1

2 "Testing the effectiveness of *rbcLa* DNA-barcoding for species discrimination in tropical
3 montane cloud forest vascular plants (Oaxaca, Mexico) using BLAST, genetic distance, and
4 tree-based methods"

5

6

7 Sonia Trujillo-Argueta[1], Rafael F. del Castillo[1+] and Abril Velasco-Murguía[1].

8 [1] Instituto Politécnico Nacional, CIIDIR, Oaxaca de Juárez, Oaxaca, México.

9

10 Corresponding Author:

11 Rafael F. del Castillo

12 Hornos 1003, Sta. Cruz Xoxocotlán, Oaxaca, 71230, México

13 Email address: fsanchez@ipn.mx

14

15

Commented [MH1]: I found this region name inconsistent through the manuscript. Consider change it into *rbc*La with italic only for "rbc".

16 **Abstract**

17 DNA-barcoding is a species identification tool that uses a short section of the genome that
18 provides a genetic signature of the species. The main advantage of this novel technique is
19 that it requires a small sample of tissue from the tested organism. In most animal groups,
20 this technique is very effective. In plants, however, the recommended standard markers,
21 such as *rbcLa,* may not always work, and their effects remain to be tested in many plant
22 groups, particularly from the Neotropical region. We examined the discriminating power
23 of *rbcLa* in 55 tropical cloud forest vascular plant species from 38 families (Oaxaca,
24 Mexico). We followed the CBOL criteria using BLASTn, genetic distance, and monophyly
25 tree-based analyses (neighbor-joining, NJ, maximum likelihood, ML, and Bayesian
26 inference, BI). *rbcLa* universal primers amplified 69% of the samples and yielded 91.30%
27 bi-directional sequences. Sixty-three new *rbcLa* sequences were established. BLAST
28 discriminates 80.8% of the genus but only 15.4% of the species. Genetic distances
29 among *Quercus, Oreopanax*, and *Daphnopsis* species were nil. Contrastingly, Ericaceae
30 (5.6%), Euphorbiaceae (4.6%), and Asteraceae (3.3%) species displayed the highest within-
31 family genetic distances. According to the most recent angiosperm classification, NJ and
32 ML trees successfully resolved (100%) monophyletic species. ML trees showed the highest
33 mean branch support value (87.3%). Only NJ and ML trees could successfully
34 discriminate *Quercus* species belonging to different subsections: *Quercus martinezii* (white
35 oaks) from *Q. callophylla* and *Q. laurina* (red oaks). The ML topology could distinguish
36 species from the Solanaceae clade that the best BLAST match could not. Also, the BI
37 topology showed a polytomy in this clade, and the NJ tree displayed low-support branch
38 values. We do not recommend genetic-distance approaches for species discrimination.
39 More published *rbcLa* sequences are necessary for BLAST to be more effective. Instead,
40 the ML tree-based analysis displays the highest species discrimination among the tree-
41 based analyses. With the ML topology in selected genera, *rbcLa* helped distinguish
42 infrageneric taxonomic categories, such as subsections, grouping affine species within the
43 same genus, and discriminating species in most cases. Since the ML phylogenetic tree
44 could discriminate 48 species out of our 55 studied species, we recommend this approach to
45 resolve tropical montane cloud forest species using *rbcLa*, as an initial step and improving
46 DNA amplification methods.

47

48 **Introduction**

49 A biodiversity inventory is crucial as a first step to protecting species and ecosystems. A
50 significant portion of global biodiversity remains unnamed. Recent estimations indicate that
51 8.7 million species of multicellular organisms occur on Earth, but about 20% of those
52 species have been described using morphological approaches since 1750 (Centre for
53 Biodiversity Genomics, CBG, 2021). Thus, it is urgent to speed up the species
54 identification process (Hvistendahl, 2021). DNA barcoding was recently proposed to
55 identify species using short-standardized sequences and only requires a small sample of
56 tissue (Hebert et al., 2003). Cytochrome oxidase 1 (*CO1*) successfully discriminates against

many animal species but does not resolve plant species. The Consortium for the Barcode of Life's (CBOL) plant working group evaluated several plastid DNA regions based on universality, sequence quality, and species discrimination, recommending using a core of a 2-locus combination of partial genes *rbcLa* + *matK* as the plant barcode (Group C Hollingsworth et al., 2009). Such a universality has not been found in all plant groups, and other studies suggest using additional loci (Kress & Erickson, 2007; Fasekas et al., 2008; Pang et al., 2012; China plant B Li et al., 2011). Moreover, *matK* may work very well for species of orchids (Lahaye et al., 2008) but not for certain fern groups (Trujillo-Argueta et al., 2021). Furthermore, in some angiosperm genera, such as *Salix* (Percy *et al*., 2014) and *Quercus* (Piredda et al., 2011), plastid markers might not work at all.

Commented [MH2]: Consider make a change for this region also, make i tinto *mat*K with only "mat" in italic

Commented [MH3]: Plastid markers are thought to be not working for those mentioned specimen. Why the authors still using *rbc*La? How about the nuclear barcode region?

On average, the resolution of the tested DNA barcoding markers for plants is not as high as barcode markers used for many animal groups (CBOL Plant Working Group, 2009; Fazekas et al., 2008). Of the possible plant markers, *rbcLa* appears to be one of the best available. Although far from perfect, the resolution of *rbcLa* was shown to be better than those tested in arid plants in the United Arab Emirates (Maloukh et al., 2017) and Saudi Arabia (Bafeel et al., 2012). Also, *rbcLa* can be a valuable tool to identify species in conditions in which other methods are impractical. For instance, this marker was successfully used for studying the patterns of root diversity in old-field communities in Ontario, Canada (Kesanakurti et al., 2011). This kind of research is encouraging, but more studies are needed to explore the resolution potential of this marker for species in ecosystems other than those of temperate regions. The Neotropics are considered the richest region in biodiversity (Gaston & Williams 1996; Thomas 1999). However, some ecosystems in the neotropics have been little explored regarding DNA barcode studies. The available studies are often limited to a few plant groups such as orchids (Lahaye et al., 2008) or ferns (Nitta, 2020; Trujillo-Argueta et al., 2021).

Commented [MH4]: How about in the old tropic region? The observed or sequenced plants are already finished or still in progress?

In Mexico, the tropical montane cloud forest (TMCFs) is a top priority ecosystem for conservation due to its high diversity, endemism richness, and anthropogenic threats (Villaseñor, 2010; Toledo-Aceves et al., 2011). However, local DNA barcoding needs to be developed. This study aims to evaluate the performance of the plant core DNA barcode *rbcLa* using universal primers for vascular plants and without using other markers in a tropical montane cloud forest of the Mixteca Baja, Oaxaca, Mexico. We followed the three above-mentioned CBOL criteria and built a barcode library of native plant species for this region.

Commented [MH5]: This sentence need more supporting information about the importance of local DNA barcoding

Methods

Species this study and study site

One hundred samples of plants belonging to different families were collected in a tropical montane cloud forest at San Miguel Cuevas, Santiago Juxtlahuaca Municipality, Oaxaca, Mexico (17°15'00.96" N, 98°02'57.34", centroid coordinates). The climate in this area is semi-humid, temperate to semi-warm (1382 mm and 16.8 C, mean annual precipitation and

Commented [MH6]: Any information about the why only those 55 species used in this study? Any criterio for the species selection?

98 temperature, Fernandez-Eguiarte et al., 2020), with soils rich in organic matter (Instituto
99 Nacional de Estadística y Geografía, 2005), and a mean altitude of 2187m. The municipal
100 council of San Miguel Cuevas granted permission to conduct our field studies on their
101 lands. Two commissioners of the communal property of this municipality, Mr. Pedro Gil
102 (2017- July 2018) and Mr. Damián Domínguez (July 2018 - June 2019), were directly
103 responsible for such permissions. Mr. Heladio Luna Rodríguez, a San Miguel Cuevas
104 Community local authority member, supervised, guided, and helped us throughout the field
105 trips. In no case was the entire plant collected. Collecting the samples did not kill the
106 plants, which were left alive in their original places. Based on The International Union for
107 Conservation of Nature (IUCN) Red List (accessed December 19th, 2021) from all 55
108 species in this study, more than half (57.7%) were not previously registered (Data Deficit
109 DD); 43.6% belong to the Least Concern (LC) category. Also, *Daphnopsis*
110 *tuerckheimiana* holds the status of Near Threatened (NT); and *Oreopanax sanderianus*, that
111 of Vulnerable species (VU).

112 Plant vouchers were determined by the following specialists: Daniel Tejero-Díez, UNAM
113 FES Iztacala, México, lycopod and ferns; Sergio Zamudio, Institute of Ecology, Veracruz,
114 México, Berberidaceae; Rafael F. del Castillo, IPN CIIDIR Oaxaca, Mexico, Pinaceae;
115 Jesús Guadalupe González Gallegos, University of Guadalajara, Mexico, Lamiaceae;
116 Socorro González Elizondo, IPN CIIDIR Durango, Mexico, Cyperaceae and Ericaceae;
117 Susana Valencia Avalos, UNAM Facultad de Ciencias, Mexico, Fagaceae; J.R. Kuethe,
118 University of Auckland, New Zealand, Passifloraceae; and Rufina García, Abril Velasco-
119 Murguía and Rafael F. del Castillo, IPN CIIDIR Oaxaca, Mexico, the rest of the specimens.
120 The herbarium vouchers were deposited at the herbarium of CIIDIR Oaxaca, Instituto
121 Politécnico Nacional (OAX), pending for registration numbers due to the pandemic crisis.
122 The species and their IUCN Red List Status are shown on Table 1.

123

124 DNA amplification and sequencing

125 Several fresh leaves from each sampled plant were collected and placed in a Ziplock® bag.
126 The samples were kept at -20°C in a freezer until processed. The number of samples
127 collected per taxon was one and occasionally two.

128 Genomic DNA was extracted from 2mg leaf tissue with FastDNA SPIN kit and FastPrep®
129 (MP Biomedicals, USA) equipment.

130 DNA concentration (ng/µl) and purity (260/280A) from the genomic DNA extracted were
131 measured with a Biophotometer (Eppendorf®). Plant core barcoding partial gene rbcLa
132 was used for amplification. We used standard primers from the Canadian Center for DNA
133 Barcoding (CCDB) (Kuzmina, 2011), rbcLa-F
134 ATGTCACCACAAACAGAGACTAAAGC (Tate & Simpson, 2003) and rbcLa-R
135 GTAAAATCAAGTCCACCRCG (Kress & Erickson, 2007). rbcLa was amplified using a
136 25µL volume of reaction mixture: 15.8 µL of nuclease-free water, 5 µl MyTaq Buffer
137 reaction (kit MyTaqDNA Polymerase Bioline), 1µL of forward primer, 1µL of reverse

Commented [MH7]: Give it a space

Commented [MH8]: The reference is not found in the references list

Commented [MH9]: Did the voucher made into herbarium specimen?
If yes, please mention about the preparation

Commented [MH10]: Any criteria during the sampling? Especially the condition of the leaf

Commented [MH11]: Give it a space

Commented [MH12]: Give it a space

Commented [MH13]: Give it a space

Commented [MH14]: Give it a space

138 primer, 0.2 µL of MyTaq Polymerase and 2µL of isolated genomic DNA template. PCR
139 reaction was carried out using an Applied Biosystems Veriti® thermocycler. We followed
140 Fasekas et al. (2012) protocols for rbcLa amplification. The PCR temperature cycling
141 program was: 94°C for 4 min; 35 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 1 min;
142 final extension of 72°C for 10 min. Amplified PCR products were detected using agarose
143 gel electrophoresis (1.2 % agarose gel TBE) under UV light by staining with GelRed
144 Nucleic Acid (Biotium). PCR products were purified using the EZ-10 Spin Column PCR
145 Products Purification Kit (Biobasic). All PCR products were sequenced by Capillary
146 Electrophoresis Sequencing (CES) in an ABI 3130xl Genetic Analyser at the Laboratorio
147 Bioquímica Molecular UBIPRO FES Iztacala UNAM and with an AB3730 at the
148 Laboratorio de Servicios Genómicos LANGEBIO- CINVESTAV.

Commented [MH15]: Give it a space

Commented [MH16]: How much the DNA concentration used?

Commented [MH17]: Give it a space

Commented [MH18]: Give it a space

Commented [MH19]: Give it a space

Commented [MH20]: Give it a space

Commented [MH21]: Give it a space

149

150 DNA Alignment

151 *rbcLa* sequence chromatograms were manually edited and assembled into contigs using
152 CodonCode Aligner v.9.0.1 http://www.codoncode.com/aligner. Consensus sequences
153 were generated and aligned using MUSCLE (Edgar, 2004). These alignments were
154 examined by eye and corrected when necessary.

155

156 BOLD and Genebank

157 Our study was registered under the name "Diversity of a humid temperate forest in Oaxaca,
158 Mexico" project code DVHTF at The Barcode of Life Data System (BOLD,
159 http://www.boldsystems.org). BOLD is a bioinformatics workbench devoted to acquiring,
160 storing, analyzing, and publishing DNA barcode records (Ratnasingham & Hebert, 2007).
161 Three files were included in the metadata submitted to BOLD: 1) Specimen data file
162 including detailed voucher information, scientific names of the taxa sampled, collection
163 dates, geographical coordinates, elevation, collectors, identifiers, and habitat. 2) An image
164 file was submitted with high-quality specimen images from each plant. 3) A trace file was
165 submitted along with primers and the direction of sequences. Sequences uploaded to
166 BOLD were edited and aligned in FASTA format and referenced by Sample IDs.
167 Sequences were also submitted to the GenBank.

168

169 Species differentiation

170 To evaluate species discrimination using rbcLa sequences, we used three approaches:

171 a) The Basic Local Alignment Search Tool for nucleotide (BLASTn) method (Altschul et
172 al. 1990), which searches against the sequence database available online by the National
173 Center for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov.
174 Identification at the genus level was considered successful when all hits with the maximum
175 percent identity scores >99% involved a single genus. Species identification was

considered successful only when the highest maximal percent identity included a single species and scored >99% (Abdullah, 2017; Bafeel et al., 2012).

b) Genetic divergence. Interspecific and intraspecific distances were analyzed in MEGAX (Kumar et al., 2018). Genetic distance was inferred from 1000 replicates, and the evolutionary distances were computed using the Kimura 2-parameter method with gaps/missing data treatment adjusted using pairwise deletion. The genetic distances (%) of families, genera, and species were analyzed in the Barcode of Life Data Systems (BOLD, www.boldsystems.org) (Ratnasingham & Hebert 2007).

Commented [MH22]: Did the parameter chosed manually or was suggested base don the data set?

c) Monophyly tree-based analyses using Neighbor-Joining (NJ), Maximum Likelihood (ML), and Bayesian Inference (BI) analysis.

NJ was analyzed in MEGAX (Kumar et al., 2018) inferred from 1000 replicates, and the evolutionary distances were computed using the Kimura 2-parameter method with gaps/missing data treatment adjusted using pairwise deletion. ML analyses were run on the IQ-TREE web server (http://iqtree.cibiv.univie.ac.at). Internal node support and bootstrap analyses were calculated using 1000 iterations. Tree inference using Bayesian analysis was run on MrBayes 3.2.2 on XSEDE via the CIPRES supercomputer cluster (www.phylo.org) for 10 million generations. The resultant ML and BI trees were visualized in the interactive Tree of Life (iTOL) (Letunic & Bork 2019). We evaluated which of the tree-based methods (NJ, ML, and MB) recovered more monophyletic species with a bootstrap/posterior probabilities support of >70% (de Groot et al., 2011).

Results

DNA Amplification and sequencing success

We could successfully amplify 69% of the botanical samples collected. We studied 38 families, of which 27 had one species and 11 families 2 to 5 species (Table 1). Of the 55 studied species, 29.1% were herbs, and 70.9% were trees and shrubs. From this subset, we could obtain high-quality bidirectional sequences (>250bp) in 91.3% of the species, using the standard primers of the CCDB for the *rbcLa* barcode.

BLAST

Using BLASTn, we obtained 100% resolution in all the 38 families studied and 80.8% in 48 genera. Only 47.3% of our 55 studied species were previously registered in the *rbcLa* sequences of the GenBank database (Figure 1). We also contributed to 13 new species in the GenBank Taxonomy Database. These species were not previously registered for any other gene sequence. With the available accessions at the GenBank, we found that *rbcLa* can unambiguously discriminate only 15.4% of the studied species at the species level (Figure 1). Just four species, *Monnina xalapensis, Cnidoscolus aconitifolius, Iresine diffusa,* and *Lophosoria quadripinnata*, were found to best BLAST match to a single

217 species with more than 99% identity. Most of our *rbcL* sequences matched from 2-12
218 species with ≥99% maximal percent identity; and seven species, *Alnus acuminata, Solanum*
219 *hispidum, Quercus laurina, Quercus callophylla, Pinus montezumae, Osmanthus*
220 *americanus,* and *Physalis phyladelphica*, matched the *rbcL* sequences in the GenBank with
221 >30 different species. The best BLAST match identifications per species for
222 the *rbcLa* plastid barcode are shown in Table 2.
223
224 A specimen data file, image file, and trace file(s) were submitted to BOLD along with
225 edited and aligned sequences for each of our 63 botanical samples (55 species and eight
226 different duplicates) and can be accessed through the BOLD DNA database
227 (http://www.boldsystems.org) under the 'DVHTF' project. Sixty-three sequences were
228 obtained in this study for *rbcLa*, BOLD Process ID, and GenBank Accession numbers
229 (Table 1).
230
231 *Genetic divergence*
232
233 The distribution of intra- and interspecific K2P distances across all taxon pairs of our 55
234 species of plants of The Mixteca Baja, Oaxaca, tropical montane cloud forest, obtained
235 from partial gen *rbcLa* are shown in Figure 2. Mean pairwise genetic distance within
236 species was 0, within genus 0.65 ± 0.07, and 1.76 ± 0.03 within families. Congeneric
237 species of *Quercus*, *Daphnopsis,* and *Oreopanax* did not show genetic divergence.
238 Contrastingly, *Solanum*, *Deppea,* and *Pinus* displayed intergeneric differences (Table 3)
239 The mean genetic divergence observed in the studied families with two or more genera is
240 shown in Table 4. The highest mean divergence values were observed in the Ericaceae,
241 Euphorbiaceae, and Asteraceae families.
242
243
244 *Monophyly tree-based analyses*
245
246 Phylogenetic tree-based analysis using Neighbor-Joining (Supplementary Fig.S1),
247 Maximum Likelihood (Figure 3), and Bayesian Inference tree (Supplementary Fig.S2) were
248 reconstructed to evaluate our 55 species discrimination using the *rbcLa* barcode region. In
249 all cases, ferns and lycopodium were used as outgroups. These tree-based methods
250 evaluated which tree rendered the greatest species resolution and whether the barcode
251 sequences generated monophyletic species (Table 5). NJ and ML phylogenetic trees
252 resolved 100% of monophyletic species using *rbcLa.* Nevertheless, the clade support value
253 > 70% with a bootstrap of 1000 replicates yielded the most robust phylogeny in the ML
254 tree (87.3%) than the one obtained in the NJ tree (70.9%). Therefore, we present the ML
255 phylogenetic tree (Figure 3). Although the BI tree showed the highest clade support value
256 (92.7%), this tree did not resolve all 55 species as monophyletic species. Two polytomies
257 were observed in the clade of the *Quercus* species and the Solanaceae clade
258 (Supplementary Fig. S2).
259
260
261 **Discussion**

262 Our study reveals the advantages and limitations of the *rbcLa* barcode region for species
263 identification of vascular plant species of a neotropical montane cloud forest. First, the
264 amplification success was not universal, but bi-directional sequencing was highly
265 successful when feasible. BLAST identification at the genus level is accurate in most cases
266 but usually not for species identification. Finally, in selected genera, this marker helped
267 distinguish infrageneric taxonomic categories, such as subsections, and helps to group
268 affine species within the same genus. Below, we discuss in detail these issues.
269
270 Multiple factors can cause the absence of DNA amplification in some samples. Since we
271 could amplify *rbcLa* in several species, the possibilities of methodological failures or
272 problems with the reactants or the lab equipment used are unlikely. One possible cause of
273 the amplification failure is DNA degradation in some samples, as those were collected in
274 the field and brought to the lab. During this time, the tissues may become degraded in some
275 species. This appears to be a plausible explanation for cases in which DNA from tissue
276 samples was successfully amplified in one individual but not in another of the same
277 species. This is the case of *Solanum nigricans* (this study) and *Dryopteris wallichiana*,
278 which could not be amplified in this study but were successfully amplified in a previous
279 study using samples from different plants (Trujillo-Argueta et al., 2021). Another
280 possibility is that the pair of *rbcLa* universal premiers used may not work for certain
281 species. Our *rbcLa* amplification success (69%) could be increased using the alternative set
282 of universal primers proposed by CCDB for gene barcode *rbcL.*
283
284 Our sequencing success (91.30%) was high and similar to those reported in other works. In
285 a study of root diversity patterns using plastid gene *rbcL*, Kesanakurti et al. (2011)
286 registered 96% amplification success with 85% sequencing success. In another study that
287 identified Sicily's most threatened plant taxa, the amplification and sequencing successes
288 were 96% and 95%, respectively (Giovino et al., 2016). In a study of the temperate flora of
289 Canada, the use of *rbcLa* gave a 91.4% sequencing success (Burgess et al., 2011).
290 Our BLAST results were higher for genus discrimination (80.77%) than the values obtained
291 for species differentiation (15.38%). Results from other regions and species are variable.
292 For example, in wild, arid plants, discrimination at genus and species levels were lower
293 than ours: 50% and 8%, respectively (Bafeel et al., 2012), but higher in a comprehensive
294 study of the local flora of Canada (91% and 44%, Braukmann et al., 2017). In a study of
295 threatened species of Sicily, the discrimination at the genus level was lower (52%) but
296 higher at the species level (48%) than our results (Giovino et al., 2016). The peculiarities of
297 the biology of the studied species may also account for the observed discrimination
298 variability. Part of our low percent species discrimination results using BLASTn can be
299 explained by low marker resolution, as was noticed in those species that matched
300 their *rbcLa* sequence with more than 30 different species in the GenBank database (*Alnus*
301 *acuminate, Solanum hispidum, Quercus laurina, Quercus callophylla, Pinus montezumae,*
302 *Osmanthus americanus and Physalis phyladelphica*). Another explanation is misidentified
303 voucher specimens in public DNA databases, an issue that several authors have
304 acknowledged (e.g., Abdullah, 2017; Burguess et al., 2011). Since it is customarily to
305 describe species based on morphological characteristics, it is possible that hybridization and
306 polyploidy, which are common in plants, may contribute to decreasing barcoding species
307 discrimination (Hollinsworht, 2011; Fasekas et al., 2008). Since more than half of the
308 species in this study (52.72%) lacked comparative data in the GenBank database, it is

Commented [MH23]: Consider to explain this sentence clearly

Commented [MH24]: During the sample collection, did the ziplock contain nothing?
Did the authors not use silica gel and put it inside the ziplock?

Commented [MH25]: Fasekas or Fazekas?

309 necessary to increase the DNA barcode database, particularly for tropical wild plant
310 species. Indeed, we contributed to new 63 *rbcLa* sequences to BOLD, its metadata, and the
311 GenBank database. Although 42 of our species already had a *rbcLa* sequence on the
312 GenBank database, new records on these species might help discover new haplotypes or
313 geographical variants (Hajibabaei et al., 2007). Even if *rbcLa* does not have high species
314 discrimination, it does for genus discrimination, which for some ecological studies might
315 be enough (e.g., Kesanakurti et al., 2011).
316
317 Our distribution of intra- and interspecific genetic divergence (Figure 2) agrees with the
318 premise that a DNA barcode must exhibit high interspecific but low intraspecific
319 divergence (Lahaye et al., 2008). The percent interspecific divergence of this study (0.65) is
320 similar to those reported in other hotspot diversity areas such as the Mediterranean Basin
321 (0.89) (Giovino et al., 2016) and Southern Africa (0.82) (Lahaye et al., 2008). The lack of
322 genetic divergence observed in the three genera of trees: *Quercus* (*Q. martinezii, Q.*
323 *laurina, and Q. callophylla*); *Oreopanax* (*O. sanderianus and O. xalapensis*),
324 and *Daphnopsis* (*D. selerorum and D. tuerckheimiana*) concurs with Smith & Donoghue
325 (2008). These authors found that the rates of molecular evolution are low in woody plants
326 with long generation times compared to herbs. In the case of oaks (*Quercus),* several
327 attempts have been made to identify species in Italy, using different plastid barcodes
328 without success since hybridization and polyploidy are expected to be high in this group
329 (Piredda et al., 2011). Null genetic divergence obtained in *Oreopanax* and *Daphnopsis*
330 (Table 3) is of concern since *Oreopanax sanderianus* and *Daphnopsis tuerckheimiana* are
331 on the red list of IUCN. The highest values of genetic distance found in the Ericaceae
332 (5.57%), Euphorbiaceae (4.59%), and Asteraceae (3.3%) families that hold many herbs and
333 shrubs species agree with the assumption that *the rbcLa* barcode has a better species
334 differentiation for non-tree species. Moreover, a study conducted in a subalpine forest in
335 Southwest China found a better DNA barcode resolution for herbs than for tree species
336 (Tan et al., 2018). However, more studies are needed to confirm this trend in other species
337 and localities.
338
339 The phylogenetic arrangements found in our study using barcode *rbcLa* concur with the
340 recent Angiosperm Phylogeny Group classification (APG IV) of flowering plants (The
341 Catalog of Life Partnership, 2017). The percent monophyletic species resolution obtained
342 in this study using NJ (100%), ML (100%), and BI (85.45%) phylogenetic trees, was higher
343 compared to 17% of species resolution found in arid wild plants using ML trees (Bafeel et
344 al., 2012), barcoding the biodiversity of Kuwait (58%) using NJ trees (Abdullah, 2017) and
345 the 71.8% registered in two biodiversity hotspots of Mesoamerica and Southern Africa,
346 using ML and BI trees (Lahaye et al., 2008). Our ML phylogenetic tree showed the most
347 robust phylogeny (87.27%), *Ocotea helicterifolia, Quercus callophylla, Quercus laurina,*
348 *Iresine difussa, Berberis lanceolata, Moussonia deppeana,* and *Osmanthus americanus*,
349 could not be resolved as monophyletic species with a clade bootstrap support value $\geq$ 70%.
350 Most of these species are trees in agreement with the assumption that rates of molecular
351 evolution are low in woody plants compared to herbs (Smith & Donoghue, 2008). For those
352 species that could not be differentiated with the ML tree, we suggest the addition of a
353 second barcode.
354

Species discrimination can be improved by using tree-based phylogenetic methods rather than BLAST analysis and genetic distance approaches. For instance, using NJ and ML phylogenetic trees, it was possible to differentiate *Quercus martinezii* from *Q. laurina* and *Q. callophylla* (Supplementary Fig.S1, Figure 3) despite unsuccessful best BLAST matches and the null genetic divergence observed in *Quercus*. Based on an updated infrageneric classification of the oaks (Denk et al., 2017), *Q. martinezii* belongs to the white oaks (subsection *Quercus*), while *Q. callophylla* and *Q. laurina* belong to the red oaks (subsection *Lobatae*). In the Solanaceae family, three out of the five studied species (*Physalis philadelphica, Solanum hispidum*, and *Solandra maxima*) share high similitude with at least 30 species using the best BLAST match results. Furthermore, using our best BI tree, we observed a polytomy in the *Solanaceae* clade (Supplementary Fig.S2), and a low discrimination value in the NJ tree. However, these species could be resolved with our ML phylogenetic tree. Taxonomic species are usually described based on morphological characteristics that can easily be altered by local adaptation, phenotypic plasticity, or neutral morphological polymorphism, which may cause a single variable species to be classified as many species (e.g., Gemeinholzer & Bachmann, 2005). On the other hand, very recent divergence and little differentiation might contribute to the inability of barcoding to separate species in some cases (Birch et al., 2017).

Conclusions

DNA barcoding using *rbcLa* can be a promising identification tool primarily at the family and genus level for vascular plant species of the neotropical montane cloud forest. We identify three major problems with the use of this technique. First, the lack of a universal amplification capability is probably associated with DNA degradation in some cases, but without ruling out other factors requiring further study. Second, the inability to detect certain morphological species is probably not related to *rbcLa* itself but to biological (e.g., polyploidy and hybridization) and technical (misidentifications or taxonomic misclassifications) problems. Third, the few available registers in the BOLD and GenBank databases (more than half of our species, 52.72%, did not have previous *rbcLa* sequence records). Indeed, we contributed new 13 species to the GenBank Taxonomy Database and 63 new sequences for *rbcLa* in BOLD and GenBank. We found preliminary evidence suggesting that the ability of the marker to discriminate species is not randomly distributed among taxa. Herb and shrub species in the Asteraceae, Ericaceae and Euphorbiaceae families showed the highest genetic distance using *rbcLa*, which can be helpful to distinguish congeneric species. Contrastingly, we detected nil genetic divergence among congeneric species in long-lived tree genera, *Quercus, Oreopanax*, and *Daphnopsis.* Nonetheless, the accuracy for discriminating species can be substantially improved using tree-based analysis. While BLAST and genetic distance approaches could not differentiate *Quercus* species, NJ and ML could successfully separate white oaks (*Quercus martinezii)* from red oaks (*Q. callophylla* and *Q. laurina)*. Also, most species in the Solanaceae family that showed unsuccessful BLAST results and low genetic distance could be discriminated against with ML phylogenetic tree. The ML phylogenetic tree showed the most robust phylogeny (87.27%) of all our 55 studied species of the tropical

400 montane cloud forest of San Miguel Cuevas in Oaxaca state, Mexico. The establishment of
401 this local barcode database will be valuable for a broad range of potential ecological,
402 conservational, and phylogenetic applications.

403

404 **Acknowledgments**

405 Raúl Rivera García and Heladio Luna, assisted in field work. We are thankful to San
406 Miguel Cuevas Community at Oaxaca State, México, for permitting access to their forest.

407

408

409 **References**

Commented [MH26]: The reference was typed inconsistently.
Consider make it according to the jorunal format, especially authors and the years of publication. Should it be typed inside the bracket or do not need bracket?

410 Abdullah M T. 2017. Conserving the biodiversity of Kuwait through DNA barcoding the
411 flora. D. Phil. Thesis, University of Edinburgh & Royal Botanic Garden Edinburgh.

412 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local
413 alignment search tool. Journal of molecular biology, 215(3), 403-410.

414 Bafeel, S. O., Arif, I. A., Bakir, M. A., Al Homaidan, A. A., Al Farhan, A. H., & Khan, H.
415 A. (2012). DNA barcoding of arid wild plants using rbcL gene sequences. Genet. Mol. Res,
416 11(3), 1934-41.

417 Birch, J. L., Walsh, N. G., Cantrill, D. J., Holmes, G. D., & Murphy, D. J. (2017). Testing
418 efficacy of distance and tree-based methods for DNA barcoding of grasses (Poaceae tribe
419 Poeae) in Australia. PLoS One, 12(10), e0186259.

420 Braukmann, T. W., Kuzmina, M. L., Sills, J., Zakharov, E. V., & Hebert, P. D. (2017).
421 Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. PLoS
422 one, 12(1), e0169515.

423 Burgess, K. S., Fazekas, A. J., Kesanakurti, P. R., Graham, S. W., Husband, B. C.,
424 Newmaster, S. G., ... & Barrett, S. C. (2011). Discriminating plant species in a local
425 temperate flora using the rbcL+ matK DNA barcode. Methods in Ecology and Evolution,
426 2(4), 333-340.

427 CBOL Plant Working Group (2009) A DNA barcode for land plants. Proceedings of the
428 National Academy of Sciences of the United States of America, 106, 12794–12797.

429 Centre for Biodiversity Genomics, University of Guelph (2021). The Global Taxonomy
430 Initiative 2020: A Step-by-Step Guide for DNA Barcoding. Technical Series No. 94.
431 Secretariat of the Convention on Biological Diversity, Montreal, 66 pages.

432 China Plant, B. O. L. G., Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., ... & Duan,
433 G. W. (2011). Comparative analysis of a large dataset indicates that internal transcribed

434 spacer (ITS) should be incorporated into the core barcode for seed plants. Proceedings of
435 the National Academy of Sciences of the United States of America, 108(49), 19641-6.

436 De Groot, G. A., During, H. J., Maas, J. W., Schneider, H., Vogel, J. C., & Erkens, R. H.
437 (2011). Use of rbcL and trnL-F as a two-locus DNA barcode for identification of NW-
438 European ferns: an ecological perspective. PLoS one, 6(1), e16371.

439 Denk, T., Grimm, G. W., Manos, P. S., Deng, M., & Hipp, A. L. (2017). An updated
440 infrageneric classification of the oaks: review of previous taxonomic schemes and synthesis
441 of evolutionary patterns. Oaks physiological ecology. Exploring the functional diversity of
442 genus Quercus L., 13-38.

443 Doumenge, C.; Gilmour, D.; Ruíz-Pérez, M., and Blockhus, J. 1995. Tropical montane
444 cloud forests: conservation status and management issues. In: Hamilton, L. S.; Juvick, J. O.,
445 and Scatena, F. N., Ed. Tropical Montane Cloud Forests. New York: Springer-Verlag; pp.
446 24-37.

447 Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
448 throughput. Nucleic acids research, 32(5), 1792-1797.

449 Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G.,
450 Husband, B. C., ... & Barrett, S. C. (2008). Multiple multilocus DNA barcodes from the
451 plastid genome discriminate plant species equally well. PloS one, 3(7), e2802.

452 Fazekas, A. J., Kuzmina, M. L., Newmaster, S. G., & Hollingsworth, P. M. (2012). DNA
453 barcoding methods for land plants. In DNA barcodes (pp. 223-252). Humana Press,
454 Totowa, NJ.

455 Fernandez-Eguiarte A., Zavala-Hidalgo, J., Romero-Centeno, R., 2020. Atlas Climático
456 Digital de México (versión 2.0). Centro de Ciencias de La Atmósfera. Universidad
457 Nacional Autónoma de México. Servicio Meteorológico Nacional, Comisión Nacional
458 Del Agua. http://atlasclimatico.unam.mx/atlas/kml/ (accessed 13 March 2020).
459

460 Gaston, K. J. and Williams, P. H. 1996. Spatial patterns in taxonomic diversity. In: Gaston,
461 K. J. Biodiversity: A biology of numbers and difference. Oxford, UK: Blackwell Science,
462 Limited 202-229.

463 Gemeinholzer, B., & Bachmann, K. (2005). Examining morphological and molecular
464 diagnostic character states of Cichorium intybus L.(Asteraceae) and C. spinosum L. Plant
465 Systematics and Evolution, 253(1), 105-123.

466 Giovino, A., Marino, P., Domina, G., Scialabba, A., Schicchi, R., Diliberto, G., ... &
467 Scibetta, S. (2016). Evaluation of the DNA barcoding approach to develop a reference data-
468 set for the threatened flora of Sicily. Plant Biosystems-An International Journal Dealing
469 with all Aspects of Plant Biology, 150(4), 631-640.

470 Group, C., Hollingsworth, P.W. et al. A DNA barcode for land plants. Proc. Natl. Acad.
471 Sci. USA, 106 31, 12794-12797 (2009).

Commented [MH27]: This reference can not be found in the text above

Commented [MH28]: This reference can not be found in the text above

472 Hajibabaei, M., Singer, G. A., Hebert, P. D., & Hickey, D. A. (2007). DNA barcoding: how
473 it complements taxonomy, molecular phylogenetics and population genetics. TRENDS in
474 Genetics, 23(4), 167-172.

475 Hebert, P. D., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological
476 identifications through DNA barcodes. Proceedings of the Royal Society of London. Series
477 B: Biological Sciences, 270(1512), 313-321.

478 Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. Proceedings of the
479 National Academy of Sciences, 108(49), 19451-19452.

480 Hvistendahl, M. (2021). A Tale of Two Worlds. Scientific American Special Edition, The
481 age of humans 30 (4), 37-41.

482 Instituto Nacional de Estadística y Geografía (INEGI), 2005. Prontuario de Informaci´on
483 Geogr´afica Municipal de los Estados Unidos Mexicanos: Santiago Juxtlahuaca,
484 Oaxaca. http://www3.inegi.org.mx/contenidos/app/mexicocifras/datos_
485 geograficos/20/20469.pdf (accessed 20 March 2020).
486
487 Kesanakurti, P. R., Fazekas, A. J., Burgess, K. S., Percy, D. M., Newmaster, S. G.,
488 Graham, S. W., ... & Husband, B. C. (2011). Spatial patterns of plant diversity below-
489 ground as revealed by DNA barcoding. Molecular Ecology, 20(6), 1289-1302.

490 Kress, W. J., & Erickson, D. L. A two-locus global DNA barcode for land plants: the
491 coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS one 2 6,
492 e508 (2007).

493 Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular
494 evolutionary genetics analysis across computing platforms. Molecular biology and
495 evolution, 35(6), 1547.

496 Kuzmina M.L. 2011. Canadian Center for DNA Barcoding (CCDB) Protocols. Primers
497 sets for plants and fungi. Available: http://www.dnabarcoding.ca

498 Lahaye, R., Van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., ... &
499 Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. Proceedings of
500 the National Academy of Sciences, 105(8), 2923-2928.

501 Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new
502 developments. Nucleic acids research, 47(W1), W256-W259.

503 Maloukh, L., Kumarappan, A., Jarrar, M., Salehi, J., El-Wakil, H., & Rajya Lakshmi, T. V.
504 (2017). Discriminatory power of rbcL barcode locus for authentication of some of United
505 Arab Emirates (UAE) native plants. 3 Biotech, 7(2), 1-7.

506 Nitta, J. H. (2008). Exploring the utility of three plastid loci for biocoding the filmy ferns
507 (Hymenophyllaceae) of Moorea. Taxon, 57(3), 725-736.

508 Pang, X., Liu, C., Shi, L., Liu, R., Liang, D., Li, H., ... & Chen, S. (2012). Utility of the
509 trnH–psbA intergenic spacer region and its combinations as plant DNA barcodes: a meta-
510 analysis. PloS one, 7(11), e48833.

511 Percy, D. M.; Argus, G. W.; Cronk, Q. C.; Fazekas, A. J.; Kesanakurti, P. R.; Burgess, K.
512 S.; Husband, B. C.; Newmaster, S. G.; Barrett, S. C. H., and Graham, S. W. 2014.
513 Understanding the spectacular failure of DNA barcoding in willows (Salix): Does this
514 result from a trans-specific selective sweep? Molecular Ecology 23:4737-4756.

515 Piredda, R., Simeone, M. C., Attimonelli, M., Bellarosa, R., & Schirone, B. (2011).
516 Prospects of barcoding the Italian wild dendroflora: oaks reveal severe limitations to
517 tracking species identity. Molecular ecology resources, 11(1), 72-83.

518 Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System
519 (http://www. barcodinglife. org). Molecular ecology notes, 7(3), 355-364.

520 Smith, S. A., & Donoghue, M. J. (2008). Rates of molecular evolution are linked to life
521 history in flowering plants. science, 322(5898), 86-89.

522 Tan, S. L., Luo, Y. H., Hollingsworth, P. M., Burgess, K. S., Xu, K., Li, D. Z., & Gao, L.
523 M. (2018). DNA barcoding herbaceous and woody plant species at a subalpine forest
524 dynamics plot in Southwest China. Ecology and Evolution, 8(14), 7195-7205.

525 Tate, J. A., & Simpson, B. B. (2003). Paraphyly of Tarasa (Malvaceae) and diverse origins
526 of the polyploid species. Systematic Botany, 28(4), 723-737.

527 The Catalogue of Life Partnership (2017). APG IV: Angiosperm Phylogeny Group
528 classification for the orders and families of flowering plants. Checklist dataset
529 https://doi.org/10.15468/fzuaam accessed via GBIF.org on 2022-02-21.

530 Thomas, W. W. 1999. Conservation and monographic research on the flora of Tropical
531 America. Biodiversity & Conservation 8:1007-1015.

532 Toledo-Aceves, T.; Meave, J. A.; González-Espinosa, M., and Ramírez-Marcial, N. 2011.
533 Tropical montane cloud forests: Current threats and opportunities for their conservation and
534 sustainable management in Mexico. Journal of Environmental Management 92:974-981.

535 Trujillo-Argueta, S., Del Castillo, R. F., Tejero-Diez, D., Matias-Cervantes, C. A., &
536 Velasco-Murguía, A. (2021). DNA barcoding ferns in an unexplored tropical montane
537 cloud forest area of southeast Oaxaca, Mexico. Scientific reports, 11(1), 1-12.

538 Villaseñor, J. (2010). El bosque húmedo de montaña en México y sus plantas vasculares:
539 catálogo florístico-taxonómico. Comisión Nacional para el Conocimiento y Uso de la
540 Biodiversidad- Universidad Nacional Autónoma de México. 40 pp. México, D.F.