Erzincan Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
2021, 14(1), 270-283
ISSN: 1307-9085, e-ISSN: 2149-4584
Araştırma Makalesi

Erzincan University
Journal of Science and Technology
2021, 14(1), 270-283
DOI: 10.18185/erzifbed.786884
Research Article

An in-silico approach for the evaluation of six DNA barcodes by using Maximum Likelihood (ML) and TaxonDNA approaches for some Polygalaceae

Deniz Aygören Uluer[1*]

[1]Ahi Evran University, Çiçekdağı Vocational College, Department of Plant and Animal Production, Çiçekdağı, Kırşehir/Turkey

Geliş / Received: 28/08/2020, Kabul / Accepted: 08/02/2021

Abstract

Polygalaceae is a large family with a cosmopolitan distribution, comprising ca. 1,200 species in 27 genera. However, similar to many plant groups, the identification of Polygalaceae species mostly depends on floral and fruit characteristics; therefore, DNA barcoding could easily warrant the correct identification of sterile material. In the current study, the utility of six widely employed plant DNA barcode loci, namely rbcL, matK, trnL-F region (including trnL intron+trnL-F intergenic spacer), the entire ITS (ITS1+5.8S+ITS2) as well as subunits ITS1 and ITS2 have been explored by performing Maximum Likelihood (ML) and TaxonDNA analyses. The results have shown that, while none of the six loci reviewed here completely fulfils the ideal DNA barcoding criteria; yet, matK region is the most useful DNA barcode for the Polygalaceae.

Keywords: DNA barcoding, *ITS*, *mat*K, Polygalaceae, *rbc*L, *trn*L-F.

Altı DNA bölgesinin bazı Polygalaceae'de Maximum Likelihood (ML) ve TaxonDNA yöntemleri ile in slico olarak değerlendirilmesi

Öz

Polygalaceae kozmopolit bir yayılıma sahip, 27 cins içinde yaklaşık 1,200 türe sahip geniş bir familyadır. Ancak birçok bitki grubu gibi, Polygalaceae türlerinin tanımlanması genellikle çiçek ve meyva karakterlerine bağlıdır. Bu nedenle DNA barkodlama yontemi steril materyalin doğru tanımlanmasını sağlayabilir. Bu çalışmada altı tane yaygın olarak kullanılan DNA barkodu, *rbçL*, *mat*K, *trn*L-F DNA bölgesi (*trn*L intron+*trn*L-F dahil), tüm *ITS* (ITS1+5.8S+ITS2), ITS1 ve ITS2 DNA bölgelerinin, Maximum Likelihood (ML) ve TaxonDNA yöntemleri ile DNA barkodu olarak performansları değerlendirilmiştir. Sonuçlar göstermiştir ki, altı DNA bölgesinden hiçbiri ideal değildir, ancak Polygalaceae familyası için en uygun DNA barkodu *mat*K gen bölgesidir.

Anahtar Kelimeler: DNA barkodlama, *ITS*, *mat*K, Polygalaceae, *rbc*L, *trn*L-F.

1. Introduction

Polygalaceae is the second largest family in order Fabales with a nearly cosmopolitan distribution and ca. 1,200 species in 27 genera (Pastore et al., 2017). The family is classified into four tribes, Carpolobieae, Moutabeae, Polygaleae and Xanthophylleae. Genus Polygala with ca. 500 species accounts for around half the species in the family; however, Monnina, Muraltia, Securidaca and Xanthophyllum are other

species-rich genera. Similar to many plant groups, the identification of Polygalaceae species mostly depends on floral and fruit characteristics; therefore, DNA barcoding could easily warrant the correct identification of sterile material (Aygoren Uluer and Alshamrani, 2019).

DNA barcoding is a novel, cost-effective and rapid taxonomic method to identify organisms by the use of short-standardized gene region(s) (Hebert et al., 2003a, b) where morphological identification is challenging (e.g., Shi et al., 2011) or not possible due to the condition of the material (i.e., the growth phase of the plant, sterile materials) (e.g., Liu et al., 2018). Moreover, this technique is also helpful in identifying cryptic species, controlling the traffic of endemic and endangered species (i.e., conservation) (e.g., Hebert et al., 2004; Hajibabaei et al., 2007), discriminating herbal medicine and food ingredients from their cheaper substitutes (i.e., adulterants) (e.g., Xin et al., 2013), tracing dietary components of animals (e.g., Kartzinel et al., 2015), detecting misidentified species (Han et al., 2010; Kuzmina et al., 2012), in conservation, forensic investigations and biodiversity inventories; for not only professional taxonomists but also for non-experts (e.g., customs officers and forensic specialists). Therefore, even a partially effective DNA barcode would be beneficial in many areas (Chen et al., 2010).

An ideal DNA barcode should be a single locus, relatively short (~700 bp), easily amplifiable with a single pair of universal primers (i.e., has relatively conserved flanking regions) and standard PCR conditions, easily sequencible and alignible across lineages with little manual editing, able to provide high discrimination power with high interspecific and less intraspecific variation (Kress et al., 2005; Cowan et al., 2006; Newmaster et al., 2006). However, in contrast to animals, a universal plant DNA barcode could not be designated for plants to date due to several problems, such as the number of informative characters, extent hybridization, introgression and genome duplication (Hollingsworth et al., 2011). Therefore, several combinations of markers, both coding and non-coding, from nuclear and plastid genomes have been proposed to date (Kress and Ericson, 2007). In many cases, instead of only one region, a combination of at least two regions or employing nuclear regions with higher substation rates has found to be more efficient (Hajibabaeei et al., 2007).

While the Consortium for the Barcode of Life (CBOL) (2009) has recommended the use of matK+rbcL as the standard (i.e., core) DNA barcode over five other coding and non-coding regions (i.e., namely rpoC1, rpoB, psbA–trnH, psbK–psbI, and atpF–atpH), the barcode only has 72% discrimination power in many plant groups (CBOL Plant Working Group, 2009), and is not discriminative enough in particularly woody groups (Clement and Donoghue, 2012). Therefore, mainly due to this low species discrimination of the first proposed regions, PCR and primer problems, the use of supplementary DNA barcodes or replacement of standard barcodes with supplementary DNA barcodes has been suggested (Hollingsworth et al., 2011; Clement and Donoghue, 2012). Later, the use of some of these regions, such as the internal transcribed spacers of nuclear ribosomal DNA (nr ITS/ITS1/ITS2) and matK have overcome the use of standard barcode combination, namely matK+rbcL.

In the current study, the utility of DNA barcoding of the Polygalaceae family has been explored by performing Maximum Likelihood (ML) and TaxonDNA (Meier et al., 2006) analyses with six popular DNA barcodes, namely the entire ITS, ITS1, ITS2, rbcL, matK and trnL-F region.

2. Materials and Methods

2. 1. Sequence editing, alignment, and phylogenetic analyses

The rbcL, matK, trnL-F region (including trnL intron+trnL-F intergenic spacer), the entire ITS as well as subunits ITS1 and ITS2 sequences of Polygalaceae were obtained from GenBank. Where possible, the sequences from the same voucher specimen were used. Sequences were assembled and aligned using the Geneious alignment option in Geneious Pro 4.8.4 (Kearse et al., 2012). All indels were scored as missing data.

The discriminatory power for all DNA regions was evaluated at genus level (please note that since only one sequence from most of the Polygalaceae species was deposited in GenBank, instead of species level, genus level comparisons were employed) by performing two analytical methods, the sequence similarity method (TaxonDNA, Meier et al., 2006) and phylogenetic analyses (Maximum Likelihood, ML method).

For these analyses, two different matrices have been used: 1) for the ML analyses, the matK data matrix comprised 51 sequences, the rbcL data matrix comprised 58 sequences, the trnL-F region matrix contained 73 sequences, the entire ITS data matrix comprised 64 sequences, both ITS1 and ITS2 matrices contained 62 sequences, from 109 individuals of 18 genera of Polygalaceae and one outgroup (Table 1). The National Center for Biotechnology Information (NCBI/ GenBank) accession numbers for these DNA sequences are provided in Appendix 1. 2) For the TaxonDNA analyses, all subsequent DNA sequences were downloaded for each DNA region (i.e., all sequences deposited in GenBank for each region), for all Polygalaceae genera. The matK data matrix comprised 216 sequences, the rbcL data matrix comprised 362 sequences, the trnL-F region matrix contained 353 sequences, the entire ITS data matrix comprised 671 sequences, the ITS1 and ITS2 matrices contained 589 and 655 sequences, respectively (Table 1).

ML analyses were performed using RAxML version 7.0.4 (Stamatakis et al., 2008). Arachis hypogea was defined as the outgroup. Default ML search options were selected with 100 bootstrap replicates. Cut-off of 50% was used to define support for “successful” resolution of monophyletic genera.

Table 1. Alignment length of the ingroup taxa and the number of total individuals sampled for the Polygalaceae family, for *mat*K, *rbc*L, *trn*L intron and *trn*L-F intergenic spacer, *ITS*, ITS1 and ITS2 for the TaxonDNA and Maximum Likelihood (ML) analyses.

	TaxonDNA Analyses		ML Analyses	
Region	Aligned ingroup length (bp)	Total individuals	Aligned ingroup length (bp)	Total individuals
*rbc*L	1404	362	1399	58
*mat*K	2412	216	1944	51
*trn*L+ *trn*L-F	1666	353	1132	73
ITS	1039	671	1047	64
ITS1	421	589	458	62
ITS2	433	655	294	62

For the TaxonDNA analyses, the computer software TaxonDNA (Meier et al., 2006) was used to calculate intra- and interspecific variation within the data set and to evaluate the potential of these loci to serve as a DNA barcode. To test whether accurate species assignments can be made among Polygalaceae samples, "best match" and "best close match" functions of TaxonDNA were explored. Six data sets, namely, *rbc*L, *mat*K, *trn*L-F region, the entire *ITS*, ITS1 and ITS2, were analyzed by using a minimum overlap of 300 bp. Since the total evidence tree (i.e., *rbc*L+ *mat*K+ *trn*L-F region+*ITS* tree), was not different than the only *mat*K tree, this was not included in the TaxonDNA analyses.

3. Results

For the ML analyses, the outgroup-excluded alignment of *rbc*L was 1399 bp long, outgroup-excluded alignments of *mat*K were 1944 bp, *trn*L-F region was 1132 bp, *ITS* was 1047 bp, ITS1 was 458 and ITS2 was 294 bp. On the other hand, for the TaxonDNA analyses, the outgroup-excluded alignment of *rbc*L was 1404 bp long, outgroup-excluded alignment of *mat*K was 2412 bp, *trn*L-F region was 1666 bp, *ITS* was 1039 bp, ITS1 was 421 and ITS2 was 433 bp. All these alignment details for all datasets are summarized in Table 1.

3. 1. Phylogenetic analyses

Polygalaceae family was monophyletic in all ML analyses; however, in terms of retrieving monophyletic genera, while ITS1 and ITS2 were the least successful DNA regions (four and three non-monophyletic genera, respectively), followed by the entire *ITS* and *trn*L-F region (one non-monophyletic genus for each) (Table 2). The *mat*K and the *rbc*L regions did not yield any non-monophyletic genera; however, it was noteworthy that compared to the *mat*K region, *rbc*L region mostly yielded lower support values. Combining all regions definitely resulted in higher support values (please note that *Polygala*, *Bredemeyera* and *Moutabeae*, were non-monophyletic in the previous studies such as Forest et al. (2007); Pastore et al., (2017)) (Table 2, Figure 1).

Table 2. Identification success of *mat*K, *rbc*L, *trn*L-F region (including *trn*L intron and *trn*L-F intergenic spacer), *ITS*, ITS1 and ITS2 using maximum likelihood (ML) method (i.e., monophyletic genera in the ML trees). Crosses (X) indicate the species was not monophyletic in the ML tree. Empty cells indicate that data are not available.

	*mat*K	*rbc*L	*trn*L+*trn*L-F	*ITS*+*mat*K+*rbc*L+ *trn*L +*trn*L-F	*ITS*	ITS1	ITS2
Acanthocladus			94%	99%			
Atroxima		94%	86%	97%	100%	97%	92%
Badiera			91%	65%	100%	89%	99%
Bredemeyera	99%		-	-	-	-	79%
Carpolobia	50%	97%	66%	95%	100%	97%	98%
Comesperma		95%	-	51%	53%	-	-
Diclidanthera			94%	92%			
Epirixanthes	99%			100%	100%	99%	99%
Monnina	97%	100%	99%	98%	90%	-	-
Moutabea		68%	88%	92%			
Muraltia	100%	99%	100%	100%	94%	94%	88%
Salomonia	100%		100%	100%	100%	100%	100%
Securidaca	100%	51%	95%	100%	-	-	-
Xanthophyllum	100%	72%	87%	100%	100%	-	76%
Polygala	-	-	-	-	-	-	-

3. 2. TaxonDNA analyses

For the “best match” and “best close match” criteria; most successfully identified/least ambiguous results were obtained from the entire *ITS* region (40.08% and 36.95%, respectively), following by ITS1 (36.08% and 30.58%, respectively) and ITS2 (36.78% and 33.84%, respectively) (Table 3). In contrast to the *mat*K plastid region (5.09%), the number of ambiguous sequences was very high for these three regions (between 11.02% and 24.26%). Moreover, *rbc*L and *trn*L-F regions yielded the lowest successfully identified and highest ambiguous matching results according to the “best match” and “best close match” criteria.

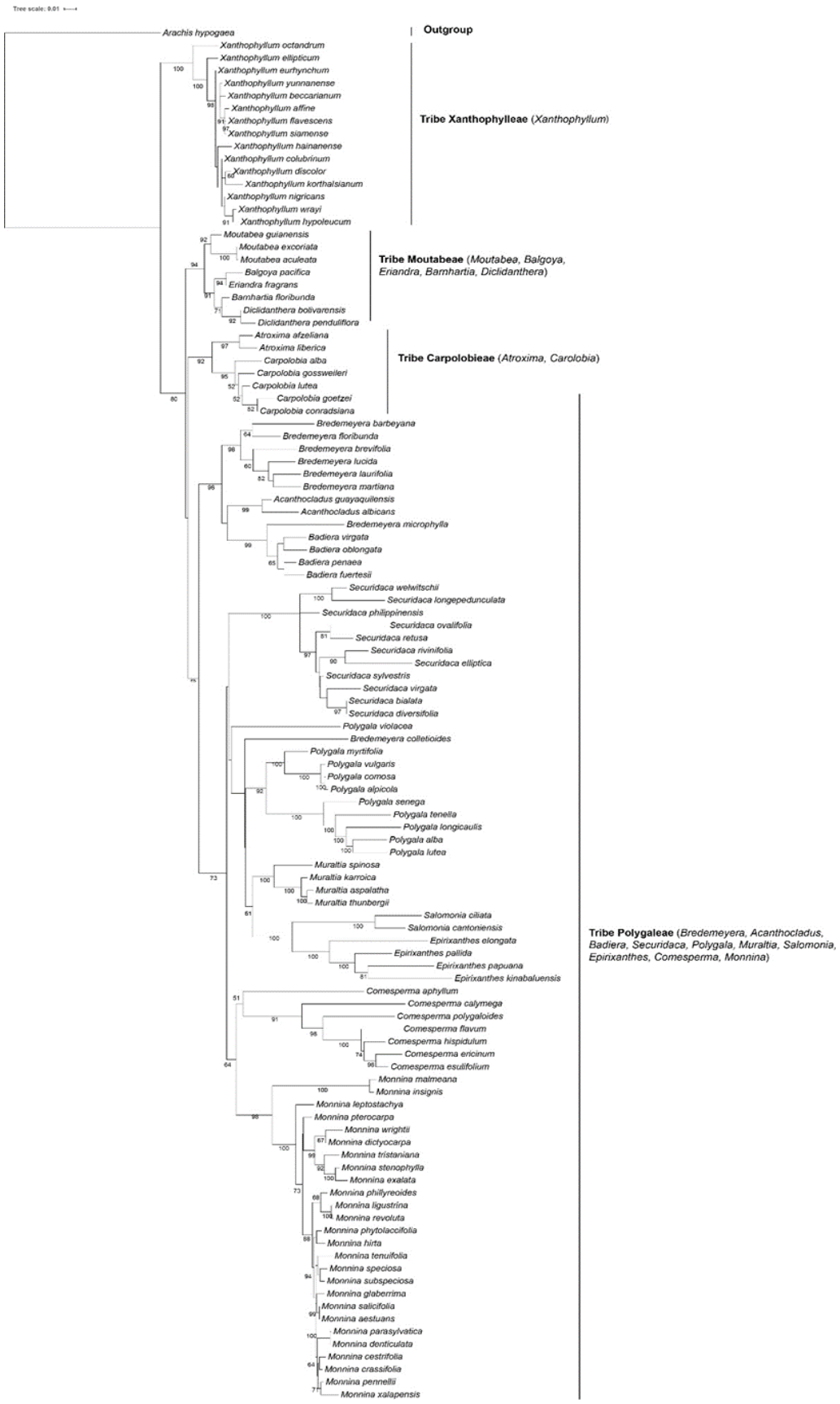

Figure 1. Phylogenetic relationships within Polygalaceae inferred from maximum likelihood analysis of *mat*K+*rbc*L+*ITS*+*trn*L-F region. Bootstrap values are indicated below branches. Four tribes of the Polygalaceae family, Carpolobieae, Moutabeae, Polygaleae and Xanthophylleae are shown.

Table 3. Identification success based on the "best match, best close match, ambiguous and misidentified" according to TaxonDNA (Best match= correct identifications according to names regardless of the similarity of sequences; best close match=correct identifications according to names and similarity of sequences; Ambiguous= several equally good best matches; Misidentified= wrong identifications when the names are mismatched) (Meier et al., 2006).

	Best Match			Best Close Match		
Region	Successfully identified	Ambiguous	Misidentified	Successfully identified	Ambiguous	Misidentified
*rbc*L	19.33%	40.05%	40.60%	19.33%	40.05%	39.77%
*mat*K	31.48%	5.09%	63.42%	26.85%	5.09%	48.14%
*trn*L+*trn*L-F	12.94%	26.58%	60.47%	12.70%	25.17%	55.29%
ITS	40.08%	12.07%	47.83%	36.95%	11.02%	26.52%
ITS1	36.08%	16.49%	47.42%	30.58%	12.71%	20.79%
ITS2	36.78%	24.26%	38.94%	33.84%	20.55%	23.49%

4. Discussion

In the present study, *rbc*L plastid gene coding region was easily alignable and the number of monophyletic genera in the ML analysis was promising; however, it was noteworthy that, compared to the *mat*K region, the support values of these monophyletic genera were not high (Table 2). Furthermore, according to the "best match" and "best close match" criteria of the TaxonDNA analyses (Meier et al., 2006), the region yielded the second-lowest correct identification rate, after the *trn*L-F region (Table 3). Similarly, both the ML and TaxonDNA analyses with the *trn*L-F region (Table 2 and 3) showed that the identification success of the region is limited. Moreover, the region was not easily alignable, which is one of the desired characteristics of an ideal DNA barcode (Kress et al., 2005). While both the *rbc*L and *trn*L-F regions show high amplification and sequencing success rate with universal primers across many plant species (Taberlet et al., 2007; CBOL Plant Working Group, 2009); yet, the results of this study have shown that neither *rbc*L nor *trn*L-F is not suitable for Polygalaceae DNA barcoding.

On the other hand, due to its rapidly evolving nature, the *ITS* region (ITS1+5.8S+ITS2) has been used particularly in low-level plant taxonomic studies (Baldwin, 1992); yet, many problems related to the entire *ITS* region have also been reported, such as fungal contamination, cloning requirement in some cases, incomplete concerted evolution, gene conversions, poor PCR and sequencing success, lack of universal primers and having several INDELS which causes misleading phylogenetic information (Alvarez and Wendel, 2003; Kress et al., 2005; Cowan et al., 2006; Chase et al., 2007; CBOL Plant Working Group, 2009; Hollingsworth et al., 2011). Therefore, short internal transcribed spacers with ease of amplification and sequencing, ITS1 and ITS2, were suggested useful DNA barcoding candidates in many plant groups. (Han et al., 2013; Zhu et al., 2018).

ITS1 and ITS2 sub-regions are one of the most widely used phylogenetic markers at both species and genus levels, because of their high copy number, high rate of evolution and reticulate evolution, which allow researchers to amplify the region easily and reveal not only the phylogenetic relationships within a plant group but also the phylogenetic patterns such as hybridization and homoplasy (Baldwin et al., 1995; Yao et al., 2010; Marghali et el. 2015).

However, in the current study, while the entire *ITS*, ITS1 and ITS2 regions yielded the highest correct identification rates, according to the "best match" and "best close match" criteria of the TaxonDNA analyses (Meier et al., 2006) (Table 3); yet neither the entire *ITS*, nor the ITS1 and ITS2 subunits were efficient in species discrimination owing to the high number of non-monophyletic genera in the ML trees (Table 2). Furthermore, due to several insertions/deletions (Schlötterer et al., 1994), the alignment of these two subunits was not easy as *rbc*L or *mat*K. Therefore, it is clear that the entire *ITS*, ITS1 and ITS2 subunits are not ideal DNA barcoding candidates for the Polygalaceae family.

While plastid-core barcode Intron Group II maturase *mat*K, is one of the most rapidly evolving coding regions in plants (CBOL Plant Working Group, 2009), and shows a high level of species discrimination ability among angiosperms (Lahaye et al., 2008); yet, it is well known that both the length of the *mat*K coding region and not perfectly conserved primer-binding sites across plants make the amplification and sequence recovery difficult with universal primers and universal PCR conditions for most of the plant groups, as well as working with degraded materials (Chase et al., 2007; Newmaster et al., 2008; CBOL Plant Working Group, 2009; Parveen et al., 2016; Kuzmina et al., 2012). In the present study, the ML analysis yielded the best results, in terms of the number of monophyletic genera, next to the *rbc*L+*mat*K+*ITS*+*trn*L-F analysis (Table 2). Moreover, the TaxonDNA (Meier et al., 2006) analyses showed that the correct identification rate of this plastid coding region is close to the ITS1 and ITS2 subunits, with a lower ambiguous identification rate, compared to the entire *ITS*, ITS1 and ITS2 subunits (Table 3). Furthermore, the region was easy to align. Therefore, I suggest that, among the six possible DNA barcodes here, the plastid *mat*K region is the best candidate for the Polygalaceae.

Last but not least, combining *rbc*L, *mat*K, *trn*L-F and *ITS* regions did not yield a better resolved (i.e., number of monophyletic genera) ML tree (Table 2, Figure 1) (please note that, due this reason, *rbc*L+*mat*K+*trn*L-F+*ITS* was not included in the TaxonDNA analyses). Therefore, similar to the *rbc*L gene, sequencing these four regions would be resource waste.

In summary, as a result of the ML and TaxonDNA analyses of the current study, while none of the six loci reviewed here completely fulfils the ideal DNA criteria for Polygalaceae, I recommend using the *mat*K region as a DNA barcode for *Polygalaceae*; however, if there are amplification and/or sequencing problems, sequencing only the entire *ITS* region; or if a particular species of the family is in interest, adding *ITS* region might be helpful. Yet, caution must be handled

when working with the *ITS* region, because as a result of gene duplication and incomplete concerted evolution, the existence of paralogous sequences could still be a problem (Alvarez and Wendel 2003).

5. Acknowledgements

I am grateful to the anonymous reviewers for their helpful comments.

6. Reference List

Alvarez, I., and Wendel, J.F. 2003. Ribosomal *ITS* sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution,* 29(3):417–434. doi:10.1016/S1055-7903(03)00208-2.

Aygoren Uluer, D. and Alshamrani, R., 2019. DNA barcoding of a complex genus, *Aesculus* L. (Sapindaceae) reveals lack of species-level resolution. *Botany*, 97(9), 503-512.

Baldwin, B.G. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. *Molecular Phylogenetics and Evolution,* 1(1): 3–16. doi:10.1016/1055-7903(92)90030-K.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America,* 106(31): 12794–12797. doi: 10.1073/pnas.0905845106.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., van den Berg, C., Madrinan, S., Petersen, G. et al. 2007. A proposal for a standardised protocol to barcode all land plants. *Taxon,* 56(2): 295–299. doi:10.1002/tax.562004.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L. et al. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS One*, 5(1): e8613. doi:10.1371/journal.pone.0008613.

Clement, W.L., and Donoghue, M.J. 2012. Barcoding success as a function of phylogenetic relatedness in *Viburnum*, a clade of woody angiosperms. *BMC Evolutionary Biology,* 12(1): 73. doi:10.1186/1471-2148-12-73.

Cowan, R.S., Chase, M.W., Kress, W.J., and Savolainen, V. 2006. 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon*, 55(3): 611–616. doi:10.2307/25065638.

Forest, F., Chase, M.W., Persson, C., Crane, P.R. and Hawkins, J.A., 2007. The role of biotic and abiotic factors in evolution of ant dispersal in the milkwort family (Polygalaceae). *Evolution,* 61(7), 1675-1694.

Hajibabaei, M., Singer, G.A., Hebert, P.D., and Hickey, D.A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics,* 23(4): 167–172. doi:10.1016/j.tig.2007.02.001.

Han, J., Liu, C., Li, M., Shi, L., Song, J., Yao, H., Pang, X. and Chen, S., 2010. Relationship between DNA barcoding and chemical classification of *Salvia* medicinal herbs. *Chinese Herbal Medicines*, 2(1), 16-29.

Han, J., Zhu, Y., Chen, X., Liao, B., Yao, H., Song, J., Chen, S. and Meng, F., 2013. The short ITS2 sequence serves as an efficient taxonomic sequence tag in comparison with the full-length *ITS*. *BioMed Research international*, 2013.

Hebert, P.D., and Ratnasingham, S., and deWaard, J.R. 2003a. Barcoding animal life: *cytochrome c oxidase subunit 1* divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, 270 (Suppl 1): S96–S99. doi:10.1098/rsbl.2003.0025.

Hebert, P.D., Cywinska, A., and Ball, S.L., and deWaard, J.R. 2003b. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270(1512): 313–321. doi:10.1098/rspb. 2002.2218.

Hebert, P.D., Penton, E.H., Burns, J.M., Janzen, D.H., and Hallwachs, W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41): 14812–14817. doi:10.1073/pnas.0406166101.

Hollingsworth, P.M., Graham, S.W. and Little, D.P., 2011. Choosing and using a plant DNA barcode. *PloS One*, 6(5), p.e19254.

Kartzinel, T.R., Chen, P.A., Coverdale, T.C., Erickson, D.L., Kress, W.J., Kuzmina, M.L., Rubenstein, D.I., Wang, W. and Pringle, R.M., 2015. DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences of the United States of America*, 112(26), 8019-8024.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S. et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12): 1647– 1649. doi:10.1093/bioinformatics/bts199.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., and Janzen, D.H. 2005. Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23): 8369–8374. doi: 10.1073/pnas.0503123102.

Kress, W.J. and Erickson, D.L., 2007. A two-locus global DNA barcode for land plants: the coding *rbc*L gene complements the non-coding *trnH-psbA* spacer region. *PloS One*, 2(6), p.e508.

Kuzmina, M.L., Johnson, K.L., Barron, H.R. and Hebert, P.D., 2012. Identification of the vascular plants of Churchill, Manitoba, using a DNA barcode library. *BMC Ecology*, 12(1), 1-11.

Lahaye, R., Van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G. and Savolainen, V., 2008. DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8), 2923-2928.

Liu, J., Milne, R.I., Möller, M., Zhu, G.F., Ye, L.J., Luo, Y.H., Yang, J.B., Wambulwa, M.C., Wang, C.N., Li, D.Z. and Gao, L.M., 2018. Integrating a comprehensive DNA barcode reference

library with a global map of yews (*Taxus* L.) for forensic identification. *Molecular Ecology Resources*, 18(5), 1115-1131.

Marghali, S., Fadhlaoui, I., Gharbi, M., Zitouna, N. and Trifi-Farah, N., 2015. Utility of ITS2 sequence data of nuclear ribosomal DNA: Molecular evolution and phylogenetic reconstruction of Lathyrus spp. *Scientia Horticulturae*, 194, 313-319.

Meier, R., Shiyang, K., Vaidya, G. and Ng, P.K., 2006. DNA barcoding and taxonomy in *Diptera*: a tale of high intraspecific variability and low identification success. *Systematic Biology*, 55(5), 715-728.

Newmaster, S.G., Fazekas, A.J., and Ragupathy, S. 2006. DNA barcoding in land plants: evaluation of *rbc*L in a multigene tiered approach. *Botany*, 84(3): 335–341. doi:10.1139/b06-047.

Newmaster, S.G., Fazekas, A.J., Steeves, R.A.D., and Janovec, J. 2008. Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, 8(3): 480–490. doi:10.1111/j.1471-8286.2007.02002.x.

Parveen, I., Gafner, S., Techen, N., Murch, S.J. and Khan, I.A., 2016. DNA barcoding for the identification of botanicals in herbal medicine and dietary supplements: strengths and limitations. *Planta Medica*, 82(14), 1225-1235.

Pastore, J.F.B., Abbott, J.R., Neubig, K.M., Whitten, W.M., Mascarenhas, R.B., Mota, M.C.A. and van den Berg, C., 2017. A molecular phylogeny and taxonomic notes in *Caamembeca* (Polygalaceae). *Systematic Biology*, 42(1), 54-62.

Shi, L.C., Zhang, J., Han, J.P., Song, J.Y., Yao, H., Zhu, Y.J., LI, J.C., Wang, Z.Z., Xiao, W., Lin, Y.L. and Xie, C.X., 2011. Testing the potential of proposed DNA barcodes for species identification of Zingiberaceae. *Journal of Systematics and Evolution*, 49(3), 261-266.

Schlötterer, C., Hauser, M.T., von Haeseler, A. and Tautz, D., 1994. Comparative evolutionary analysis of rDNA ITS regions in *Drosophila*. *Molecular Biology and Evolution*, 11(3), 513-522.

Stamatakis, A., Hoover, P., and Rougemont, J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, 57(5): 758–771. doi:10.1080/10635150802429642.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. and Willerslev, E., 2007. Power and limitations of the chloroplast *trn L* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, 35(3), e14-e14.

Xin, T., Yao, H., Gao, H., Zhou, X., Ma, X., Xu, C., Chen, J., Han, J., Pang, X., Xu, R. and Song, J., 2013. Super food *Lycium barbarum* (Solanaceae) traceability via an internal transcribed spacer 2 barcode. *Food Research International*, 54(2), 1699-1704.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y. et al. 2010. Use of ITS2 region as the universal DNA barcode for plants and animals. *PloS One*, 5(10): e13102. doi:10.1371/journal.pone.0013102.

Zhu, S., Li, Q., Chen, S., Wang, Y., Zhou, L., Zeng, C. and Dong, J., 2018. Phylogenetic analysis of *Uncaria* species based on internal transcribed spacer (*ITS*) region and ITS2 secondary structure. *Pharmaceutical Biology*, 56(1), 548-558.

7. Appendix I: Taxon sampling for Maximum Likelihood (ML) analyses of Polygalaceae. A dash indicates the region was not sampled.

	GenBank accessions			
Taxon name	**Entire ITS**	***mat*K**	***rbc*L**	***trn*L-F**
Polygalaceae				
Acanthocladus albicans	KU682359.1	KU682399.1	_	KU682396.1
Acanthocladus guayaquilensis	_	_	AM234190.1	AF366972.1
Atroxima afzeliana	GQ888877.1	EU604049.1	AM234175.1	AM234273.1
Atroxima liberica	GQ888878.1	_	AM234174.1	AF366941.1
Badiera fuertesii	GQ888879.1	_	_	GQ889057.1
Badiera oblongata	GQ888880.1	_	_	GQ889059.1
Badiera penaea	GQ888881.1	_	KJ082129.1	GQ889060.1
Badiera virgata	GQ888882.1	_	_	GQ889061.1
Balgoya pacifica	_	_	_	AF366942.1
Barnhartia floribunda	_	_	AM234168.1	AM234271.1
Bredemeyera barbeyana	_	MK956827.1	_	MK956828.1
Bredemeyera colletioides	_	_	AM234171.1	AF366947.1
Bredemeyera martiana	KU682360.1	KU682400.1	_	KU682386.1
Bredemeyera microphylla	_	_	AM234173.1	AF366948.1
Bredemeyera floribunda	GQ888883.1	EU596520.1	EU644699.1	AF366945.1
Bredemeyera brevifolia	MK726264.1	_	_	MK737631.1
Bredemeyera laurifolia	MK726267.1	MK754433.1	_	MK737632.1
Bredemeyera lucida	GQ888884.1	_	_	AF366946.1
Carpolobia alba	GQ888885.1	EU604053.1	AM234176.1	AF366949.1
Carpolobia conradsiana	_	JX517551.1	JX572380.1	_
Carpolobia goetzei	GQ888886.1	_	AM234177.1	AM234274.1
Carpolobia gossweileri	_	_	MN366721.1	_
Carpolobia lutea	_	KC627733.1	KC628046.1	_
Comesperma hispidulum	_	_	AM234178.1	AF366952.1
Comesperma polygaloides	KT220736.1	_	KT207920.1	KT207923.1
Comesperma ericinum	KT220747.1	_	L29492.1	_
Comesperma aphyllum	GQ888887.1	_	_	GQ889066.1
Comesperma calymega	GQ888888.1	_	KT207922.1	GQ889067.1
Comesperma esulifolium	GQ888889.1	EU596516.1	AM234179.1	GQ889068.1
Comesperma flavum	GQ888890.1	_	_	GQ889069.1
Diclidanthera bolivarensis	_	_	_	AF366954.1
Diclidanthera penduliflora	_	_	_	AF366955.1
Eriandra fragrans	GQ888891.1	EU604051.1	AM234170.1	AM234272.1
Epirixanthes elongata	KM982699.1	KR002165.1	_	_
Epirixanthes kinabaluensis	KM982700.1	KR002166.1	_	_
Epirixanthes pallida	KM982709.1	KR002175.1	_	_
Epirixanthes papuana	KM982710.1	KR002176.1	_	_

Monnina aestuans	_	EU604037.1	EU644698.1	_
Monnina cestrifolia	GQ888893.1	_	_	GQ889072.1
Monnina crassifolia	GQ888894.1	_	_	AF366956.1
Monnina denticulata	GQ888895.1	_	_	GQ889074.1
Monnina dictyocarpa	_	_	AM234183.1	AF366961.1
Monnina exalata	KU682362.1	KU682422.1	_	_
Monnina glaberrima	_	EU604039.1	EU644697.1	_
Monnina hirta	GQ888896.1	_	AM234181.1	AF366957.1
Monnina insignis	MK726270.1	KU682425.1	_	MK737634.1
Monnina malmeana	_	_	AM234180.1	AF366960.1
Monnina leptostachya	GQ888897.1	_	AM234182.1	AF366962.1
Monnina ligustrina	GQ888898.1	_	_	GQ889077.1
Monnina parasylvatica	GQ888899.1	_	_	GQ889078.1
Monnina pennellii	_	EU604036.1	EU644696.1	_
Monnina phillyreoides	GQ888900.1	_	AM234185.1	AF366958.1
Monnina phytolaccifolia	_	EU596519.1	EU644695.1	_
Monnina pterocarpa	_	_	AM234186.1	AF366963.1
Monnina revoluta	GQ888901.1	_	_	GQ889080.1
Monnina salicifolia	_	EU604038.1	EU644694.1	_
Monnina speciosa	GQ888902.1	_	_	GQ889081.1
Monnina stenophylla	GQ888903.1	KU682426.1	_	KU682424.1
Monnina subspeciosa	GQ888904.1	_	_	GQ889083.1
Monnina tenuifolia	GQ888905.1	_	_	GQ889084.1
Monnina tristaniana	GQ888906.1	_	_	GQ889085.1
Monnina wrightii	GQ888908.1	_	_	GQ889086.1
Monnina xalapensis	GQ888909.1	EU604047.1	AM234184.1	AM234275.1
Moutabea aculeata	_	_	AM234169.1	AF366964.1
Moutabea excoriata	KU682358.1	KU682421.1	_	KU682385.1
Moutabea guianensis	_	JQ626362.1	JQ625841.2	MK797457.1
Muraltia aspalatha	AJ812638.1	AM889729.1	GQ248649.1	AJ842823.1
Muraltia karroica	AJ812629.1	KR002177.1	_	AJ842815.1
Muraltia spinosa	KM982712.1	KR002178.1	KF724313.1	_
Muraltia thunbergii	AJ812637.1	AM889730.1	GQ248650.1	AJ842822.1
Polygala comosa	MK095538.1	EU362027.1	AM234211.1	GQ889120.1
Polygala tenella	_	EU604030.1	EU644687.1	_
Polygala senega	AJ812649.1	EU604031.1	AM234189.1	AF366992.1
Polygala violacea	GQ889040.1	EU604035.1	EU644686.1	AF366987.1
Polygala alpicola	GQ888923.1	EU604041.1	AM234191.1	AM234277.1
Polygala longicaulis		EU604042.1	EU644688.1	MK797499.1
Polygala alba	GQ888920.1	KT456918.1	KT458054.1	GQ889099.1
Polygala lutea	GQ888982.1	MK986522.1	KJ773769.1	AF366991.1
Polygala myrtifolia	MK976845.1	MK986525.1	AJ829699.1	_
Polygala vulgaris	MT796511.1	MK986551.1	AJ829703.1	GQ889221.1
Salomonia ciliata	KM982715.1	KR002181.1	_	AF366997.1
Salomonia cantoniensis	KM982714.1	KR002180.1	KX527454.1	AF366996.1
Securidaca bialata	_	_	EU644682.1	_

Securidaca diversifolia	GQ889047.1	KJ594020.1	AM234225.1	MK797601.1
Securidaca elliptica	GQ889048.1	_	_	GQ889227.1
Securidaca longepedunculata	GQ889049.1	JX517755.1	JF265595.1	GQ889228.1
Securidaca ovalifolia	GQ889050.1	_	_	GQ889229.1
Securidaca philippinensis	_	KU853082.1	KU853151.1	KU853209.1
Securidaca sylvestris	_	JQ588838.1	JQ593518.1	_
Securidaca retusa	GQ889051.1	EU604029.1	EU644681.1	GQ889230.1
Securidaca rivinifolia	GQ889052.1	_	_	GQ889231.1
Securidaca virgata	GQ889053.1	_	AM234226.1	AF367000.1
Securidaca welwitschii	_	_	AM234227.1	AF367001.1
Xanthophyllum hypoleucum	GQ889054.1	_	_	GQ889233.1
Xanthophyllum wrayi	GQ889055.1	_	MG784899.1	GQ889234.1
Xanthophyllum hainanense	KP092743.1	HQ415290.1	KP094662.1	_
Xanthophyllum flavescens	KR532727.1	AB924997.1	KR530226.1	_
Xanthophyllum siamense	KR532728.1	KR531651.1	KR530229.1	_
Xanthophyllum yunnanense	KR532736.1	KR531655.1	KR530237.1	_
Xanthophyllum affine	_	MH332507.1	AM234228.1	AF367002.1
Xanthophyllum beccarianum	_	KU519700.1	_	_
Xanthophyllum colubrinum	_	AB924705.1	AB925488.1	_
Xanthophyllum discolor	_	KJ709134.1	KJ594934.1	_
Xanthophyllum ellipticum	_	LC151410.1	KU853189.1	KU853248.1
Xanthophyllum eurhynchum	_	KX302358.1	MN592544.1	_
Xanthophyllum korthalsianum	_	MG784957.1	_	_
Xanthophyllum nigricans	_	MG784958.1	MG784903.1	_
Xanthophyllum octandrum	_	JN564163.1	AM234229.1	KC428626.1
Outgroup				
Arachis hypogea	AF156675.2	MH428819.1	U74247.1	AY651848.1