LPQ 13899 and *H. sp.* TDP 16995) were both collected in the Amazon as well (Brazil and Peru, respectively), while the rest of the species are also organised in three groups: 1) a Guayanan lineage of *H. flavum* and *H. petraeum*, with a potential inclusion of the Amazonian *H. excelsum*, as proposed by the results of the combined dataset analysis (only sequenced for chloroplast loci), 2) a cluster of the *H. alagoanum* accessions, a species typical of the Brazilian Atlantic coast Rain Forest, and a 3) unresolved group with a combination of taxa from the Brazilian Atlantic coast (*H. janeirense, H. heringerianum*) and from the Brazilian Amazon rain forest (*H. grazielanum*, and potentially *H. modestum*, according to the results from the Bayesian analysis).

When the phylogenetic hypothesis obtained with the molecular analysis was contrasted with morphological features, there was no character (or set of characters) that would define, straightforwardly as unequivocal synapomorphies, each of the clades within *Hymenolobium*; a similar case to that of "cryptic" species within *Andira* reported by Pennington (2003). From a palynological study of the genus, Gurgel et al. (2000) described the genus as stenopalynous, which means that pollen characteristics do not show significant variation within the group. However, from a superficial exploration of the macromorphological traits in the descriptions of the species (Ducke, 1936; Rizzini, 1969; Mattos, 1979; Lima 1982; Lima, 1988), it seems that those taxa from Clade A tend to have

longer inflorescences, and bigger flowers and fruits, compared to those from Clade B. Within Clade A, the fact that they share smaller and more numerous leaflets in the compound leaves reinforces the grouping of *H. mesoamericanum* and *H. sericeum*; additionally, both species have in common with *H. nitidum* the greyish terminal branches that are pubescent at the newest growth.

The case of Clade B was similar in the sense that there was no evident morphological character that is shared by all species of the group, that is not present in those of the Clade A, besides appearing to have smaller sized inflorescences, flowers and fruits. Nevertheless, one feature that is present in many species of the B group is the presence of yellow trichomes, particularly on the adaxial surface of the leaflets. These yellow hairs are present in the rachis of the inflorescence of *H. heterocarpum* and *H. flavum*; the pubescence in the upper surface of the leaflets, particularly of tomentose type, is also a character that reinforces the clustering of *H. flavum, H. petraeum*, and potentially *H. excelsum.* Palynological evidence would also support the placement of *H. excelsum* in the clade of *H. flavum* since both share micro-reticulate pollen (Gurgel et al. 2000). The rest of the species from the clade share a combination of characters *H. grazielanum, H. modestum* and *H. alagoanum* have glabrous and lustrous adaxial surfaces of the leaflets, *H. grazielanum* and *H. alagoanum* also have in common the colour of the flower, *H. modestum* and *H. janeirense* share the presence of brownish hairs in the calyx; however the species that may have the strongest similarity within the clade could be *H. modestum* and *H. alagoanum,* both with adaxially lustrous, coriaceous leaflets, and similar sized inflorescences and flowers. Lima (1982) acknowledged this resemblance when he was describing *H. modestum*, and was going to propose it as synonym of *H. alagoanum*, however considering "the remarkable differences of the yellowish pubescence and the disparity in fruit consistency" he decided to keep them as a different species.

4.2.4. Congruence between chloroplast and nuclear data

From a visual comparison, the gene trees obtained based on *mat*K and ITS regions varied greatly in respect to resolution level for both genera; still the few strong supported clades resolved in the *mat*K based phylogeny were all present in the ITS tree, with the exception of the lineage that grouped *A. cujabensis*, *A. galeottiana*, *A. ormosioides* (DC 2221) and *A. vermifuga* (DMN 1889) in the matK tree Although these differences might

reflect biological processes such as potential hybridization with subsequent introgression or lineage sorting, in these cases it is suspected that the problem at lie in misidentification of the specimens of *A. cujabensis* and *A. ormosioides*. Both species are very difficult to distinguish from A. vermifuga when sterile (Pennington, 2003 and pers. comm.). It has not been possible to check the voucher specimens during this research project, but this will be a priority task before publication of these data.

The ITS gene tree also shows a separated distribution of the *A. humilis* accessions included in Clade D1 (RTP 239, 246, 269) and D2 (RTP 268). Pennington (1995, 2003) reported the presence of intraspecific cpDNA polymorphism for this species and also for *A. carvalhoi*, specifying that the intraspecific genetic variation of these taxa is likely to be the result of introgressive hybridization. However, Pennington (2003) states that if the existence of hybrids is assessed by the presence of specimens with intermediate characteristics (which may not be necessarily the case as exposed by McDade, 1990), then hybridization between the species of *Andira* is not a common process due to the rare observation of specimens with intermediate features.

Another interesting aspect worth mentioning is the pattern observed with species from the Clades C 2.1 and C2.2. These lineages form part of a monophyletic group C 2 that is divided in a subgroup of only *A. inermis* accessions (GAK 6619, MC 3579) and a second one which includes a combination of samples of *A. inermis* (CEH 1673, RTP 589, SB 347) with *A. multistipula* and *A. jaliscensis.* In an unpublished study Pennington and Lavin give detailed examples of how several widespread species representative of rain forests and savannas in the neotropics have in their distribution range daughter species that are geographically confined, and therefore frequently present a pattern of non-monophyly of conspecific DNA sequences in the phylogenies. *A. inermis* is considered as an example of this, where the distinctive morphological characterization of the species from throughout the range of distribution means that hybridization is an unlikely reason to explain the phylogenetic relationships of *A. inermis, A. jaliscensis* and *A. multistipula*. *Andira inermis* is proposed to be the species from which the nested *A. multistipula* and *A. jaliscensis* originated (Pennington 2003); therefore they suggest that the non-monophyly of the conspecific sequences in this case is a consequence of the retention of the ancestral genetic polymorphism of *A. inermis.*

Finally, in the ITS tree it seemed that *A. macrothyrsa* was another case of a species with polyphyletic accessions, since the sample *A. macrothyrsa* TDP 13550 was located within the D1 clade, instead of next to the rest of the *A. macrothyrsa* accessions in the basal region of the tree. Nevertheless, this accession is likely to be a misidentified specimen of *A. surinamensis* (Pennington pers. com.), which is vegetatively similar to *A. macrothyrsa,* but the lack of reproductive characters of the sample hindered the accurate determination of the specimen.

4.2.5. Combinability and advantages of combining the data

The method used in this study in order to assess the combinability of the different datasets was the conditional combination approach (de Queiroz, 1993; Bull et al., 1993); the idea behind it is to minimize the error in estimating relationships of the taxa by preventing the grouping of divergent data (Hardig et al., 2000). In this approach the data are analysed separately and the level of incongruence is tested previous to the fusion of the partitions. It is preferable to estimate similarity between the dataset from different methods, e.g. assessing topological conflict (Colles, 1980; Rohfl, 1982), estimating character congruence (Mickevich and Farris, 1981) and /or with significance tests for heterogeneity, and in this study topological conflict was assessed from visual examination and congruence of the data was estimated performing the partition homogeneity test. Although topologies of the trees obtained based on the separate analyses of the different regions varied greatly in respect to the resolution they provided, none of the relationships proposed, and supported with moderate posterior probability (values higher than 81%), were in conflict. Furthermore, the outcome of the partition homogeneity test indicated that there was not any significant incongruence between the data. In conclusion, one would expect that the combination ITS and cp data should not result in a potential source of error in the inference of phylogenetic relationships of the taxa.

The separate and combined analysis of the chloroplast data resulted in unresolved trees compared to the more structured one generated by the combined analysis including ITS regions. This is why in this section of the discussion when I mention a combined analysis I will be making reference to the more resolved topology obtained by assembling the sequences from of all regions. The outcome of merging the data from the different loci had a different effects in the phylogenetic signal of each genus: in *Andira* the structure of the tree

remained quite similar (with a few different clades that had not a strong support), but in the case of *Hymenolobium* much of the phylogenetic structure evidenced in the separate ITS analysis was lost after combining the data with the chloroplast sequences. Although the accuracy of the estimated phylogenetic hypothesis generally increases with the addition of congruent characters, the loss of resolution for *Hymenolobium* in the combined analysis may be a result of the dilution of the variable ITS characters by random or systematic errors coming from the less informative chloroplast regions (Bull et al 1993). Even so, the addition of the chloroplast data had the advantage that it incorporated those species that could not be sequenced for ITS due to logistic drawbacks (e.g. lack of material availability, time limitations).

4.3. Biogeographical considerations of *Hymenolobium* and *Andira*

Lavin and colleagues (2005) reported that a dated phylogeny for the legume family suggesting that the divergence of the *Andira* clade and the core dalbergioid crown node is aged 55.5-57.3 Ma, and that a subsequent separation of *Hymenolobium* and *Andira* occurred approximately 9.3-29.8 Ma. A biogeographic study of *Andira* dated a phylogeny based on ITS sequences which are used here (Skema, 2003), the estimated age ranges that I am using in the remainder of the discussion are those calculated for *Andira* based on the rates of substitution estimated for *Inga* (Richardson et al 2001). The genus *Andira* has been characterized by having both recent and ancient events of diversification (Simon et al. 2009), with two major speciation events: one in the mid-Neogene (7-18Ma) that resulted in the radiation of the *Andira* species from the Basal Clade, and a second one involving the diversification of the Clade D lineage around the mid-Pleistocene to Holocene (0-6 Ma). As for *Hymenolobium*, there is no previous study in which speciation events of the genus could be distinguished within a temporal frame.

It is worth noticing that both genera seem to have a similar ecological history in that oldest lineages in *Hymenolobium* (all species of Clade A and the ancestor of *H. mesoamericanum*) and *Andira* (*A. unifoliolata, A. parviflora, A. micrantha, A. macrothyrsa*, and the ancestor of *A. cujabensis – A. cordata*) are rainforest species. In the case of *Andira,* this early radiation is estimated to have happened during the mid-Neogene (7-18Ma). Both genera also have Atlantic rainforest clades nested in an apical position (Clade B2 and B3 in *Hymenolobium;* Clade D2 in *Andira*) that are lacking strong sequence divergence; all of

which is suggestive of a recent radiation burst of those lineages in the rain forests of the Mata Atlantica. In the case of *Andira*, the radiation of this clade is dated to 0-6 Ma; and related to ecological isolation mechanisms (e.g., some *Andira* species such as *A. carvalhoi* and *A.nitida* are endemic to the sandy resting vegetation on the Atlantic coast) or allopatric divergence due to Plio-Pleistocene climatic events in the Atlantic coast of Brazil (Fjedsa, 1994) .

Finally, another biogeographical pattern shared by both is the presence of apically nested Central American species, suggesting three migrations of ancient Amazonian lineages from South America to Central America; such is the case of *H. mesoamericanum, A. galeottiana* and the lineage *A. inermis - A. jaliscensis*. The estimated age of divergence of *A. multistipula* and *A. inermis - A. jaliscensis* is around 0.5-1.5 Ma, this recently diversification of the group postdates the closure of the Isthmus of Panama (~3-3.5 Ma Burnham and Graham, 1999), so it is tempting to suggest that there was a dispersion event in which the ancestor of the *A. inermis – A. jaliscensis* lineage gradually migrated to Central America and that a subsequent *in situ* speciation event lead to the differentiation of *A. inermis* and *A. jaliscensis*. The case of *A. galeottiana* could be similar, since the radiation of the clade D2 is relatively recent (0-6 Ma), the time ranges of diversification of the lineage overlaps to those of the formation of the land connection between South and Central America; otherwise, it is known that the fruits of both *A. vermifuga* and *A. galeottiana* are adapted to water dispersal (Pennington, 2003), so the possibility of the migration of the *A.* galeottiana ancestor from South America to Central America through water dispersion may be considered as an alternative.

The evolutionary history of *Andira* is suggested to be dynamic, with a combination of recent and ancient events of lineage divergence. In these terms, a deeper understanding of the diversification of *Andira*, and presumably of *Hymenolobium*, will require further systematic studies with a focus not only of historical factors but including ecological aspects as well, that have been proven to be strongly related to the phylogenetic structure of many tropical clades (Hughes et al 2013).

5. CONCLUSIONS AND FURTHER WORK

This study shows *Hymenolobium* and *Andira* are both monophyletic. The fast evolving ITS regions resolved the relationships in both genera but it seems possible that the

lack of resolution in the chloroplast gene trees reflects their slower evolution. There was no phylogenetic incongruence between the strongly supported clades in each of the phylogenies. The few inconsistencies found were related to the disparity of taxon sampling for each of the regions, and the potential misidentifications of three accessions, in the combined total evidence phylogenetic approach.

The phylogenetic relationships within *Hymenolobium* suggest the presence of two lineages, but there is no distinctive morphological feature that defines these clades. It appears however that there is an ecological and geographical pattern in the genus, which happens to be shared with *Andira*, involving an Amazonian origin with a dispersal of some elements to Central America and the Brazilian Atlantic coast followed by a rapid radiation.

The *Andira* clade represents an interesting group to study many evolutionary and biogeographical aspects. Comparative work including low-copy DNA sequences may be useful to improve the lack of resolution in the phylogenies based on *mat*K and *trn*L. It would also be sensible to assess the evidence of potential introgression and gene duplication within the genera in order to reinforce the accuracy of the phylogenies here presented. In these terms, it would be essential to elaborate a dated phylogeny of the whole clade and include it in a phylogenetic study focused in the potential ecological features that may be influencing the evolutionary and distribution patterns of the group.

6. ACKNOWLEDGMENTS

I would like to thank my supervisor, Toby Pennington, for his guidance through the project and for giving me the opportunity to work with an interesting group, as it is the *Andira* Clade. I also would like to thank my co-advisors Kyle Dexter and James Richardson for their valuable comments and suggestions in the elaboration of the manuscript. I would like to acknowledge Laura Forrest, Michelle Hollingsworth, Ruth Hollands for their teaching in the laboratory; Michael Moeller for the patience in the many, many, hours building and analysing the data sets; Alexandra Clark and Cynthia Skema for providing many of the DNA sequences used in this study; and Domingos Cardoso for providing also some of the material and for helping me through the understanding of the *mat*K alignment and analysis. I am also grateful to Tiina Sårkinen, my MSc mentor, for her support and friendship during this course; to David Harris for allowing me to volunteer in the herbarium and join him in the passionate

work of herbarium curatorship; to Louis Ronse De Craene for his help and guidance as director, and his contagious enthusiasm as a lecturer of the course; and to all RBGE staff for making of this year a positive experience. Lastly, I would like to render a special thanks to my MSc colleagues that have been my "botanical family" through all the year.

7. REFERENCES

Alvarez, I. & Wendel, J. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular phylogenetics and evolution*. **29**: 417–434.

Antonelli, A. & Sanmartín, I. 2011. Why are there so many plant species in the Neotropics? *Taxon* **60:** 403-414.

Bardon, L. Champagne, J., Dexter, K., Sothers, C., Prance, G. & Chave, J. 2012. Origin and evolution of the Chrysobalanaceae family: insights into the evolution of plants in the Neotropics. *Botanical Journal of the Linnean Society* **171**: 19-37.

Barroso, G., Peixoto, A., Ichaso, C., Costa, C., Guimarães, E. & de Lima, H. 1991. *Sistemática de Angiospermas do Brasil*, vol 2. Imprensa Universitária, Minas Gerais, Brazil.

Bellstedt, D., Linder, H. & Harley E. 2001. Phylogenetic relationships in *Disa* based on non-coding *trn*L-*trn*F chloroplast sequences: Evidence of numerous repeat regions. *American Journal of Botany*. **88**: 2088-2100.

Bentham, G. 1860. A synopsis of the Dalbergieae, a tribe of the Leguminosae. *Journal of the Proceedings of The Linnean Society* **IV**(Supplement): 1-134.

Bentham, G. 1876. Leguminosae. In Mart. Fl. Bras. **15:**274.

Bull, J., Huelsenbeck H., Cunningham C., Swofford, D., & Waddell, P. 1993. Partitioning and combining data in phylogenetic analysis. *Systematic Biology* **42:**384-397.

Burnham, R. & Graham, A. 1999. The history of Neotropical vegetation: new developments and status. *Annals of the Missouri Botanical Garden* **86**: 546-589.

Cardoso, D., de Queiroz, L., Pennington, R., de Lima, H., Fonty, E., Wojciechowski, M. & Lavin, M. 2012. Revising the phylogeny of papilionoid legumes: new insights from

comprehensively sampled early-branching lineages. *American Journal of Botany* **99**: 1991-2013.

Cardoso, D., Pennington, R., de Queiroz, L., Boatwright, J., Van Wyk, B., Wojciechoski, M. & Lavin, M. 2013. Reconstructing the deep-branching relationships of the papilionoid legumes. *South African Journal of Botany* **89**: 58-75.

Chen, S. Xia, T., Wang, Y., Liu, J. & Chen, S. 2005. Molecular Systematics and Biogeography of Crawfurdia, Metagentiana and Tripterospermum (Gentianaceae) based on nuclear ribosomal and plastid DNA sequences. *Annals of Botany* **96**: 413-414.

De Queiroz, A. 1993. For consensus (sometimes). *Systematic Biology* **42:** 368-372.

Doyle, J.J. 1992. Trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany*. **17**: 144-163.

Doyle, J.J. 1995. DNA data and legume phylogeny: A progress report. In: M.D. Crisp and J.J. Doyle [eds.], Advances in legume systematics, part 7, Phylogeny, 11–30. Royal Botanic Gardens, Kew, UK.

Doyle, J. & Doyle, 1987. A rapid DNA isolation method for small quantities of fresh tissue. *Phytochemical Bulletin* **19:**11-15.

Doyle J. & Doyle J. 1990. Isolation of plant DNA from fresh tissue. *Focus* **12**: 13-15.

Ducke, A. 1936. Notes on the species of *Hymenolobium*: giant trees of Brazilian Amazonia. *Tropical Woods* **47**:1-7.

Feliner, G. & Rosselló, J. 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*. **44**: 911-919.

Ferreira, G., Hopkins, M. & de Secco, R. 2004. Contribuição ao conhecimento morfológico das espécies de leguminosae comercializadas no estado do Pará, como "angelim". Acta Amazonica. **34**: 219-232.

Fjeldså, J. 1994. Geographical patterns for relict and young species of birds in Africa and South America and implications for conservation priorities. *Biodiversity and Conservation* **3:** 207-226.

Gielly, L. & Taberlet, P. 1996. A phylogeny of the European gentians inferred from chloroplast *trn*L (UAA) intron sequences. *Botanical Journal of the Linnean Society*. **120:**57-75.

Gurgel, E., Carreira, L. & Pereira, M.N. 2000. Leguminsoas da Amazônia Brasileira – XII. O Pólen do Gênero *Hymenolobium* Benth. (Leguminosae Papilionoideae). *Boletim do Museu Paraense Emílio Goeldi* **16**: 111-119.

Haffer, J. 1969. Speciation in Amazonian forest birds *Science* **165:** 131-137.

Hardig, T., Soltis, P., Soltis, D. 2000. Diversification of the North American shrub genus *Ceanothus* (Rhamnaceae): Conflicting Phylogenies from nuclear ribosomal DNA and chloroplast DNA. *American Journal of Botany* **87**: 108-123.

Hu, J., Lavin, M., Wojciechowski, M., & Sanderson, M. 2000. Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on *mat*K sequences, and implications for evolutionary patterns in Papilionoideae. *American Journal of Botany* **87:** 418-430.

Hughes, C., Eastwood, R. & Bailey D. 2006**.** From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society.* **361**: 211–225.

Hughes, C. Pennington, R. &Antonelli, A. 2013. Neotropical plant evolution: assembling the big picture. *Botanical Journal of the Linnean Society* **171:**1-18.

Johnson, L. & Soltis, D.1995. Phylogenetic inference in Saxifragaceae *sensu stricto* and *Gilia* (Polemoniaceae) using *mat*K sequences. *Annals of the Missouri Botanical Garden* **82:** 149-175.

Kajita, T., Ohashi, H., Tateishi, Y., Bailey, D. & Doyle, J. 2001. *rbc*L and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. *Systematic Botany* **26**: 515 – 536 .

Lavin, M., Pennington, R., Klitgaard, B., Sprent, J., de Lima, H. & Gsasson, 2001. The Dalbergioid Legumes (Fabaceae): Delimitation of a Pantropical Monophyletic Clade. *American Journal of Botany* **88**: 503-533.

Lavin, M., Herendeen, P. & Wojciechowski, M.F. 2005 . Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Systematic Biology* **54:** 575 – 594.

Lewis , G., Schrire, B., Mackinder, B., Lock, M. [eds.] 2005 . Legumes of the world. Royal Botanic Gardens, Kew, UK.

Lima, H. de. 1982. Consideraçoes taxonomicas sobre o genero *Hymenolobium* Bentham (LEG. FAB.). *Acta Amazonica* **12:** 41-48.

Lima, H. de. 1988. A New Species of *Hymenolobium* (Leguminosae-Papilionoideae) From Central America. *Annals of the Missouri Botanical Garden*. **75**: 1145-1147.

LPWG [The Legume Phylogeny Working Group]. 2013 . Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* **62**: 217-248.

Mattos**,** N. F. 1979. Ô genero *Hymenolobium* Bentham (Leguminosae) no Brasil. *Roessleria* **3**:13-53.

Möller, M. & Cronk, Q. 1997. Phylogeny and disjunct distribution: evolution of Saintpaulia (Gesneriaceae). *Proceedings of the Royal Society of London, Series B Series B.* **264***:* 1827-1836.

McDade, L. 1990. Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications fro cladistics analysis. *Evolution* **44:**1685-1700.

Mickevich, M. & Farris, J. 1981. The implications of congruence in *Menidia*. *Systematic Zoology* **30:** 351-370.

Nikon, K. & Carpenter, J. 1996. On consensus, collapsibility, and clade concordance. *Cladistics*. **12**: 305-321.

Pamilo, P. & Nei, M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*. **5:** 568-583.

Pennington, R. 1995. Cladistic Analysis of Chloroplast DNA Restriction Site Characters in Andira (Leguminosae: Dalbergieae). *American Journal of Botany* **82**:526-534.

Pennington, R. 1996. Molecular and morphological data provide resolution at different hierarchical levels in *Andira*. *Systematic Biology* **45:** 496-515.

Pennington, R. 2002. Proposal to conserve the name *Andira, nom. cons.* (Fabaceae) with a conserved type. *Taxon* **51**: 385-386.

Pennington, R.T. 2003. Monograph of *Andira* (Leguminosae-Papilionoideae). *Systematic Botany Monographs* **64**:1-143.

Pennington, R., Lavin, M., Ireland, H., Klitgaard, B., Preston, J. & Hu, J. 2001. Phylogenetic relationships of basal Papilionoid legumes based upon sequences of the chloroplast *trn*L intron. *Systematic Botany* **26**: 537-556.

Penninton, R. & Lavin, M. 2015. The nature of woody plant species in different neotropical forest biomes.Submitted.

Polhill, R. 1981. Papilionoideae. In R.M. Polhill and P.H. Raven [eds.], Advances in legume systematics, part 1, 191–208. Royal Botanic Gardens, Kew, UK.

Reiseberg. L. & Brunsfeld. S. 1992. Molecular evidence and plant introgression. In P.S. Soltis, D.E. Soltis, & J.J. Doyle [eds.] Molecular systematics of plants, 151-176. Chapman and Hall, New York, USA.

Richardson, J., Pennington, R., Pennington, T. & Hollingsworth. 2001. Rapid diversification of a species-rich genus of Neotropical rain forest trees. *Science* **293:** 2242-2245.

Ritland, K. & Clegg, M. 1987. Evolutionary Analysis of Plant DNA Sequences. *The American Naturalist* **130:** S74-S100.

Rizzini, C. 1969. Especies novas de Arvores do Planalto Central Brasileiro. *Anais da Academia Brasileira de Ciencias.* **41**: 239-245.

Rohlf, F.J. 1982. Consensus indices for comparing classifications. Mathem?caticalBio sciences: **59:** 131-144.

Rull, V. 2008. Speciation timing and Neotropical biodiversity: the Tertiary-Quaternary debate in the light of molecular phylogenetic evidence. *Molecular Ecology* **17:** 2722-2729.

Särkinen, T., Pennington, R., Lavin, M., Simon, M. & Hughes, C. 2012. Evolutionray islands in the Andes: persistence and isolation explain high endemisms in Andean dry tropical forests. *Journal of Biogeography* **39:** 884-900.

Shaw, J., Lickey, E., Beck, J., Farmer, S., Liu, W., Miller, J., Siripun, K., Winder, C., Schilling, E. & Small, R. 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.

Silva, M., Lisboa, P. & Lisboa, R.C.L. 1977. *Nomes vulgares de Plantas Amazônicas. Belém*. INPA, Brazil.

Simon, M., Rether, R., de Queiroz, L., Skema, C., Pennington, R. & Hughes, C. 2009. Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by *in situ* evolution of adaptations to fire. *Proceedings of the National Academy of Sciences* **106**: 20359–20364.

Skema, C. 2003. Phylogeny and Biogeography of Andira *Lam. nom. cons.* A Case Study in Neotropical Speciation. MSc thesis, University of Edinburgh, Edinburgh.

Soltis, D. & Kuzoff, R. 1995. Discordance between nuclear and chloroplast phylogenies in the Heuchera group (Saxifragaceae). *Evolution* **49**: 727-742.

Soltis, D. & Soltis, P.S. 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. In: P.S. Soltis, D.E. Soltis, and J.J. Doyle [eds.], Molecular systematics of plants II: DNA sequencing. Kluwer Academic Publishers. New York, USA.

Sprent , J. 2001. *Nodulation in legumes*. Royal Botanic Gardens, Kew, UK.

Stebbins, G. 1974. *Evolution above the species level.* Harvard University Press. Cambridge, Massachussets, USA.

Steele, P. & Wojciechoski, M. 2003. Phylogenetic analyses of tribes Trifolieae and Vicieae, based on sequences of the plastid gene *mat*K (Papilionoideae: Leguminosae). In: B.B. Klitgaard and A.Bruneau [eds.], Advances in legume systematics, part 10, Higher Level Systematics, 355-370. Royal Botanic Garden, Kew, UK.

Swofford, D., 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods), versión 4. Sinnauer associates, Sunderland, Massachussets, USA.

Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. 1991. *Plant Molecular Biology* **17:** 1105-1109.

Torke, B. & Schaal, B. 2008. Molecular phylogenetics of the species-rich Neotropical genus Swartzia (Leguminosae, Papilionoideae) and related genera of the swartzioid clade. *American Journal of Botany* **95**: 215-228.

Wendel, J., Stewart, J. & Rettig J. 1991. Molecular evidence for homoploid reticulate evolution in Australian species of *Gossypium.* Evolution. **45:** 694-711.

Wendel, J., & Doyle, J. 1998. Phylogenetic incongruence: window into genome history and speciation. In: P.S. Soltis, D.E. Soltis, and J. J. Doyle [eds.], Molecular systematics of plants II: DNA sequencing. Kluwer Academic Publishers. New York, USA.

Whitten, W., Williams, N. & Chase, M. 2000. Subtribal and generic relationships of Maxillariaceae (Orchidaceae) with emphasis on Stanhopeinae: combined molecular evidence. *American Journal of Botany.* **87**: 1842-1856.

Wojciechowski, M., Lavin, M. & Sanderson, M. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *mat*K gene resolves many well-supported subclades within the family. *American Journal of Botany* **91**: 1846 – 1862.

Wolfe, K. 1991. Protein-coding genes in chloroplast DNA: compilation of nucleotide sequences, data base entries and rates of molecular evolution. In: Vasil [ed] Cell Culture and Somatic Cell Genetics of Plants, Vol. 7BI, pp. 467-482. Academic Press, San Diego, USA.

Wolfe, K., Morden, C. & Palmer, J. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences* **89**: 10648-10652.

Yokota, Y., Kawata, T., Iida, Y., Kato, A. & Tanifuji, S. 1989. Nucleotide sequences of the 5.8S rRNA gene and the internal transcribed spacer regions in carrots and broad bean ribosomal DNA. *Journal of Molecular Evolution* **29**: 294-301.

APPENDIX I. Voucher specimen and locality information.

[1]AEHS = A.E.H. Salles, AMC = A.M. de Carvalho, B & C = Brond & Cogollo, BH = B. Hoffman, BVR = B.V. Rabelo, CEH = C.E. Hughes, D & L = Delgado & Lavin, D & M = Davidse & Miller, DC = D. Cardoso, DCo = D. Coomes, DMN= D.M. Neves, GAK = M. Gardener & S. Knees, H & L = Holst & Liesner, HCL = H.C. de Lima, JGC-S = J.G. Carvalho-Sobrinho, JJ = J. Jardin, JM = J. Magalanes, LPQ = L.P. de Quieroz, LR = L. Rico, MC = M. Cheek, MS = M. Sugiyama, RTP = R.T. Pennington, SB = S. Bridgewater, TDP = T.D. Pennington, WR = W. Rodrigues.

[2]Source of data for the phylogenetic analysis: sequence = sequences previously generated by Alexandra Clark and Cynthia Skema; DNA = dried leaves, GenBank = sequences taken from GenBank followed by the accession number.

[3]CEPEC = Centro de Pesquisas do Cacau, Bahia Brazil; K = Royal Botanic Garden, Kew, UK; FHO = Forest Herbarium, Oxford, UK; INPA, Instituto de Pesquisas de Amazonia, Manaus, Amazonas, Brazil; E = Royal Botanic Garden Edinburgh; UB = Universidade de Brasília, Distrito Federal, Brazil; U = National Herbarium Nederland, Utrecht, Nehterlands; US = Smithsonian Institution, Washington, D.C., USA.

[4]Information about the locality of collection for many accessions was not available.

+ Denotes an accession that was removed from the matrix.

			DNA Region				
Species	**Collector[1]**	**Number**	**ITS[2]**	***mat*K[2]**	***trn*L[2]**	**Herbarium[3]**	**Locality of collection[4]**
Andira anthelmia	RTP	227	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira anthelmia	JJ	568	Sequence	DNA		CEPEC	Bahia, Brazil
Andira anthelmia	LPQ	13832			GenBank JX275930.1	HUEFS	Bahia, Brazil
Andira carvalhoi	LPQ	13835		GenBank JX295958.1/JX295959.1	GenBank JX275925.1	HUEFS	Bahia, Brazil
Andira carvalhoi	RTP	233	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira cordata	RTP	264	Sequence	DNA	DNA	CEPEC, FHO, K	Bahia, Brazil
Andira cujabensis	DMN	1889		Sequence			Brazil

Andira cujabensis	RTP	503	Sequence	DNA	DNA	E, UB	Goias, Brazil
Andira fraxinifolia	RTP	236	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira fraxinifolia	MS	889	Sequence	DNA	DNA	K, SP	Sao Paulo, Brazil
Andira galeottiana	D & L	8214	Sequence				Veracruz, Mexico
Andira galeottiana	Lavin	8214		GenBank AF142681.1	GenBank AF208893.1		
Andira galeottiana+	LR	*s.n.*	Sequence			K	Oaxaca, Mexico
Andira grandistipula	BH	1992	Sequence	DNA	DNA	US, U	Guyana
Andira humilis	DC	2447		GenBank JX295960.1/ JX295961.1	GenBank JX275924.1	HUEFS	Minas Gerais, Brazil
Andira humilis	RTP	239	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira humilis	RTP	246	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira humilis	RTP	268	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira humilis	RTP	269	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira inermis	CEH	1673	Sequence			FHO, K	Oaxaca, Mexico
Andira inermis	GAK	6619	Sequence			E	St. Andrew, Trinidad
Andira inermis	LA	8239		GenBank GQ429072.1			
Andira inermis	MC	3579	Sequence		DNA	K	SW Province, Cameroon
Andira inermis	RTP	589	Sequence			E, INBIO	Puntarenas, Costa Rica
Andira inermis	SB	347	Sequence			E	Orange walk, Belize
Andira inermis	TDP	13358	Sequence			K	Puntarenas, Costa Rica
Andira inermis	Bridgewater	347		GenBank JF501102.1			Belize
Andira jaliscensis	JM	4404	Sequence	DNA	DNA		Jalisco, Mexico
Andira legalis	HCL	7480		Sequence			Brazil
Andira legalis	DC	2214		GenBank JX295893.1	GenBank JX275923.1	HUEFS	Cultivated, Rio de Janeiro, Brazil
Andira legalis	RTP	307	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira legalis+	LFGS	51		GenBank JF491264.1			Brazil
Andira macrothyrsa	RTP	523	Sequence			E, QCNE	Napo, Ecuador
Andira macrothyrsa	RTP	1207	Sequence	DNA	DNA	E	Loreto, Peru
Andira macrothyrsa	TDP	13550	Sequence			K	Loreto, Peru

Andira marauensis	AMC	*s.n.*	Sequence			CEPEC	Bahia, Brazil
Andira marauensis	LPQ	13857		GenBank JX295899.1	GenBank JX275929.1	HUEFS	Bahia, Brazil
Andira micrantha	WR	11180	Sequence	DNA	DNA		
Andira micrantha	KGD	6889		Sequence			Brazil
Andira multistipula	RTP	537	Sequence	DNA	DNA	E	Napo, Ecuador
Andira multistipula+	KGD	5542		Sequence			
Andira nitida	AMC	3309	Sequence	DNA		CEPEC	Bahia, Brazil
Andira nitida	LPQ	13860			GenBank JX275922.1	HUEFS	
Andira nitida	RTP	292	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira nitida	RTP	301	Sequence			CEPEC, FHO, K	Bahia, Brazil
Andira nitida+	HCL	7479		Sequence			Brazil
Andira ormosioides	DC	2221		GenBank JX295962.1/JX295963.1	GenBank JX275927.1	HUEFS	
Andira ormosioides	HCL	4831	Sequence			RB	Rio de Janeiro, Brazil
Andira parviflora	KGD	6920		Sequence			Brazil
Andira parviflora	WR	11179	Sequence			INPA, K	Manaus, Brazil
Andira praecox	BVR	3199	Sequence			K	Amapa, Brazil
Andira surinamensis	GAK	6630	Sequence			E	St. Andrew, Trinidad
Andira surinamensis	RTP	433	Sequence	DNA	DNA	FHO, K, U, US	Guyana
Andira surinamensis	RTP	463	Sequence			FHO, K, U, US	Guyana
Andira taurotesticulata	RTP	525	Sequence	DNA	DNA	E	Napo, Ecuador
Andira tervequinata	H&L	20674	Sequence			E	Bolivar, Venezuela
Andira trifoliolata	DCo	81	Sequence			K	Atabapo, Venezuela
Andira unifoliolata	WR	11186	Sequence			INPA, K	Manaus, Brazil
Andira vermifuga	RTP	906		Sequence			
Andira vermifuga	RTP	271	Sequence	DNA	DNA	CEPEC, FHO, K	Bahia, Brazil
Hymenolobium alagoanum	DMN	1205	Silica dried	Silica dried	Silica dried	HUEFS, K, RB	Reserva Natural Vale Linhares, Espírito Santo, Brazil
Hymenolobium alagoanum	HCL	7522		Sequence			Brazil
Hymenolobium alagoanum	JGC-S	2939	DNA	GenBank JX295906.1	GenBank JX275936.1	HUEFS	Bahia, Brazil

Hymenolobium alagoanum	RTP	224	DNA	DNA	DNA	E	Bahia, Brazil
Hymenolobium excelsum	HCL	7443		Sequence			Brazil
Hymenolobium flavum	RTP	451	DNA	DNA	DNA	K	Guyana
Hymenolobium flavum+	DMN	2021		Sequence			Brazil
Hymenolobium grazielanum	HCL	3266	DNA	GenBank JX295907.1	GenBank JX275937.1	HUEFS	Amazonas, Sao Gabriel da Cachoeira, Brazil
Hymenolobium heringerianum	AEHS	4265	DNA	GenBank JX295910.1	GenBank JX275940.1	HUEFS	Distrito Federal, Brasilia, Brazil
Hymenolobium heringerianum	HCL	7462	Silica dried	Silica dried	Silica dried		Brazil
Hymenolobium heringerianum	HCL	7460		Sequence			
Hymenolobium heterocarpum	LPQ	13899	DNA	GenBank JX295901.1	GenBank JX275931.1	HUEFS	Manaus, Amazonas, Brazil
Hymenolobium heterocarpum	LPQ	13905		GenBank JX295902.1		HUEFS	Manaus, Amazonas, Brazil
Hymenolobium janeirense var. janeirense	DC	2192	DNA	GenBank JX295904.1	GenBank JX275934.1	HUEFS	Rio de Janeiro, Brazil
Hymenolobium janeirense var. stipulatum	DC	2266	DNA	DNA	GenBank JX275935.1	HUEFS	Santa Terezinha, Bahia, Brazil
Hymenolobium mesoamericanum	RTP	614	DNA	DNA	DNA	E	La Selva Biological Station, Costa Rica
Hymenolobium modestum	KGD	6878		Sequence	DNA		Manaus, Amazonas, Brazil
Hymenolobium modestum+	HCL	6813			GenBank JX275938.1	RB	Terra Santa, Para Brazil
Hymenolobium nitidum	WR	11177	DNA	DNA	DNA	INPA, K	Manaus, Amazonas, Brazil
Hymenolobium nitidum+	TB	HD998		Sequence			Peru
Hymenolobium cf. petraeum	DMN	2000		Silica dried	Silica dried	HUEFS, K, RB	Estacao Cientifica Ferreira Penna, Para, Brazil
Hymenolobium petraeum	DC	2860	DNA	GenBank JX295909.1	GenBank JX275939.1	HUEFS	Roraima, Brazil
Hymenolobium pulcherrimum	WR	11181	DNA	DNA	DNA		
Hymenolobium pulcherrimum+	KGD	6973		Sequence			Brazil
Hymenolobium sericeum	DC	2901	DNA	GenBank JX295903.1	GenBank JX275933.1	HUEFS	Amazonas, Brazil
Hymenolobium sericeum+	HCL	7157			GenBank JX275932.1	RB	Parauapebas, Para, Brazil
Hymenolobium sp	TDP	16995	DNA	GenBank JQ619987.1	DNA	K	Loreto, Peru
Hymenolobium sp	DMN	1965	Silica dried	Silica dried	Silica dried	HUEFS, K, RB	
Hymenolobium sp	WR	11183		DNA	DNA		Manaus, Brazil

APPENDIX II. Pennington's (2003) phylogenetic reconstruction of *Andira* based on cpDNA restriction site and morphological data.